Molecular Phylogenetics and Evolution 48 (2008) 1106–1119

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

A novel supermatrix approach improves resolution of phylogenetic relationships in a comprehensive sample of danthonioid grasses

Michael D. Pirie [a,*], Aelys M. Humphreys [a], Chloe Galley [a], Nigel P. Barker [b], G. Anthony Verboom [c], David Orlovich [d], Suzy J. Draffin [d], Kelvin Lloyd [e], C. Marcelo Baeza [f], Maria Negritto [f], Eduardo Ruiz [f], J. Hugo Cota Sanchez [g], Elizabeth Reimer [g], H. Peter Linder [a]

[a] *Institute for Systematic Botany, University of Zurich, Switzerland*
[b] *Molecular Ecology and Systematics Group, Department of Botany, Rhodes University, Grahamstown, South Africa*
[c] *Department of Botany, University of Cape Town, Cape Town, South Africa*
[d] *Department of Botany, University of Otago, New Zealand*
[e] *Landcare Research, Private Bag 1930, Dunedin, New Zealand*
[f] *Departamento de Botánica, Universidad de Concepción, Concepción, Chile*
[g] *Biology Department, University of Saskatchewan, Saskatchewan, Canada*

ARTICLE INFO

Article history:
Received 18 January 2008
Revised 15 May 2008
Accepted 22 May 2008
Available online 2 July 2008

Keywords:
Coding and non-coding DNA sequence data
Conflict
Danthonioideae
Missing data
Phylogeny reconstruction
Poaceae
Sampling strategy
Supermatrix

ABSTRACT

Phylogeny reconstruction is challenging when branch lengths vary and when different genetic loci show conflicting signals. The number of DNA sequence characters required to obtain robust support for all the nodes in a phylogeny becomes greater with denser taxon sampling. We test the usefulness of an approach mixing densely sampled, variable non-coding sequences (*trnL*-F; *rpl16*; *atpB-rbcL*; ITS) with sparsely sampled, more conservative protein coding and ribosomal sequences (*matK*; *ndhF*; *rbcL*; 26S), for the grass subfamily Danthonioideae. Previous phylogenetic studies of Danthonioideae revealed extensive generic paraphyly, but were often impeded by insufficient character and taxon sampling and apparent inter-gene conflict. Our variably-sampled supermatrix approach allowed us to represent 79% of the species with up to c. 9900 base pairs for taxa representing the major clades. A 'taxon duplication' approach for taxa with conflicting phylogenetic signals allowed us to combine the data whilst representing the differences between chloroplast and nuclear encoded gene trees. This approach efficiently improves resolution and support whilst maximising representation of taxa and their sometimes composite evolutionary histories, resulting in a phylogeny of the Danthonioideae that will be useful both for a wide range of evolutionary studies and to inform forthcoming realignment of generic delimitations in the subfamily.

1. Introduction

Molecular phylogenetic analysis is easy when branch-lengths are similar and genetic loci are congruent. With increasing length variation, such as in groups including clades that have undergone rapid diversifications (e.g., Bakker et al., 1998; Richardson et al., 2001), it becomes progressively more difficult to resolve all the nodes on a tree. Conversely, levels of character sampling necessary to resolve certain clades may be greatly in excess of those required to obtain robust support for many others. The overall level of character sampling necessary for resolution increases with increasing density of taxon sampling (Bremer et al., 1999). Taxon sampling nevertheless needs to be dense for many studies, particularly if morphology does not provide a robust assessment of group membership. Inter-locus incongruence poses a further challenge to phylogeny reconstruction, as conflicting data partitions often cannot be combined to improve resolution (Bull et al., 1993), and even assessing the full extent of conflict between gene trees is only possible to within the limits of the resolution they afford individually (Hughes et al., 2006). Potential difficulties for molecular phylogenetic analysis are thus compounded in groups with variable diversification rates, complex distributions of morphological attributes, and reticulate evolutionary histories.

Various solutions to these problems have been proposed. Numerous studies have sampled multiple sequences from single linkage regions, such as plastid genomes (e.g. Olmstead and Swere, 1994; Galley and Linder, 2007), which can be combined, irrespective of reticulations in the species phylogeny, to recover a resolved tree. Some authors advocate sampling of multiple nuclear encoded gene regions (Hughes et al., 2006). Such markers can be more variable than chloroplast genes in plants and can potentially be used to address the issue of reticulating species phylogeny, a benefit not provided by plastid trees. However, they present

* Corresponding author. Fax: +41 (0)44 634 8404.
E-mail address: michael.pirie@systbot.uzh.ch (M.D. Pirie).

1055-7903/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2008.05.030

challenges in assessments of orthology and may be relatively costly and time consuming to produce when applied to large groups. Although in general sampling more characters (Rosenberg and Kumar, 2001) or under certain circumstances increasing taxon sampling (Rydin and Kallersjö, 2002; Zwickl and Hillis, 2002; Rokas and Carroll, 2005) has been shown to improve phylogenetic accuracy, neither of these approaches directly addresses the differences in data quality required for resolution at different levels in a phylogeny. In contrast, a potentially more efficient sampling strategy has been proposed, in which a 'scaffold' of conservative sequences, representing a small number of placeholder taxa, is combined with a larger number of more variable sequences representing the entire study group (Wiens et al., 2005; Wiens, 2006). The resulting data matrix can be described as a 'supermatrix'; i.e., the matrix is not complete and many of the characters may be coded as missing for many of the taxa (Sanderson et al., 1998).

Phylogenetic studies of the grass subfamily Danthonioideae demonstrate each of the above mentioned problems. Danthonioideae comprises approximately 288 recognised species, representing just fewer than 3% of the estimated 10,000 species of grasses (family Poaceae). The species of Danthonioideae are distributed across the world's temperate regions, but are particularly diverse and abundant in the Southern Hemisphere. The monophyly of Danthonioideae is well established on the basis of DNA sequence data (Barker et al., 1995, 1999; GPWG, 2001) and the recognition of the group as a coherent entity is also corroborated by the presence of haustorial synergids in the embryos of the member species (Verboom et al., 1994). It is a member of the PACCAD/PACCMAD clade (Panicoideae, Arundinoideae, Chloridoideae, Centrothecoideae, Aristidoideae and Danthonioideae; GPWG, 2001; including the Micrairoideae; Sánchez-Ken et al., 2007). Bouchenak-Khelladi et al. (2008), in large multi-plastid gene region analyses of the grasses, resolved Danthonioideae as sister to Chloridoideae with moderate bootstrap support and high Bayesian posterior probability.

A number of recent phylogenetic studies have used chloroplast DNA sequences for phylogenetic reconstruction in Danthonioideae, often in combination with the highly variable nuclear encoded ITS region of ribosomal DNA (Barker et al., 1995, 2000, 2003, 2007; Verboom et al., 2006; Galley and Linder, 2007). These studies have revealed extensive variation in branch-lengths, leaving parts of the tree stubbornly unresolved despite significant inputs of data (Barker et al., 2007). A number of these areas of uncertainty are also associated with apparent conflict between nuclear and chloroplast encoded sequence markers (Verboom et al., 2006; Barker et al., 2007). Finally, in the absence of well resolved phylogenetic trees, interpretation of patterns of morphological variation in Danthonioideae has been controversial. Phylogenetic trees based on morphological characters revealed rampant homoplasy and consequently little support for groupings (Linder and Verboom, 1996; Linder and Davidse, 1997; Barker et al., 2003). This complexity is clearly reflected in the unstable taxonomic history of the subfamily (Linder and Barker, 2005; Verboom et al., 2006) and extensive para- and polyphyly of the currently delimited genera as shown by phylogenetic analysis of nucleotide sequence data (Barker et al., 2003, 2007; Verboom et al., 2006).

A robust phylogeny of the subfamily will be useful for addressing a range of evolutionary questions, and a new generic classification of Danthonioideae based on clade membership is warranted. To date, ca.158 species of Danthonioideae have been included in molecular phylogenetic studies (Verboom et al., 2006; see Barker et al., 2007; Galley and Linder, 2007), representing 54% of the total species numbers, including 89% of the *Pentaschistis* clade but only 30% of the rest of the subfamily. The difficulty and controversy associated with assigning danthonioid species to genera on the basis of their morphology means that a high proportion of the species need to be represented in the phylogeny in order to maximize the predictive value of the resulting classification. Based on the already identified conflict between chloroplast and nuclear tree topologies, it would appear that adequate representation of the Danthonioideae phylogeny requires not just one, but at least two densely sampled and sufficiently resolved gene trees. In this paper we evaluate the performance of a supermatrix approach in resolving the phylogenetic relationships of Danthonioideae, based on chloroplast and nuclear DNA sequence markers.

2. Materials and methods

2.1. Taxon sampling and DNA sequence data

We included all species that could be sampled. Each species was represented by a single sample, with the exception of monotypic genera (or isolated species that might be recognised as such) which were represented by multiple samples, as were taxonomically problematic species and those with disjunct distributions. In the latter cases, samples were chosen to represent the different morphological forms or geographic regions, respectively. Sequences, and the samples from which they originated, were obtained from previous studies (Barker et al., 1995, 2000, 2003,2007; Verboom et al., 2006; Galley and Linder, 2007), or from newly collected, silica dried samples, and further samples were extracted from herbarium material. In total, 256 samples of 227 ingroup species (plus 14 sub-specific taxa) were analysed, representing 79% of the currently recognised danthonioid grass species (see Table 1: genera and species numbers; and Appendices 1 and 2: accessions tables). New data were generated for three outgroup taxa (*Centropodia glauca*, *Merxmuellera papposa* and *M. rangei* [Chloridoideae]). Protein coding sequences (see below) of a further nine outgroup taxa, eight representing all the grass subfamilies most closely related to Danthonioideae (PACCMAD clade), plus one more distant Poaceae outgroup, *Hordeum* (Pooideae; BEP clade) were obtained from previous studies and from GenBank (see Appendix 1).

We used two sampling strategies: non-coding regions with comprehensive taxon sampling to assign species to clades, and protein coding and nuclear ribosomal gene regions for a representative subset of taxa, selected on the basis of preliminary results from non-coding regions, to infer inter-clade relationships.

Non-coding regions (*trnL-F*, *rpl16*, *atpB-rbcL*, and ITS) were selected because of their high variability, including length variation, and in order to maximise overlap with existing data sets (*trnL-F*: Verboom et al., 2006; *trnL* intron: Barker et al., 2007; *trnL-F*, *rpl16* and *atpB-rbcL*: Galley and Linder, 2007). We sampled these regions for as many taxa as possible, thus optimising the alignment and maximising the information available in alignment gaps (coded separately).

