Elsevier Editorial System(tm) for Molecular Phylogenetics and Evolution

Manuscript Draft

Manuscript Number: MPE-15-437R1

Title: Integration of complete chloroplast genome sequences with small amplicon datasets improves phylogenetic resolution in Acacia

Article Type: Research Paper

Keywords: integrative systematics;
whole chloroplast genome;
Acacia;
ExaBayes;
RAxML

Corresponding Author: Ms. Anna Williams,

Corresponding Author's Institution: Kings Park and Botanic Garden

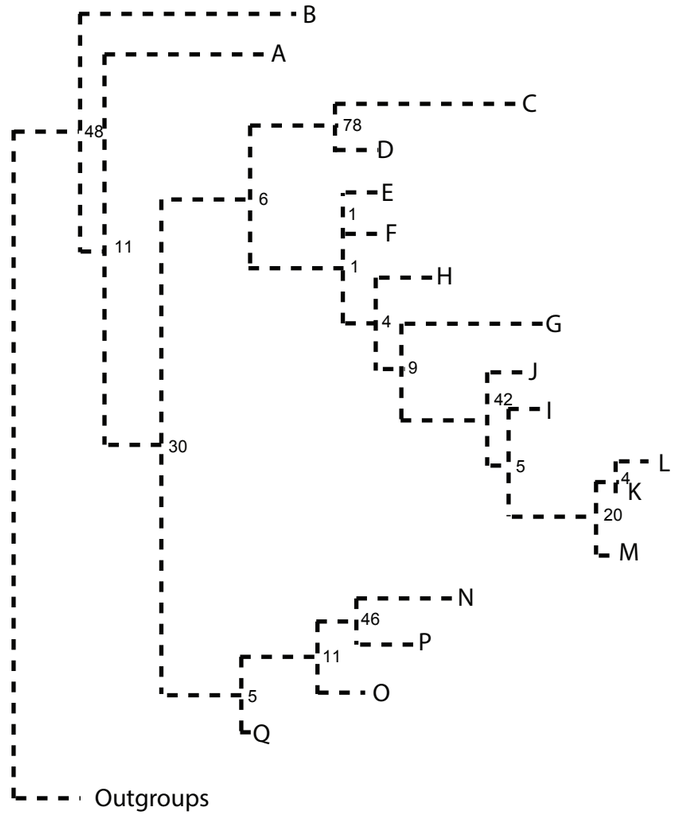
First Author: Anna Williams

Order of Authors: Anna Williams; Joseph T Miller; Ian Small; Paul G Nevill; Laura M Boykin

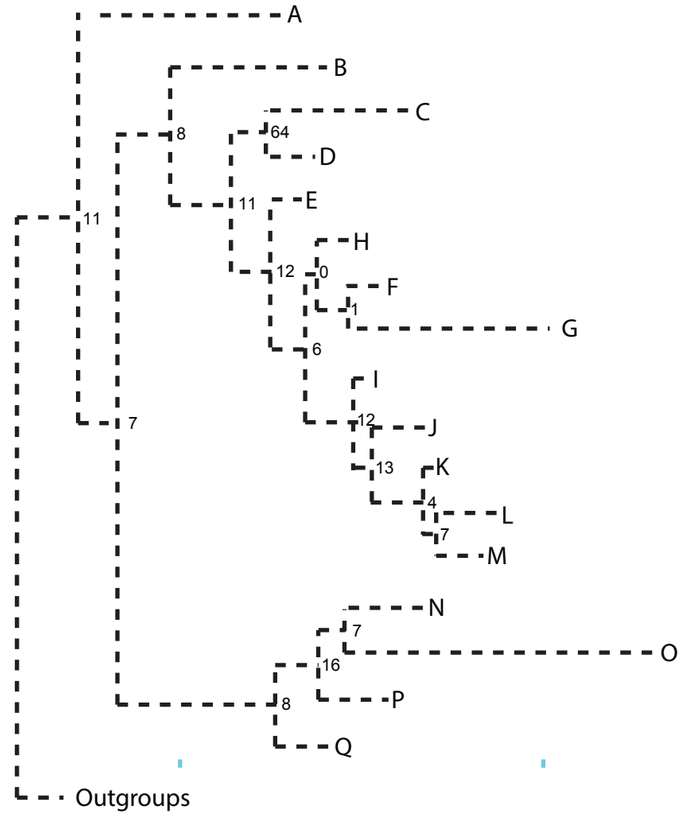
Abstract: Combining whole genome data with previously obtained amplicon sequences has the potential to increase the resolution of phylogenetic analyses, particularly at low taxonomic levels or where recent divergence, rapid speciation or slow genome evolution has resulted in limited sequence variation. However, the integration of these types of data for large scale phylogenetic studies has rarely been investigated. Here we conduct a phylogenetic analysis of the whole chloroplast genome and two nuclear ribosomal loci for 65 Acacia species from across the most recent Acacia phylogeny. We then combine this data with previously generated amplicon sequences (four chloroplast loci and two nuclear ribosomal loci) for 508 Acacia species. We use several phylogenetic methods, including maximum likelihood bootstrapping (with and without constraint) and ExaBayes, in order to determine the success of combining a dataset of 4,000 bp with one of 189,000 bp. The results of our study indicate that the inclusion of whole genome data gave a far better resolved and well supported representation of the phylogenetic relationships within Acacia than using only amplicon sequences, with the greatest support observed when using a whole genome phylogeny as a constraint on the amplicon sequences. Our study therefore provides methods for optimal integration of genomic and amplicon sequences.