Protein coding regions *matK* (including non-coding flanking *trnK* spacer regions), *rbcL*, and *ndhF* and (partial) sequences of the 26S ribosomal RNA gene were sampled for selected representatives of the major clades, identified by non-coding data, in order to resolve relationships between those clades. By selecting at least two taxa of each clade, bracketing the basal node, we incorporated both a test of the phylogenetic signal and a safeguard against laboratory error. *matK* and *ndhF* have been shown to be sufficiently variable to be useful for phylogeny reconstruction at relatively low taxonomic levels in numerous plant groups (Kim and Jansen, 1995; Baum et al., 1998). *rbcL* is less variable, but as an existing *rbcL* matrix was available (Barker et al., 2007), this was expanded to include most of the same taxa. Sequences from the 26S gene have been shown to be as variable, or slightly less so, than *rbcL* in seed plants (Kuzoff et al., 1998). Despite this relatively low variability, 26S is part of the same tandem repeat region of nuclear

Table 1
Genera of Danthonioideae: distribution, species numbers, and numbers sampled for this study. Generic concepts follow Conert (1970) for *Merxmuellera*; Linder and Verboom (1996) for *Rytidosperma* clade (sensu Barker et al., 2000; though a number are currently treated as *Merxmuellera* or *Danthonia*), and Barker (1995) for *Pseudopentameris*. Numbers in parentheses represent additional subspecies and varieties, i.e. excluding autonyms (which have not always been described)

Genus	Continental distribution	No. species	No. species sampled
Austrodanthonia H.P.Linder	Australia, New Zealand, New Guinea	31 (1)	22 (0)
Chaetobromus Nees	Africa	1 (2)	1 (1)
Chionochloa Zotov	New Zealand, Australia	24 12)	23 (0)
Cortaderia Stapf	New Zealand, South America, New Guinea	24 (1)	17 (0)
Danthonia DC.	South & North America, Europe, Africa, India	28 (6)	18 (1)
Joycea H.P.Linder	Australia	3	3
Karroochloa Conert et A.M.Türpe	Africa	4	3
Lamprothyrsus Pilger	South America	2	2
Merxmuellera Conert	Africa	18 (1)	14 (1)
Notochloe Domin	Australia	1	1
Notodanthonia Zotov	Australia, New Zealand	5	4
Pentameris Beauv.	Africa	9 (1)	8 (0)
Pentaschistis (Nees) Stapf	Africa	70 (6)	64 (4)
Plinthanthesis Steud.	Australia	3	3
Prionanthium Desvaux	Africa	3	2
Pseudopentameris Steud.	Africa	4	4
Rytidosperma Steud.	Australia, New Guinea, New Zealand, South America	41 (3)	24 (0)
Schismus P.Beauv.	Africa	5	4
Tribolium Desv.	Africa	12	10
Total		288	227 (79%)

ribosomal DNA as ITS. It may therefore be possible to combine 26S with ITS in phylogenetic analyses despite incongruence with cpDNA markers, to obtain a better resolved nrDNA tree. In this way the already demonstrated conflict with the cpDNA may be more readily assessed.

2.2. DNA extraction, PCR, sequencing, and sequence alignment

Total genomic DNA was extracted using DNeasy plant mini-kits (Qiagen GmbH, Hilden, Germany). Primers used for PCR amplification and sequencing are presented in Table 2.

PCR protocols were as follows: cpDNA: per 25 μl reaction we included 2.5 μl buffer, 2.0 μl MgCl, 4.0 μl dNTPs, 1 μl 0.4% BSA, 0.1 μl taq polymerase, 0.5 μl each of 10 μM solutions of the two primers, and 1 μl DNA template. The PCR programs were of an initial 4 min: 94 °C followed by 35 cycles of 30 s: 94 °C; 1 min: 50–55 °C; 1.5–3 min. (longer for longer products, and higher temperature when signs of non-specific amplification were observed): 72 °C; and a final extension of 7 min: 72 °C. PCR reagents for ITS and 26S amplification followed the cpDNA protocol, with the addition of 1 μl DMSO. The PCR program for ITS was also the same as for the cpDNA, but the 26S program followed Muellner et al. (2003): 3 min: 95 °C followed by a single cycle of 1 min: 95 °C; 1 min 45 °C; 1 min: 72 °C, and 35 cycles of 1 min: 94 °C; 1 min: 48 °C; 1 min: 72 °C; and a final extension of 10 min: 72 °C. PCR products were purified using Gene Elute PCR purification kits (Sigma-Aldrich, Inc.; St. Louis, MO, USA) and cycle-sequenced with the PCR and additional primers (Table 2) using Applied Biosystems (Foster City, CA, USA) Big Dye terminator kits. Cycle-sequence products were analysed by electrophoresis using an automatic sequencer 3130XL Genetic Analyzer (Applied Biosystems). In all cases it was possible to directly amplify and sequence PCR products without an intermediate cloning step: no multiple signals were observed in the resulting sequence trace files.

The matrix of non-coding sequence data comprised 257 *trnL-F* sequences; 229 *rpl16*; 218 *atpB-rbcL*; and 217 ITS (see Table 3 and Appendix 1). Due to difficulty in amplification of some lower quality samples it was not possible to sample all taxa for all markers. Sixty-two taxa were sampled for *matK* and *rbcL*, 63 for *ndhF* and 35 for 26S (see Appendix 1). DNA sequences were edited using Sequencher 4.6 (Gene Codes Corporation; MI, USA) and aligned manually using MacClade (Maddison and Maddison, 2005). Areas of the alignments in which the assessment of homology was ambiguous were excluded from the analyses. Gaps in the alignments were coded as present/absent characters, following the simple gap coding method of Simmons and Ochoterena (2000) as implemented by SeqState (Muller, 2006).

2.2.1. Testing for congruence and combining data partitions

Parsimony analyses (see below) were performed on each of the markers separately. The resulting topologies were inspected for conflicting nodes with 70% or higher bootstrap support (BS). We refer to nodes with less than 70% BS as unsupported; those between 70 and 79% as weakly supported; 80 –89% as moderately supported and >90% as strongly supported. Where no supported conflict was found, data partitions were combined. Samples with

Table 2
DNA sequence regions and primers used for PCR amplification and sequencing

Marker	Primers PCR	Primers sequencing	Reference/sequence
trnL-F	C/F, C/Danth_intR, Danth_intF/F	C, F	Taberlet et al. (1991); this study: Danth_intF = 5′-AGA ATT ATT GTG AAT CCA TTC C-3′; Danth_intR = 5′-GAT TAC TCA ATA TTC GAT TGG-3
rpl16	F71/*R1000	F71, *R1000	Baum et al., (1998), *Galley and Linder (2007)
atpB-rbcL	f1c/r1a2	f1c, r1a2	Hardy and Linder (2005)
ndhF	1/1318R, 972/2110R	1, 1165R,	Olmstead and Sweere (1994)
matK	*mk_F1 or s51F/*mk_R1	s51F, W, 1210R, 7B, 9R, *mk_R1	Hilu et al. (1999), * Moline and Linder, 2005)
rbcL	Z1/R3, F2/1374R	Z1, R3, F2, 1374R	Barker et al. (2007)
ITS	L/4	L, 4	Baum et al. (1998)
26S	N-nc26S1/1229Rev	N-nc26S1, 1229Rev	Kuzoff et al. (1998)

Table 3
Comparison of variable and informative characters for coding and non-coding cpDNA and nrDNA markers, and numbers of informative characters per taxon and per sequence added (given that different markers were obtained with differing numbers of sequences) based on the 35 taxa subset sampled for 26S

	Parsimony informative bases	Informative indels	Total informative characters (C)	C/T	C/seq
cpDNA: non-coding	274 (9.8%)	49	323	9.2	1.53
cpDNA: coding	456 (8.3%)	18	474	13.5	0.97
Nuclear: non-coding	123 (19.3%)	0	123	3.5	1.76
Nuclear: coding	74 (6.5%)	0	61	1.7	0.87
Total cpDNA	730 (9.0%)	67	797	22.8	1.14
Total nuclear	197 (10.7%)	0	197	5.6	1.41
Grand total	927 (9.3%)	67	994	28.4	1.18

missing data for one or more partitions, due to low quality of DNA in some cases, were included in the combined analyses, leading to a total of 260 samples for combined non-coding cpDNA and combined coding and non-coding cpDNA (excluding eight outgroups for which only coding data was used). The combined nrDNA matrix comprised 216 taxa. Where conflicting nodes were found (these were only between the cp- and nrDNA), the corresponding inconsistently placed taxa were duplicated in the matrix. One taxon copy was represented by the corresponding cpDNA sequences only, with the nrDNA partition coded as missing data; the other taxon copy by nrDNA only with the cpDNA coded as missing (= 'total combined DNA'). This procedure is discussed in more detail in a forthcoming paper (Pirie et al. unpublished manuscript). The positions of the following taxa were subject to such conflict (numbered as follows in Fig. 1): (1) *Tribolium pusillum* (1 sample); (2) *T. ciliare* (1 sample); (3) *Notochloe microdon* (3 samples); (4) a clade including all South American species of *Cortaderia* plus *Lamprothyrsus*, without New Zealand *Cortaderia*, *C. archboldii* or *C. pilosa*, (in total 11 species, representing the same conflict); (5) *Danthonia alpina* (1 sample); (6) *Chionochloa australis* (1 sample); (7) *Merxmuellera arundinacea* (2 samples); and (8–17) 10 species of *Pentaschistis* (each representing a separate incidence of conflict). These 30 samples, representing 27 species, were thus represented twice in the matrices, bringing the number of 'taxa' included in the total combined DNA matrix up to 290.

2.2.2. Parsimony analysis

Parsimony analysis was performed using the software package PAUP* 4.0b10 (Swofford, 2000), assuming unordered character state transformation (Fitch parsimony; Fitch, 1971) and equal weights. A two stage heuristic search strategy was used to find the set of most parsimonious trees. First, 100 parsimony ratchet searches (Nixon, 1999) of 100 generations each were performed, using PAUPrat (Sikes and Lewis, 2001), to find islands of shortest trees. The resulting shortest trees were then used as starting trees for a second round of heuristic searching, in order to sample more trees from those islands, using TBR branch swapping and limiting the number of trees saved to 10,000. Branch support was estimated using bootstrap analyses of 500 replicates with 'full' heuristic searches of 50 random addition sequences, TBR, saving 10 trees each time.

2.2.3. Bayesian analysis

The combined datasets were analysed using Bayesian inference, as implemented in MrBayes 3.12 (Huelsenbeck and Ronquist, 2001). In the analyses of ITS, 26S, combined non-coding, combined coding cpDNA and combined nrDNA (as well as preliminary analyses of combined cpDNA, and total combined DNA) the data were partitioned according to the separate markers used, with a separate partition for binary coded indels characters. In the final analyses of combined cpDNA and total combined DNA the data were partitioned instead according to non-coding versus gene regions and cpDNA versus nrDNA, i.e.; into three and five partitions respectively (including indels). In this way we aimed to reduce the complexity of the analyses, increasing the proportion of topology relative to substitution parameter change proposals in the MCMC chains, in the hope of achieving convergence faster. Both rates and substitution models were allowed to vary across the partitions. Priors for the number of parameters in the DNA substitution models were applied to each partition (as determined using ModelTest 3.06 (Posada and Crandall, 1998), with the topology in each case derived from a randomly selected most parsimonious tree). In each case this corresponded to models with NST = 6, gamma distributed rates and proportion of invariable sites. The single outgroup allowed by MrBayes was *Hordeum* (for analysis of coding cpDNA data) or *Centropodia glauca* (otherwise). Prior probabilities for all topologies were assumed to be equal.

Two independent MCMC analyses were set to run indefinitely with four simultaneous MCMC chains. One tree per 1000 generations was saved. We checked for convergence during the runs by comparing the mean log likelihoods (LnL; minus burnin) and estimated sample sizes (ESS) of parameters using TRACER (Rambaut and Drummond, 2003). Runs were considered to have converged when their post-burnin mean LnL were the same and the combined ESS for all parameters was >100. We attempted to accelerate convergence in the supermatrix analyses by (a) reducing the number of partitions, as above; and (b) providing starting trees for one or more of the four chains in each case as follows:

(1) All chains were started with a shortest tree from the parsimony search. (2) All chains were started with a shortest tree from the parsimony search, but that topology was randomly perturbed in each case using the "nperts" command in MrBayes (nperts = 10). (3) One chain was started with a shortest tree from the parsimony search, the other three were random. (4) One chain was started with a shortest tree from parsimony search and two further chains were started with topologies modified from that tree to place duplicated taxa according to (a) the cpDNA signal and (b) the nrDNA signal. The fourth chain was random.

Burn-in values were determined empirically from the likelihood values and 50% majority rule consensus trees together with approximations of the posterior probability (PP) for the observed bipartitions were calculated from the non-burnin trees pooled from the independent runs.

3. Results

3.1. Alignment, variability, and information content of the different markers

Dense sampling of non-coding markers made homology assessment straightforward in most parts of the alignment. The low length variability of coding markers made their alignment unambiguous both within the sparsely sampled ingroup and between the ingroup and more distant outgroups. Overall, coding regions were less variable than non-coding regions. The more variable of the coding regions (*matK* and *ndhF*) provided a similar proportion of informative characters (10.4% and 9.6% respectively) to non-coding

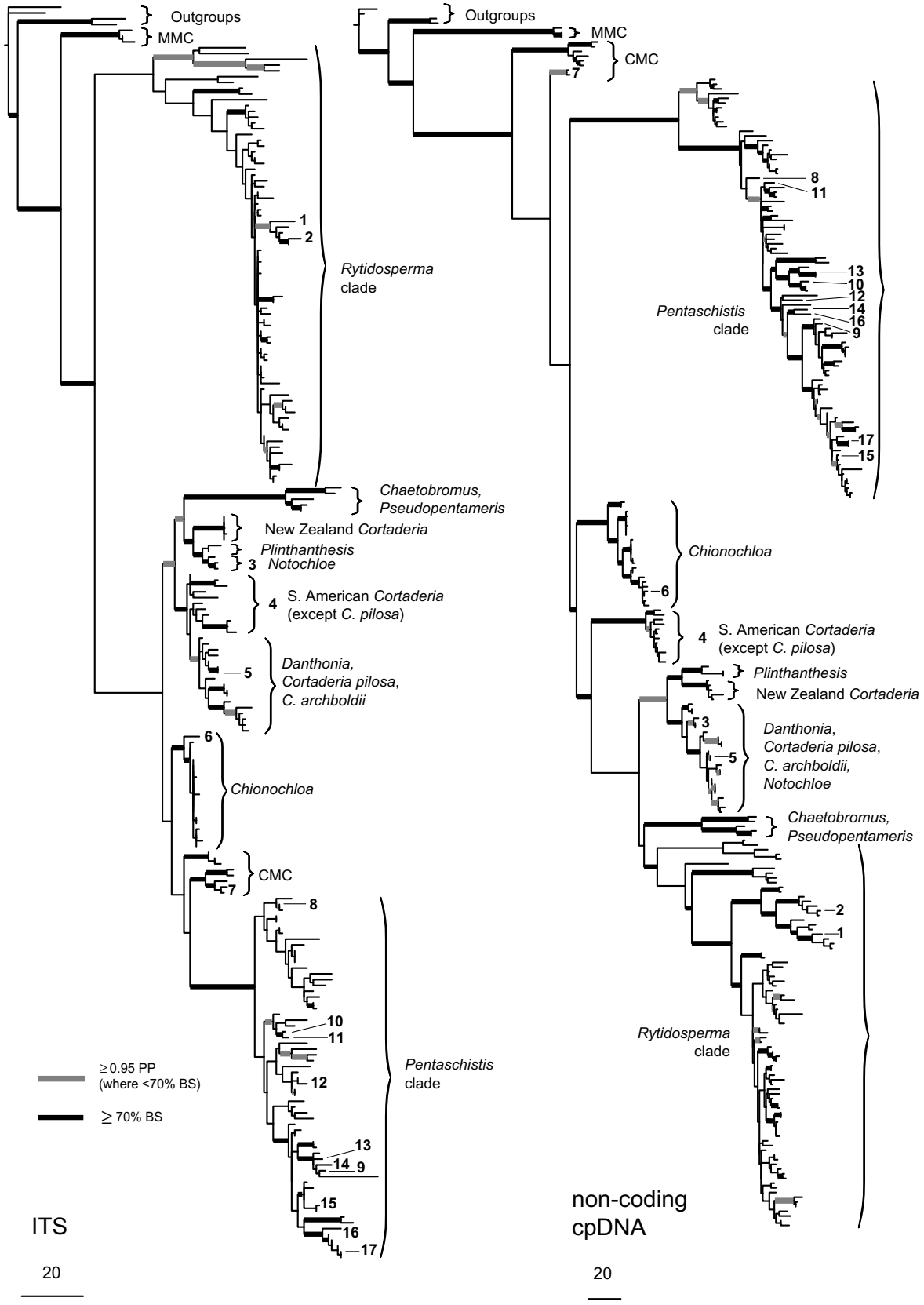

Fig. 1. Phylograms arbitrarily selected from the most parsimonious trees found in heuristic search of ITS and of combined non-coding cpDNA data. Nodes subject to bootstrap support ⩾70% are indicated by thicker branches, those subject to <70% BS, but ⩾0.95 Bayesian posterior probabilities, are indicated with thicker grey branches. For ease of presentation, taxon names have been removed. Clades and taxa as referred to in the text are indicated with brackets and named, and taxa with conflicting positions numbered as follows: (1) *Tribolium pusillum*; (2) *T. ciliare*; (3) *Notochloe microdon*; (4) *South American Cortaderia* (except *C. pilosa*); (5) *Danthonia alpina*; (6) *Chionochloa australis*; (7) *Merxmuellera arundinacea*; and (8) *Pentaschistis basutorum*; (9) *P. reflexa*; (10) *P. horrida*; (11) *P. malouinensis*; (12) *P. velutina*; (13) *P. aristidoides*; (14) *P. chippendalliae*; (15) *P. pictigluma var. mannii CG267*; (16) *P. densifolia*; and (17) *P. pseudopallescens*.

cpDNA markers (average 9.8%; Table 3) for the 35 taxa subset, but this was around half that of ITS (19.3%). It must be noted that the figures for the non-coding data result from an alignment which is based on the complete sampling: if the 35 taxon subset of sequences had been aligned without reference to the length variability of the entire sample, we would expect the quality of the homology assessment to suffer. 26S and *rbcL* were less variable, providing 6.5% and 5.5% parsimony informative characters, respectively, for the same taxon subset. When including all the sampled taxa, the numbers of informative characters for the non-coding cpDNA markers more than doubled, and informative indels increased almost fourfold (Tables 3 and 4). However, the ratio of informative characters to taxa (C/T; Table 4) declined from c. 9:1 to c. 3:1. The number of informative characters in ITS increased by little more than 50% with the full taxon sampling, and the ratio of informative characters to taxa decreased from 3.5:1 to 1:1. This much lower ratio is likely due to the short length of ITS sequences.

3.2. Parsimony analysis

Analysis of the individual cpDNA markers resulted in topologies without significantly supported incongruence (⩾70% BS). These were therefore combined in three datasets: 'combined non-coding'; 'combined coding' and 'combined cpDNA' (including all taxa: a large proportion with *matK, ndhF* and *rbcL* coded as missing data; clade support represented in Fig. 3). Combining the cpDNA data resulted in an increase in the ratio of informative characters to taxa (i.e., terminals) from 1:1 for the non-coding cpDNA regions individually to c. 3:1 when these were combined, and further to c. 6:1 with all cpDNA data (Table 4). This was reflected in improved resolution and higher support values. The increase resulting from including the coding regions was gained with a relatively small sequencing effort. Different numbers of sequences were required to span the full extent of the different markers, but on average one coding cpDNA sequence yielded 0.84 informative characters, compared with 0.87 for 26S; 0.57 for non-coding cpDNA; and 0.49 for ITS (C/seq; Table 4).

The combined non-coding cpDNA data (Fig. 1) provided support for major clades, plus some resolution between closely related species, whilst the combined coding cpDNA data (Fig. 2) provided support for relationships between those major clades. Most clades recovered with support in the non-coding and coding analyses individually were also recovered when the two matrices were combined, albeit sometimes with reduced support. Clade support (both BS and PP from Bayesian analysis; see below) for 10 selected 'spine' nodes, A–I (Fig. 3), according to different combinations of the data, is presented in Table 5. All nodes supported by coding cpDNA alone were supported also when combined with the non-coding regions, although of the 10 nodes detailed in Table 5, support for four decreased by ⩾10% BS.

The nuclear markers ITS and 26S also showed no supported incongruence, although the numbers of informative characters, and thus degree of resolution, were lower than for the cpDNA analyses (Figs. 1 and 2). Combining the two markers increased the ratio of informative characters to taxa only marginally, from 1.0:1 (ITS) to 1.3:1 (ITS plus 26S; Table 4). Combined analysis resulted in a topology congruent with those of the individual analyses, but without significant increase in support values. Clade support from combined nrDNA analyses is shown in Fig 3.