*Graphical Abstract (for review)

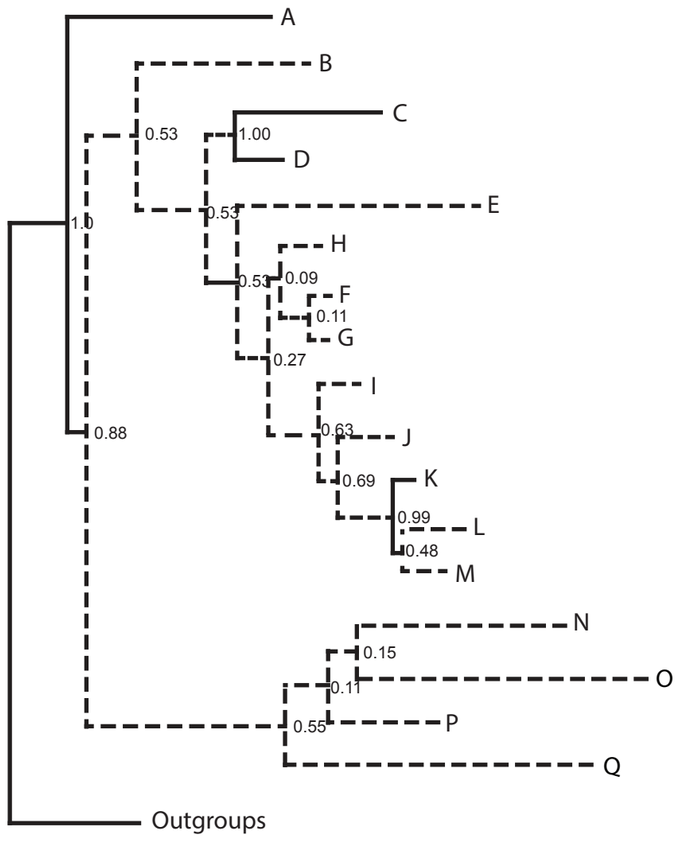
(a) Six gene small amplicon sequence tree



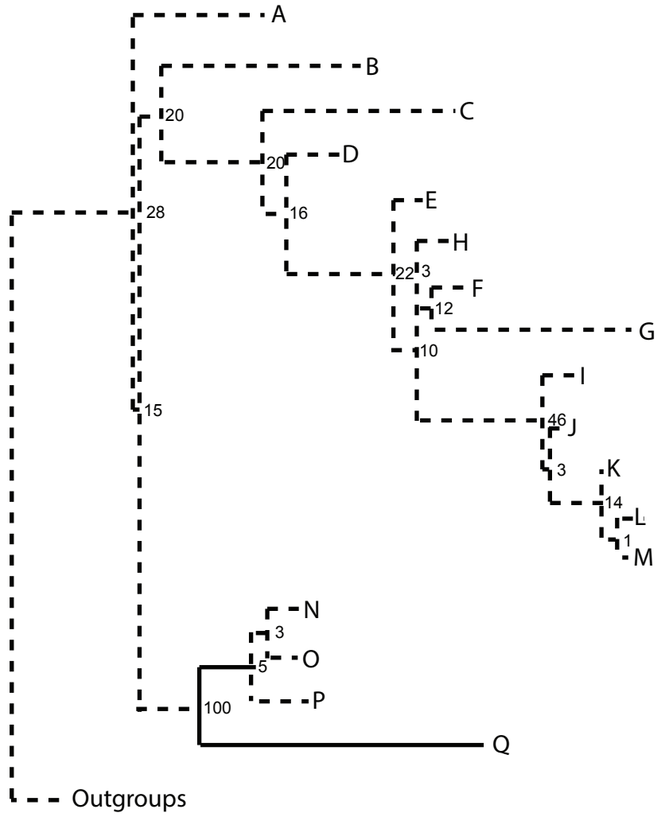
(b) Super matrix analysis (RAxML)