The number of informative characters in the total combined DNA supermatrix was the sum of the cpDNA (1457) and nrDNA (287) supermatrices, which improved the ratio of informative characters to taxa greatly with respect to the nrDNA dataset (from c. 1:1 to 6:1), but much less so with respect to the cpDNA (5.6:1–6.0:1). Combination of cpDNA and nrDNA resulted in a topology congruent with both combined cpDNA and nrDNA analyses (when only nodes with support ⩾70% are considered). It included strong support for the alternative positions of most of the conflicting taxa and greatly increased overall resolution and support with respect to the nrDNA tree. Support values decreased overall with respect to the cpDNA tree. Node support according to the different combined analyses is presented in Fig. 3. We compared bootstrap support values for nodes recovered in the total combined DNA analysis with the higher of the two support values for those nodes according to cpDNA and nrDNA individually (treating all values <50% the same): For nine nodes support increased by ⩾10%, whilst for 29 nodes it decreased by ⩾10%, on combination of the nrDNA and cpDNA data. However, only 1 of 10 spine nodes detailed in Table 5 decreased by ⩾10% BS in the total combined DNA analysis compared to the combined cpDNA analysis. Twelve nodes with support ⩾70% in one or other of the individual analyses were not recovered in the total combined analysis: three in the *Rytidosperma* clade (combined cpDNA), eight in the *Pentaschistis* clade (four each in the combined nrDNA and cpDNA analyses), and the crown node of *Plinthanthesis* (cpDNA). Seven of these 12 collapsed nodes subtended duplicated taxa. In the total combined analysis, no novel clades were recovered with support ⩾70%.

3.3. Bayesian analysis

Bayesian analysis of coding cpDNA, 26S and ITS matrices each converged within two runs of 5 million generations each, with bur-

Table 4
Numbers and proportions of informative characters for the markers used, and numbers of informative characters per taxon and per sequence added, based the full taxon sampling for each marker (excluding more distant outgroups available only for coding cpDNA markers)

	No. taxa (T) (no. of seqs)	Sequence length, alignment (entire/included)	Parsimony informative bases	Pars. inf. indels	Total inf. chars. (C)	C/T	C/seq
trnL-F	257 (514)	c. 950, (1334/1232)	219 (23.1%)	65	284	1.1	0.55
rpl16	229 (458)	c. 900, (1378/1312)	206 (22.8%)	78	284	1.2	0.62
atpB-rbcL	218 (436)	c. 950, (1174/1076)	184 (19.4%)	49	233	1.1	0.55
Non-coding cpDNA	260 (1408)	c. 2800	609 (21.8%)	192	801	3.1	0.57
ndhF	56 (224)	2064	257 (12.5%)	4	261	4.7	1.17
matK	57 (342)	1893	275 (14.5%)	24	299	5.3	0.87
rbcL	53 (212)	1338	96 (7.2%)	0	96	1.8	0.45
Coding cpDNA	57 (778)	5295	628 (11.9%)	28	656	11.5	0.84
All cpDNA	260 (2186)	c. 8095	1237 (15.3%)	220	1457	5.6	0.67
ITS	217 (434)	c. 700 (757/639)	213 (33.3%)	0	213	1.0	0.49
26S	35 (70)	1138	74 (6.5%)	0	61	1.7	0.87
Nuclear	216 (504)	c. 1838	287 (15.6%)	0	287	1.3	0.57
Grand total	290 (2690)	c. 9933	1524 (15.3%)	220	1744	6.0	0.65

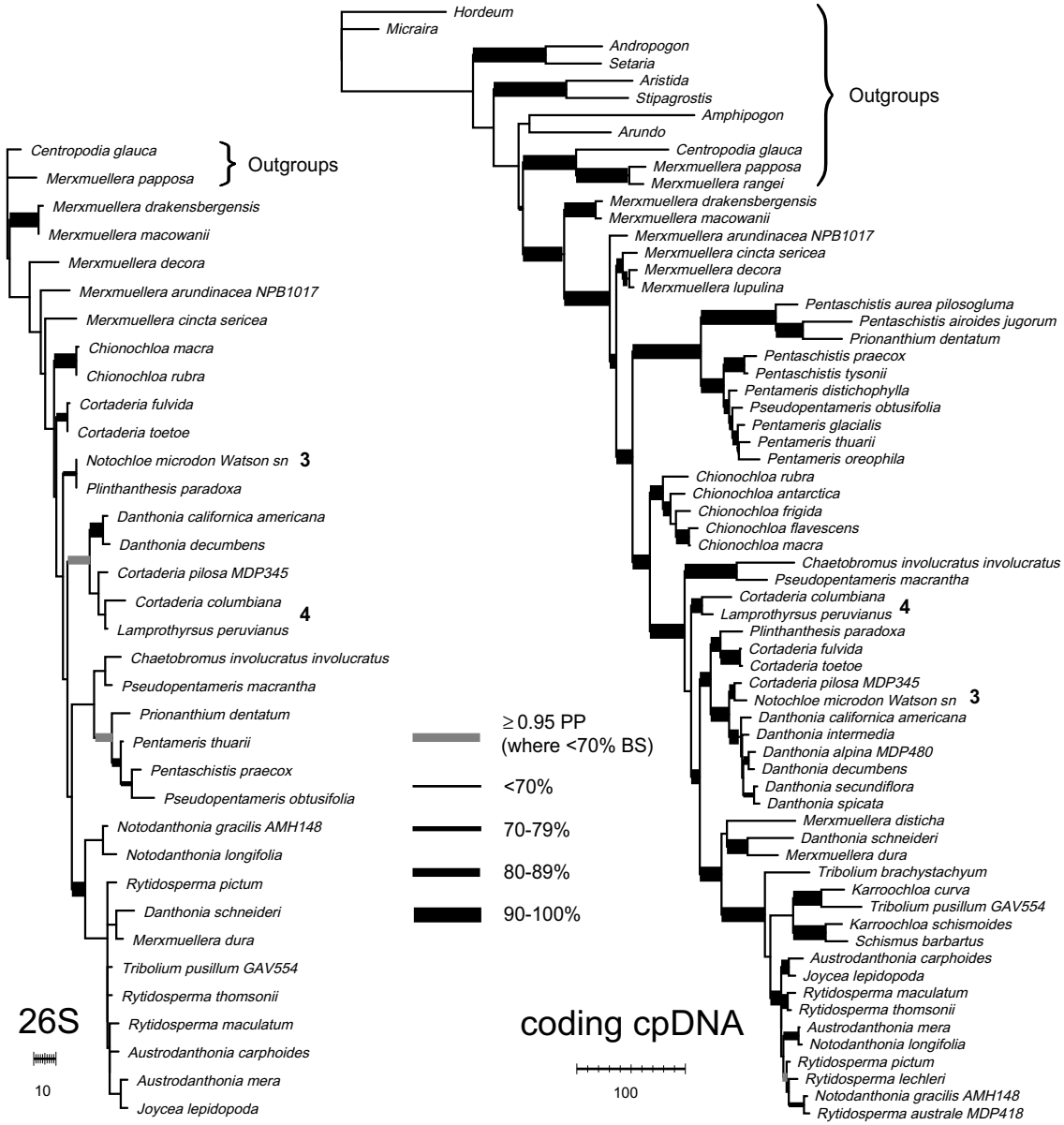

Fig. 2. Phylograms arbitrarily selected from the most parsimonious trees found in heuristic search of 26S and of combined coding cpDNA (excluding taxa for which only rbcL was available). Bootstrap support is indicated by the thickness of the branches, summarised in categories: <70% (unsupported); 70–79% (weak support); 80–89% (moderate support); 90–100% (strong support). Nodes subject to <70% BS, but ⩾0.95 Bayesian posterior probabilities, are indicated with grey branches. Note the conflicting positions of South American Cortaderia (including Lamprothyrsus; clade (4) and Notochloe microdon (3).

nin periods of between 50,000 and 500,000 generations. We report the ESS values for the total tree length parameter (TL), as these were consistently the lowest. The ESS of TL for coding cpDNA was 161; for 26S: 192; and for ITS: 115. Analyses of non-coding cpDNA (ESS TL: 132) and combined nrDNA (ESS TL: 102) converged within two runs of 10 million generations each, with burnin periods of 1 million generations.

Independent analyses of the combined cpDNA and total combined DNA data partitioned according to the markers (i.e., seven and nine partitions respectively) failed to converge, despite running for over 20 million generations over a number of weeks. Further analyses employed the simplified data partitioning and one or more parsimony derived starting trees, as described above. Five analyses of combined cpDNA all reached the same LnL plateau after between 50,000 and 3,500,000 generations. These runs were allowed to continue for between 5.6 and 10 million generations each (35.6 million generations in total), at which point the combined output yielded ESS values >100 for all parameters (TL: 126). The 30,900 post-burnin trees were then pooled.

Analyses of the total combined data did not all reach the same LnL plateau. After burnin periods of 2.4 and 1.6 million generations respectively, the mean LnL calculated for runs 1 and 2 (started with unperturbed and perturbed parsimony trees, respectively) were both −56,730. Runs 3 and 4 (1 parsimony and three random starting trees; 1 parsimony, two modified parsimony and one random tree, respectively) reached different suboptimal LnL plateaus after apparent burnin periods (run 3: c. 10.5 million generations, burnin 3.5 million generations, LnL: −56,770; run 4: c. 8.4 million generations, burnin 600,000 generations, LnL: −56,760). Inspection of the trees sampled by these runs revealed inconsistent placement of duplicated taxa. Runs 1 and 2 were allowed to continue for c. 18 million generations each at which point ESS values for all parameters were >100 (TL: 124). A total of c. 30,300 post-burnin trees were pooled.

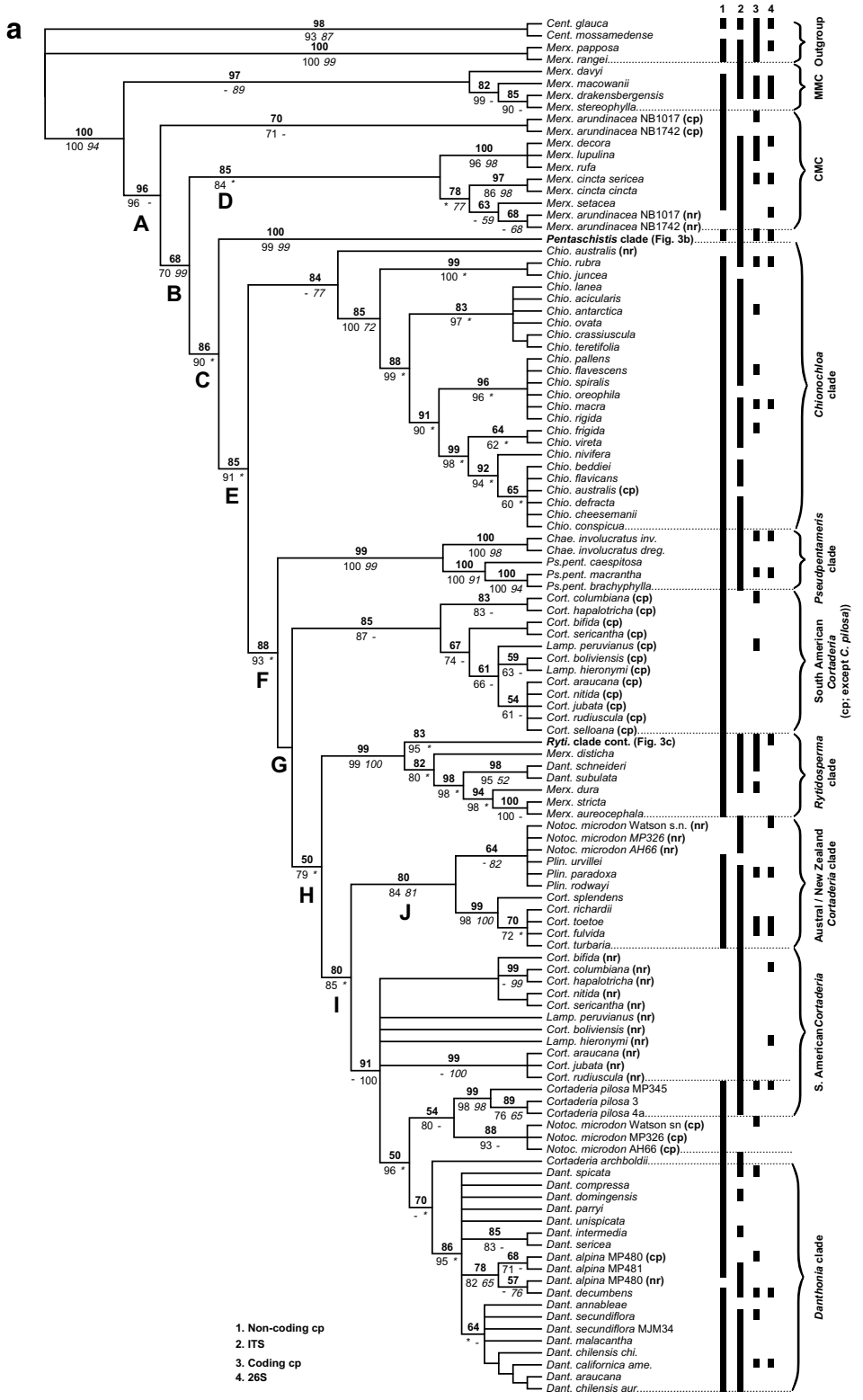

Fig. 3. a–c. Strict consensus of most parsimonious trees found in heuristic search of the total combined DNA sequence data. Bootstrap support (BS) values for supermatrix analyses are indicated above the branches (total combined data in bold) and below (combined cpDNA in normal font; combined nrDNA in italics). Where nodes are not represented in a particular partition (i.e. due to missing data or conflict) these are denoted by "-"; where BS is less than 50%, they are represented by "*". Where support values are not indicated, those nodes received less than 50% BS in all analyses. nrDNA and cpDNA partitions (where treated separately due to conflict) are represented by '(nr)' and '(cp)' respectively. Selected 'spine' nodes as referred to in the text are labelled A–K. The presence of data partitions (non-coding cpDNA; ITS; coding cpDNA; 26S) per taxon, is represented by the black vertical bars. (a) Danthonioideae (Cent. = Centropodia; Merx. = Merxmuellera; Chio. = Chionochloa; Chae. = Chaetobromus; Ps.pent. = Pseudopentameris; Cort. = Cortaderia; Lamp. = Lamprothyrsus; Ryti. = Rytidosperma; Dant. = Danthonia; Notoc. = Notochloe; Plin. = Plinthanthesis). Clades referred to in the text are indicated with brackets and named following Barker et al. (2007). (b) and(c); (b) The Pentaschistis clade (Pentas. = Pentaschistis; Pentam. = Pentameris; Prion. = Prionanthium; Ps.pent. = Pseudopentameris). (c) The Rytidosperma clade continued (Karr. = Karroochloa; Trib. = Tribolium; Schi. = Schismus; Aust. = Austrodanthonia; Ryti. = Rytidosperma; Joyc. = Joycea; Notod. = Notodanthonia).

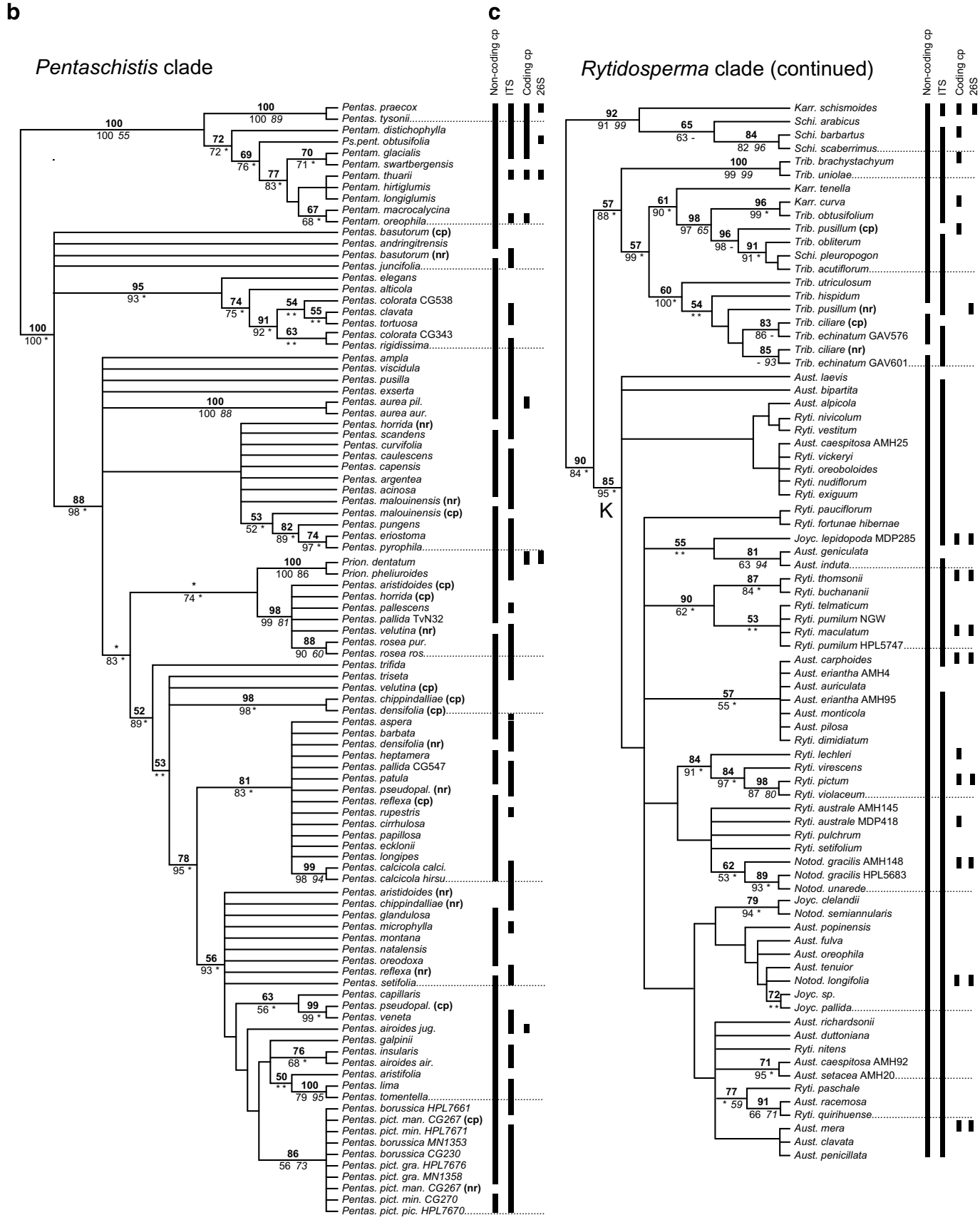

Fig 3. (*continued*)

Results were congruent with parsimony analysis of the same data (where only nodes supported by ⩾70% BS and 0.95 PP are taken into account), but Bayesian topologies were slightly more resolved. In contrast to the parsimony supermatrix analyses, in

Table 5
Bootstrap support (left) and posterior probabilities (PP; right) for ten 'spine' nodes (Fig. 3a) according to non-coding and coding cp and nr DNA separately and combined in supermatrix analyses

Data/Node	A	B	C	D	E	F	G	H	I	J
Non-coding cpDNA	96/1.00	<50/<0.50	64/0.84	79/1.00	81/1.00	91/1.00	<50/*0.90	<50/0.99	67/0.98	79/0.99
Coding cpDNA	100/1.00	93/1.00	99/1.00	91/1.00	97/1.00	100/1.00	65/0.77	97/1.00	97/1.00	99/1.00
Combined cpDNA	96/1.00	70/1.00	90/1.00	84/1.00	91/1.00	93/1.00	<50/*0.80	79/1.00	85/1.00	84/1.00
ITS	–	95/1.00	<50/<0.50	<50/<0.50	<50/<0.50	<50/<0.50	<50/<0.50	<50/<0.50	<50/<0.50	88/0.99
26S	–	<50/0.89	55/0.59	<50/<0.50	<50/<0.50	<50/<0.50	<50/<0.50	<50/<0.50	<50/<0.50	<50/0.86
ITS/26S	–	99/1.00	52/0.87	<50/<0.50	<50/<0.50	<50/<0.50	<50/<0.50	<50/<0.50	<50/<0.50	81/1.00
Total combined	96/1.00	68/1.00	86/1.00	85/1.00	85/1.00	88/1.00	<50/*0.85	50/1.00	80/1.00	80/1.00

Values <50% are not reported. Node 1 corresponds to a lineage represented here only by cpDNA data, and is therefore not reported for nrDNA. PP support for an alternative resolution of node G (following Fig. 4) is indicated with an asterisk.

which BS for spine nodes declined on combination of the coding/non-coding and cpDNA/nrDNA datasets, Bayesian PP for these clades remained high. Nodes supported by $\geqslant$0.95 PP, but <70% BS (separate non-coding and coding cpDNA, ITS and 26S) are indicated in Figs. 1 and 2, detailed results for the total combined DNA analyses are presented in Fig 4, and PP values for spine nodes according to each of the data combinations are summarised in Table 5.

4. Discussion

4.1. Advantages of the supermatrix approach: efficient sampling strategies

Wiens et al. (2005) and Wiens (2006) argue for the use of different sampling strategies to resolve relationships at different levels in a phylogeny. In 'top down' and 'bottom up' approaches they sampled many taxa for a single fast evolving mitochondrial gene region, and slower evolving nuclear encoded markers and morphology for fewer taxa (Wiens et al., 2005). This sampling strategy is particularly appropriate in study groups where the level of character sampling necessary to achieve adequate resolution differs between clades. All things being equal, more sequence data is necessary to infer the phylogeny of a rapid radiation of taxa than that of a group with longer time-lapses between speciation events. By limiting extra character sampling to exemplar taxa, for example, for already sufficiently resolved clades, sequencing effort can be concentrated on other areas of the topology. The benefit of such a strategy is most apparent in cases where different areas of the topology are best addressed with different kinds of data.

Radiations can be recent in origin, with low levels of variation between species even for the most variable of markers. They can also be more ancient, with as a result higher levels of inter-species variation, but with comparably low proportions of that variation representing informative synapomorphies. Different kinds of sequence data may be more appropriate for phylogeny reconstruction given these different scenarios: rapidly evolving non-coding markers may contain the most information for recently evolved groups, but can present a number of problems when comparing more distantly related taxa. Patterns of insertions, deletions and repetitive motifs, which between more closely related taxa can be highly informative (as for example found in the grass-specific insert in the cpDNA RPCO2 region; Barker et al., 1999), reduce the proportion of the sequence data which can be directly compared between more divergent lineages and pose problems for alignment and coding of gap characters. Markers with strongly imposed structural constraints, such as protein coding genes, may be less variable, but still a better tool in such cases.

This distinction is well illustrated in the Danthonioideae phylogenetic reconstruction presented here. Rapidly evolving non-coding markers recovered a number of major clades, but failed to resolve relationships between some of those clades (i.e. the spine of the tree) and between some closely related species (see Fig. 1). To improve resolution between closely related species, it would seem sensible to sample further maximally variable markers (Hughes et al., 2006), with dense taxon sampling, as has been done in a study of the *Pentaschistis* clade (Galley and Linder, 2007). Resolving the spine of the tree involves inferring relationships between well supported clades, which in principle can be represented by a relatively small number of individual taxa. The results presented in Fig. 1 and Table 3 show that in our Danthonioideae dataset, even the non-coding data is informative at this level, and in fact yields around 50% more information per sequence generated than protein coding data. However, the dense taxon sampling necessary to obtain a satisfactory alignment greatly inflates the number of sequences required to obtain this result (Table 4). By sequencing easy-to-align protein coding genes for a relatively small number of taxa representing the major clades, and combining this with the densely taxon-sampled non-coding matrix in a supermatrix, we resolved most of the spine relationships (see Fig. 3). This targeted, 'top down' and 'bottom up' sampling strategy (Wiens et al., 2005) required only a fraction of the sequencing effort that would have been necessary for comprehensive taxon coverage across the board.

4.2. Inclusion of taxa with missing or conflicting data

Taxa can be included in a supermatrix analysis even if much of the corresponding data is unavailable (Wiens, 2003; Wiens et al., 2005). This not only provides a means to improve efficiency of sampling, as above. It can be important, for example, when sequences are obtained from low quality DNA sources, as was the case for a number of rare and inaccessible taxa sampled from herbarium specimens for this study (e.g. *Cortaderia archboldii*). It is also important where reticulations in the phylogeny mean that gene sequences with a given signal are no longer present in certain taxa, having been replaced by transfer from a different lineage.

In general, we would prefer to analyse all of the available data in a single analysis, as the interaction between data partitions in a combined analysis opens the possibility of revealing hidden support for clades not recovered by the partitions separately (Nixon and Carpenter, 1996; Sanderson et al., 1998). This benefit is lost if a consensus or 'supertree' approach is adopted (Sanderson et al., 1998; Pisani and Wilkinson, 2002). Combined analysis, by any method, cannot be justified in cases, such as in the Danthonioideae, where different classes of evidence (i.e., cpDNA and nrDNA sequences) exhibit conflicting signal (de Queiroz et al., 1995). The means of addressing this problem have to date been limited to either separate analysis of conflicting data partitions, or the exclusion of taxa or data exhibiting conflicting signal. Reticulate patterns, such as the placement of hybrid taxa, have been inferred after phylogenetic analysis either by hand (e.g., Kellogg et al., 1996) or by summarising the separate gene trees using one of a number of network techniques (reviewed in Vriesendorp and Bakker, 2005). This approach is limited when levels of resolution in

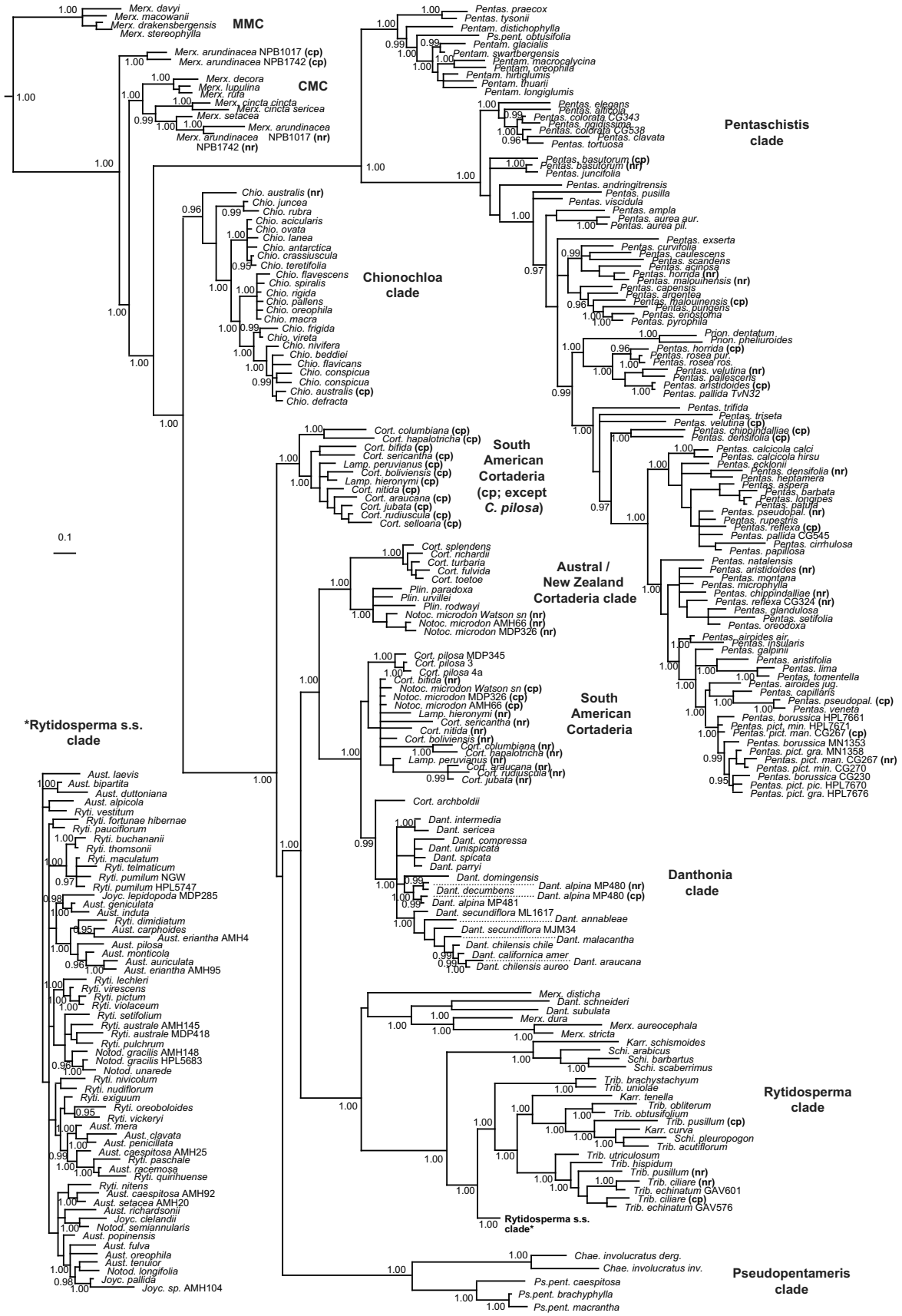

Fig. 4. Bayesian 50% majority rule consensus tree of the total combined data. Posterior probabilities ⩾0.95 are represented next to the nodes. Taxon name abbreviations follow those of Fig. 3.

individual gene trees are relatively low, as is the case with the nrDNA tree presented here. The taxon duplication strategy we have used circumvents this limitation, allowing simultaneous analysis of the non-conflicting elements of both data partitions, whilst still including the information provided by gene-tree conflict. The result is a single (sample of) bifurcating phylogenetic tree(s) which represents all of the available data.

4.3. Drawbacks of the supermatrix approach: missing data and parsimony bootstrap support

Node support values, typically bootstrap support in parsimony analyses, are not perfectly comparable across different datasets (Felsenstein, 2004). The results presented here may suggest that such difficulties are aggravated under a supermatrix approach. The importance of the ratio of informative characters to taxa has been demonstrated for 'normal' datasets (Bremer et al., 1999), and appears to be reflected in the results of combining non-coding and coding cpDNA data here: the coding data has the highest proportion of characters to taxa, and also the highest node support. In contrast combining the cpDNA and nrDNA markers resulted in a small *increase* in the proportion of informative characters to taxa (from 5.6:1 to 6.0:1), but an overall marked *decrease* in bootstrap support values (Fig. 3). A similar decrease in support was recorded by Bouchenak-Khelladi et al. (2008) in their large multi-gene phylogenetic trees of the entire grass family generated using a supermatrix approach, in comparison to trees generated from matrices with perfect parallel sampling. Further compounding factors may confuse interpretation of these results: these include the difference between targeted and arbitrarily distributed missing data and the possibility of unsupported incongruence between the partitions.

Missing data was introduced into the supermatrix in two very different ways: for the coding versus non-coding matrices, extra data was introduced for deliberately targeted taxa (i.e., bracketing the basal node [where this was known] of supported clades). Support for the relationships between such clades was fairly robust to the inclusion of further taxa.

In contrast, both failure to obtain data for difficult samples, and the taxon duplication process for conflicting taxa, resulted in a more arbitrary distribution of missing data. In the latter case, that distribution was imposed by (a) the conflict discovered between the partitions, and (b) the proportions of cpDNA and nrDNA in the dataset. Missing data were therefore distributed randomly among the taxa, and divided unequally between the taxa "created" by the taxon duplication process (83% and 17% informative characters depending on whether the cp- or nrDNA was represented, respectively). Although the overall proportion of missing data may in itself be unimportant (Wiens et al., 2005), the effect of a large proportion of missing data becomes important when characters necessary to place particular taxa in the phylogeny are missing. For example, strongly supported resolution between *Danthonia*, *Cortaderia pilosa* and the cpDNA duplicate of *Notochloe microdon* in the cpDNA tree is lost on the inclusion of nrDNA sequences of South American *Cortaderia* in the combined tree. Wilkinson (1995) argued for eliminating disruptive taxa where they have no effect on the relationships inferred for other taxa in a process termed 'safe taxonomic reduction'. We have included all the taxa in the combined analysis in order to retain the information pertinent to those taxa. The loss of support could be interpreted as a meaningful reflection of relationships when the different gene trees are taken into account. Support values in the absence of these taxa can in any case be interpreted from the individual combined cpDNA and nrDNA analyses (Fig. 3).

Alternatively, it might be argued that support should be interpreted from the posterior probabilities obtained from Bayesian analyses instead of parsimony bootstrap proportions when using a supermatrix approach. Posterior probabilities have been shown in previous studies to be more robust to missing data (Wiens et al., 2005; Bouchenak-Khelladi et al., 2008). Our results show universally high posterior probabilities for a number of nodes which were subject to high BS in the analysis of combined cpDNA, but for which BS declined markedly on combination of cpDNA and nrDNA partitions. The challenges to achieving convergence under Bayesian inference when using a supermatrix approach (discussed below) represent one possible caveat to this interpretation. A further caveat is provided by possible over-credibility of Bayesian PP as a measure of clade support (e.g., Suzuki et al., 2002). In general, it would seem sensible to regard nodes recovered under Bayesian inference but not supported by parsimony analysis with caution, irrespective of the proportion of missing data in the matrix. However, in this case the spine nodes subject to high PP in the supermatrix analyses are also supported by parsimony analysis of the coding cpDNA data alone.

4.4. Unidentified conflict

Gene trees that combine limited resolution with clear incidences of conflict present a related, but very different problem. If we take the case of the *Pentaschistis* crown group (Fig. 3b), neither the cpDNA data analysed here nor ITS gives a fully resolved tree. Of the supported nodes, a relatively high proportion reveal conflict, which may suggest that further (unsupported) incongruence has been overlooked. The overall decrease in support values, plus the failure to recover all of the nodes which received significant support according to either partition independently in the combined analysis, might be seen as evidence to support this suspicion. However, the majority of these nodes subtended one or more duplicated taxa. Duplicated taxa both decrease the proportion of characters to taxa for the clade to which they belong, and may be represented by insufficient data to be placed with confidence when compared with closely related taxa. Pirie et al. (unpublished manuscript) performed combined analyses on a subset of the data used here, having excluded conflicting taxa. They reported overall increased support values, providing evidence for the fundamental congruence between the markers. Similar analyses performed on just the *Pentaschistis* clade did not show such an increase (results not shown). We may therefore be forced to conclude that we cannot combine the data with confidence for this group.

4.5. Challenges for Bayesian inference

Extensive missing data can impact parameters that are based on summations of all characters, such as branch lengths and nucleotide composition biases (Gatesy et al., 2002). This is not the case with parsimony, in which missing data for a character only affects the influence of that character. Some authors have nevertheless reported that Bayesian inference methods were superior to parsimony in the presence of missing data (Flynn et al., 2005; Wiens et al., 2005). However, Wiens et al. (2005) also report that MrBayes analyses of their supermatrix were often very slow in reaching convergence. This effect was also observed here. In particular, we observed that multiple runs including duplicated taxa (with high proportions of missing data) were slow or failed to arrive at the same LnL plateau. In each case the difference in mean LnL was accompanied by inconsistent placement of duplicated taxa in the 50% majority rule consensus.

Reducing the number of partitions in our supermatrix analyses allowed faster exploration of tree space, but it still took a prohibitively long time for independent analyses starting from random starting trees to converge to the same mean LnL. We achieved convergence in these analyses only after dictating the topology of one

or more of the starting trees, and thereby starting the analyses in more probable regions of tree/model space. Such a strategy, although apparently effective, could potentially be at the cost of the strict independence of the runs. Better alternative models may not be sampled because they are too distant from the starting tree. Where thorough exploration of tree space is not possible, we do not know whether perturbing the user defined starting trees and including random starting topologies for some chains may solve this problem. The situation is clearer when there are competing phylogenetic hypotheses to be tested, such as the conflicting elements of the cpDNA and nrDNA gene trees presented here, or for example when the results of parsimony analysis are suspected to have been affected by long-branch attraction. In such cases competing hypotheses can themselves be represented as starting trees, and the results compared. Given these caveats, the strategy appears to have been successful: in both cpDNA and total combined DNA analyses multiple runs reached the same LnL plateau and adequately sampled all parameters of the model within a reasonable time.

4.6. Phylogeny of Danthonioideae

The combination of high representation of taxa, generally robust clade support, and the appropriate treatment of inter-locus conflict in the analyses presented here has resulted in a considerable improvement in hypotheses of phylogenetic relationships between species of Danthonioideae.

The rooting of previous studies was confirmed by analysis of coding cpDNA sequences representing the subfamilies of the PACCAD clade (Fig. 2). Our results support those of Bouchenak-Khelladi et al. (2008), suggesting a sister group relationship between Danthonioideae and Chloridoideae, rather than between Danthonioideae and Aristidoideae, as hypothesised by the Grass Phylogeny Working Group (GPWG, 2001).

Within Danthonioideae, major clades identified in earlier studies (Barker et al., 2007; indicated in Fig. 3a) were recovered here, with some differences in clade membership resulting from resolution of inter-locus conflict. The basal *Merxmuellera* assemblage (sensu Barker et al., 2007) comprises three, rather than two clades (see Fig. 3 a): the Mountain *Merxmuellera* clade (MMC; sensu Barker et al., 2007); the chloroplast genome of *M. arundinacea*; and the Cape *Merxmuellera* clade (CMC sensu Barker et al., 2007; excluding the chloroplast genome of *M. arundinacea*). Barker et al. (2007) indicated conflict between cpDNA and ITS data partitions involving the *Danthonia*, Austral/New Zealand *Cortaderia*, and South American *Cortaderia* clades. We can now recognise that both *Notochloe* and all South American *Cortaderia* species sampled here (with the exception of *C. pilosa*), have distinct chloroplast and nrDNA evolutionary histories. The nrDNA lineages of South American *Cortaderia* are included in a clade comprising the chloroplast genome of Australian *Notochloe microdon* and both data partitions of the largely New World *Danthonia* clade, New Guinean *Cortaderia archboldii*, and *Cortaderia pilosa*. It is interesting to note that *C. pilosa* is morphologically similar to, and overlaps in geographical distribution with, the rest of South American *Cortaderia*.

Resolution within a number of clades, such as the *Danthonia* clade and New Zealand *Cortaderia*, was limited. In particular, further work will be needed to address the uncertainty in the phylogeny of the substantial *Rytidosperma* clade. Variation in the DNA sequence markers used here is low within *Rytidosperma*, resulting in less resolution than in the *Pentaschistis* clade, based on the same data. The distribution pattern of the *Rytidosperma* clade (across all continents of the southern hemisphere), its apparent recent origin, and the complexity of its phylogenetic relationships makes it an ideal subject with which to further address the history of dispersal across the Southern Hemisphere. A robust phylogeny of this group may require a large input of data, and this work is ongoing (AMH).

With the exception of *Chionochloa*, *Plinthanthesis* and *Prionanthium*, all currently defined non-monotypic Danthonioideae genera are para- or polyphyletic. In some cases, the problem reflects the position of just a few taxa and is consistent between gene trees. For example, two of the 64 sampled species of *Pentaschistis* (*P. tysonii* and *P. praecox*; Fig. 3b) render *Pentaschistis* paraphyletic with respect to *Pentameris*. *Pseudopentameris obtusifolia* was originally described as a species of *Pentameris*, and its treatment as such would leave both *Pentameris* and *Pseudopentameris* monophyletic. Other genera, such as *Merxmuellera* and *Cortaderia*, as previously demonstrated (Barker et al., 2003, 2007), are grossly polyphyletic. Furthermore, as a result of their apparently reticulate evolutionary history, some groups of species are monophyletic according to one dataset, but not according to the other. In this light, it is perhaps unsurprising that patterns of morphological variation, when regarded exclusively, are difficult to interpret in Danthonioideae. The representation of taxa in this study can however give us greater confidence that morphological characters diagnostic for clades will correctly predict the placement of the relatively small number of species that were not sampled. A formal taxonomic treatment addressing these issues and including detailed discussion of the morphological characters which can be used to identify species to monophyletic genera is in preparation (Linder et al., in preparation).

5. Conclusions

The results presented here demonstrate the utility of a 'top down, bottom up' approach (Wiens, 2006) to phylogeny reconstruction in a species-rich clade. Our densely taxon-sampled noncoding matrices led to the identification of a number of major clades and resolution of some relationships within those clades. These were augmented by smaller coding matrices which provided strong support for the relationships between those clades. We echo the opinion of Wiens (2006) that in determining sampling strategies for molecular phylogenetic reconstruction the choice should not be between more genes or more taxa. Both more genes and more taxa can be sampled, without a prohibitively large sequencing effort, when character sampling is correctly targeted. Our approach differs more significantly in that it addresses the issue of conflicting gene trees, which is of particular relevance to inferences of evolutionary history and to classification in Danthonioideae. The conflicting taxa duplication approach allows data from conflicting gene trees to be combined. This is useful when data of different partitions are largely compatible, but the data available for one (often nrDNA in molecular phylogenetic studies of plant groups), or both data partitions, are insufficient to provide a fully resolved tree. The result of this work is a phylogeny of the Danthonioideae which combined both dense taxon representation and robust clade support. This will be useful for a wide range of evolutionary studies and to inform forthcoming realignment of generic delimitations in the subfamily.

Acknowledgments

The project under which a major part of this research was conducted was funded by a Swiss National Science Foundation grant to HPL (3100A0–107927), and additional support for fieldwork provided by a Swiss Academy of Sciences travel grant (M.D.P. and A.M.H.). Neville Walsh, Surrey Jacobs, Keith McDougall, Henry Connor and colleagues in herbaria CANB, VIC, SA and TAS in Australia and CHR in New Zealand gave invaluable assistance during field work of M.D.P. and A.M.H. We also gratefully

acknowledge the government agencies of those countries for providing research and collection permits. Mirjam Marti, Miloš Tatarski and Melanie Ranft generated a number of DNA sequences. Additional samples were kindly provided by Alexandre Antonelli, Osvaldo Morrone and Jeanette Keeling. The manuscript was improved following the useful comments of two anonymous reviewers.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2008.05.030.

References

Bakker, F.T., Hellbrügge, D., Culham, A., Gibby, M., 1998. Phylogenetic relationships within *Pelargonium* sect. *Peristera* (*Geraniaceae*), inferred from nrDNA and cpDNA sequence comparisons. Plant Sys. Evol. 211, 273–287.

Barker, N.P., 1995. A systematic study of the genus *Pseudopentameris* (Arundinoideae: Poaceae). Bothalia 25, 141–148.

Barker, N.P., Galley, C., Verboom, G.A., Mafa, P., Gilbert, M., Linder, H.P., 2007. The phylogeny of the austral grass subfamily Danthonioideae: evidence from multiple data sets. Plant Syst. Evol. 264, 135–156.

Barker, N.P., Linder, H.P., Harley, E.H., 1995. Polyphyly of Arundinoideae (Poaceae): evidence from *rbc*L sequence data. Syst. Bot. 20, 423–435.

Barker, N.P., Linder, H.P., Harley, E.H., 1999. Sequences of the grass-specific insert in the chloroplast rpoC2 gene elucidate generic relationships of the Arundinoideae (Poaceae). Syst. Bot. 23, 327–350.

Barker, N.P., Linder, H.P., Morton, C.M., Lyle, M., 2003. The paraphyly of *Cortaderia* (Danthonioideae; Poaceae): evidence from morphology and chloroplast and nuclear DNA sequence data. Ann. Missouri Bot. Gard. 90, 1–24.

Barker, N.P., Morton, C.M., Linder, H.P., 2000. The Danthonieae: generic composition and relationships. In: Jacobs, S.W.L., Everett, J. (Eds.), Grasses: Systematics and Evolution. CSIRO, Melbourne, pp. 221–230.

Baum, D.A., Small, R.L., Wendel, J.F., 1998. Biogeography and floral evolution of Baobabs (*Adansonia*, Bombacaceae) as inferred from multiple data sets. Syst. Biol. 47, 181–207.

Bouchenak-Khelladi, Y., Salamin, N., Savolainen, V., Forest, F., Bank, M.V.D., Chase, M.W., Hodkinson, T.R., 2008. Large multi-gene phylogenetic trees of the grasses (Poaceae): Progress towards complete tribal and generic level sampling. Molec. Phylog. Evol. 47, 488–505.

Bremer, B., Jansen, R.K., Oxelman, B., Backlund, M., Lantz, H., Kim, K.-J., 1999. More characters or more taxa for a robust phylogeny — case study from the coffee family. Syst. Biol. 48, 413–435.

Bull, J.J., Huelsenbeck, J.P., Cunningham, C.W., Swofford, D.L., Waddell, P.J., 1993. Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42, 384–397.

Conert, H.J., 1970. *Merxmuellera*, eine neue Gattung der Gramineen. Senckenbergiana Biol. 51, 129–133.

de Queiroz, A., Donoghue, M.J., Kim, J., 1995. Separate versus combined analysis of phylogenetic evidence. Annu. Rev. Ecol. Syst. 26, 657–681.

Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates Inc., Sunderland, MA.

Fitch, W.M., 1971. Toward defining the course of evolution: minimum change for a specified tree topology. Syst. Zool. 20, 406–416.

Flynn, J., Finarelli, J., Zehr, S., Hsu, J., Nedbal, M., 2005. Molecular phylogeny of the Carnivora (Mammalia): assessing the impact of increased sampling on resolving enigmatic relationships. Syst. Biol. 54, 317.

Galley, C., Linder, H.P., 2007. The phylogeny of the Pentaschistis clade (Danthonioideae, Poaceae) based on chloroplast DNA, and the evolution and loss of complex characters. Evolution 61, 864–884.

Gatesy, J., Matthee, C., DeSalle, R., Hayashi, C., 2002. Resolution of a supertree/supermatrix paradox. Syst. Biol. 51, 652–664.

GPWG, 2001. Phylogeny and subfamilial classification of the grasses (Poaceae). Ann. Missouri Bot. Gard. 88, 373–457.

Hardy, C.R., Linder, H.P., 2005. Intraspecific variability and timing in ancestral ecology reconstruction: a test case from the Cape flora. Syst. Biol. 54, 299–316.

Hilu, K.W., Alice, L.A., Liang, H., 1999. Phylogeny of Poaceae inferred from *matK* sequences. Ann. Missouri Bot. Gard. 86, 835–851.

Huelsenbeck, J.P., Ronquist, F., 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755.

Hughes, C., Eastwood, R., Donovan Bailey, C., 2006. Review. From famine to feast? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction. Phil. Trans. R. Soc. B 361, 211–225.

Kellogg, E.A., Appels, R., Mason-Gamer, R.J., 1996. When genes tell different stories: the diploid genera of Triticeae (Gramineae). Syst. Bot. 21, 321–347.

Kim, K.-J., Jansen, R.K., 1995. ndhF sequence evolution and the major clades in the sunflower family. Proc. Natl. Acad. Sci. USA 99, 10379–10383.

Kuzoff, R.K., Sweere, J.A., Soltis, D.E., Soltis, P.S., Zimmer, E.A., 1998. The phylogenetic potential of entire 26S rDNA sequences in plants. Mol. Biol. Evol. 15, 251–263.

Linder, H.P., Barker, N.P., 2005. From Nees to now - changing questions in the systematics of the grass subfamily Danthonioideae. Nova. Acta Leopoldina 92, 29–44.

Linder, H.P., Davidse, G., 1997. The systematics of *Tribolium* Desv. Danthonieae: Poaceae). (Bot. Jahrb. Syst. 119, 445–507.

Linder, H.P., Verboom, G.A., 1996. Generic limits in the Rytidosperma (Danthonieae, Poaceae) complex. Telopea 6, 597–627.

Maddison, D.R., Maddison, W.P., 2005. MacClade. Version 3. Sinaur Associates, Inc., Sunderland, MA.

Moline, P.M., Linder, H.P., 2005. Molecular phylogeny and generic delimitation in the Elegia group (Restionaceae, South Africa) based on a complete taxon sampling and four chloroplast DNA regions. Syst. Bot. 30, 759–772.

Muellner, A.N., Samuel, R., Johnson, S.A., Cheek, M., Pennington, T.D., Chase, M.W., 2003. Molecular phylogenetics of Meliaceae (Sapindales) based on nuclear and plastid DNA sequences. Am. J. Bot. 90, 471–480.

Muller, K., 2006. SeqState, version 1.32. available from: http://www.botanik.uni-bonn.de/system/downloads/SeqStatus.

Nixon, K.C., 1999. The Parsimony Ratchet, a new method for rapid parsimony analysis. Cladistics 15, 407–414.

Nixon, K.C., Carpenter, J.M., 1996. On simultaneous analysis. Cladistics 12, 221–241.

Olmstead, R.G., Sweere, J.A., 1994. Combining data in phylogenetic systematics: an empirical approach using three molecular data sets in the Solanaceae. Syst. Biol. 43, 467–481.

Pisani, D., Wilkinson, M., 2002. Matrix representation with parsimony, taxonomic congruence, and total evidence. Syst. Biol. 51, 151.

Posada, D., Crandall, K.A., 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14, 817–818.

Rambaut, A., Drummond, A.J., 2003. Tracer v. 1.3, Available from: http://beast.bio.ed.ac.uk/Tracer.

Richardson, J.E., Pennington, R.T., Pennington, T.D., Hollingsworth, P.M., 2001. Rapid diversification of a species-rich genus of Neotropical rainforest trees. Science 293, 2242–2245.

Rokas, A., Carroll, S.B., 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol. Biol. Evol. 22, 1337–1344.

Rosenberg, M.S., Kumar, S., 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl. Acad. Sci. USA 98, 10751–10756.

Rydin, C., Kallersjö, M., 2002. Taxon sampling and seed plant phylogeny. Cladistics 18, 485.

Sánchez-Ken, J.G., Clark, L.G., Kellogg, E.A., Kay, E.E., 2007. Reinstatement and emendation of subfamily Micrairoideae (Poaceae). Syst. Bot. 32, 71–80.

Sanderson, M.J., Purvis, A., Henze, C., 1998. Phylogenetic supertrees: assembling the trees of life. Trends Ecol. Evol. 13, 105–109.

Sikes, D.S., Lewis, P.O., 2001. PAUPRat: PAUP* implementation of the parsimony ratchet version 1, beta. Distributed by the authors. Department of Ecology and Evolutionary Biology. University of Connecticut, Storrs, USA.

Simmons, M.P., Ochoterena, H., 2000. Gaps as characters in sequence-based phylogenetic analysis. Syst. Biol. 49, 369–381.

Suzuki, Y., Glazko, G.V., Nei, M., 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA 99, 16138–16143.

Swofford, D. L., 2000. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Sinauer.

Taberlet, P., Gielly, L., Pautou, G., Bouvet, J., 1991. Universal primers for amplification of the three non-coding regions of chloroplast DNA. Plant Mol. Biol. 17, 1105–1109.

Verboom, G.A., Linder, H.P., Barker, N.P., 1994. Haustorial synergids: an important character in the systematics of danthonoid grasses (Arundinoideae: Poaceae)? Am. J. Bot. 81, 1601–1610.

Verboom, G.A., Ntsohi, R., Barker, N.P., 2006. Molecular phylogeny of African Rytidosperma-affiliated danthonioid grasses reveals generic polyphyly and convergent evolution in spikelet morphology. Taxon 55, 337–348.

Vriesendorp, B., Bakker, F.T., 2005. Reconstructing patterns of reticulate evolution in angiosperms: what can we do? Taxon 54, 593–604.

Wiens, J.J., 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52, 528–538.

Wiens, J.J., 2006. Missing data and the design of phylogenetic analysis. J. Biomed. Inform 39, 34–42.

Wiens, J.J., Fetzner, J.W., Parkinson, C.L., Reeder, T.W., 2005. Hylid frog phylogeny and sampling strategies for speciose clades. Syst. Biol. 54, 719–748.

Wilkinson, M., 1995. Coping with abundant missing entries in phylogenetic inference using parsimony. Syst. Biol. 44, 501–514.

Zwickl, D.J., Hillis, D.M., 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51, 588–598.