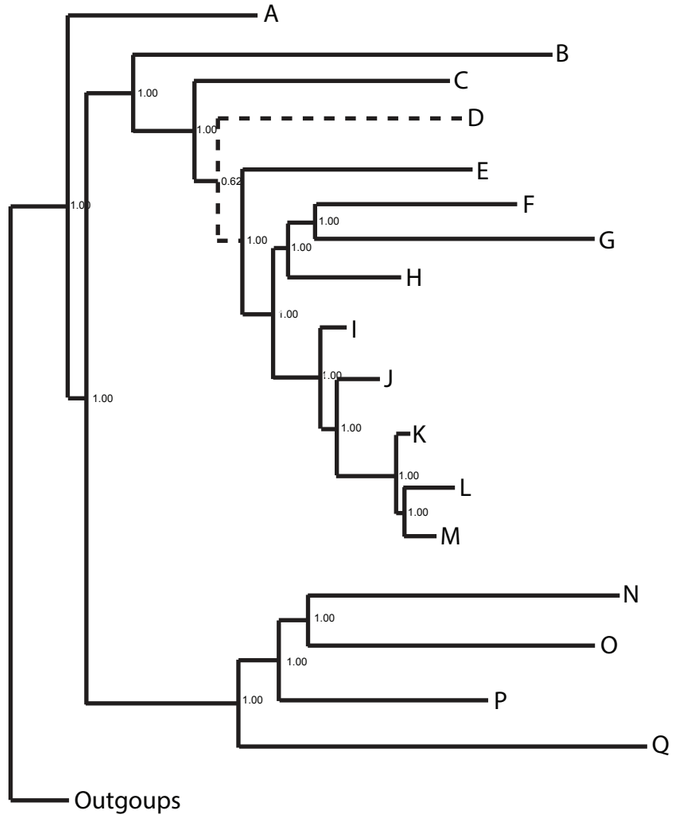
(c) Super matrix analysis (ExaBayes)



(d) Constraint tree



(e) Whole chloroplast genome tree



Integration of complete chloroplast genome sequences with small amplicon datasets improves phylogenetic resolution in *Acacia*

Anna V. Williams[a,b,c,*], Joseph T. Miller[d,e], Ian Small[a], Paul G. Nevill[f,b,c], Laura M. Boykin[a,g]

[a]*Australian Research Council Centre of Excellence in Plant Energy Biology, The University of Western Australia, Crawley, WA 6009, Australia*

[b]*Kings Park and Botanic Garden, Fraser Ave, Kings Park, WA 6005, Australia*

[c]*School of Plant Biology, The University of Western Australia, Crawley, WA 6009, Australia*

[d]*National Research Collections Australia, CSIRO National Facilities and Collections, GPO Box 1600, Canberra, ACT 2601, Australia*

[e]*Division of Environmental Biology, National Science Foundation, 4201 Wilson Blvd, Arlington, VA 22230, USA*

[f]*Department of Environment and Agriculture, Curtin University, Bentley, WA 6102, Australia*

[g]*School of Chemistry and Biochemistry, The University of Western Australia, Crawley, WA 6009, Australia*

*Corresponding author: anna.williams@graduate.uwa.edu.au

Abstract

Combining whole genome data with previously obtained amplicon sequences has the potential to increase the resolution of phylogenetic analyses, particularly at low taxonomic levels or where recent divergence, rapid speciation or slow genome evolution has resulted in limited sequence variation. However, the integration of these types of data for large scale phylogenetic studies has rarely been investigated. Here we conduct a phylogenetic analysis of the whole chloroplast genome and two nuclear ribosomal loci for 65 *Acacia* species from across the most recent *Acacia* phylogeny. We then combine this data with previously generated amplicon sequences (four chloroplast loci and two nuclear ribosomal loci) for 508 *Acacia* species. We use several phylogenetic methods, including maximum likelihood bootstrapping (with and without constraint) and ExaBayes, in order to determine the success of combining a dataset of 4,000 bp with one of 189,000 bp. The results of our study indicate that the inclusion of whole genome data gave a far better resolved and well supported representation of the phylogenetic relationships within *Acacia* than using only amplicon sequences, with the greatest support observed when using a whole genome phylogeny as a constraint on the amplicon sequences. Our study therefore provides methods for optimal integration of genomic and amplicon sequences.

Keywords: integrative systematics, whole chloroplast genome, *Acacia*, ExaBayes, RAxML

1. Introduction

Phylogenetic analysis of plant species has traditionally used highly variable DNA sequence data found throughout chloroplast introns and intergenic spacer regions (Baldauf et al., 2000; Gielly and Taberlet, 1994; Moncalvo et al., 2002; Peterson and Eernisse, 2001; Taberlet et al., 1991). However, using a small number of loci is frequently insufficient to resolve evolutionary relationships, particularly at low taxonomic levels or where recent divergence, rapid speciation or slow genome evolution has limited sequence variation (Kane et al., 2012; Parks et al., 2009; Whittall et al., 2010; Yang et al., 2013; Zhang et al., 2011). Phylogenetic resolution and support is known to depend on both the number of characters and the number of taxa included in a study (Jansen et al., 2007; Philippe et al., 2011). While utilising too few genes may result in incongruence between gene regions and will increase the capacity for error in the phylogeny (Philippe et al., 2011; Rokas and Carroll, 2005), using too few species will result in a phylogeny that is more sensitive to homoplasy. Thus, the ideal is clearly to use the maximum number of genes across the maximum number of taxa.

There has been considerable debate regarding the most efficient way in which to increase resolution in phylogenies and to reduce error (Graybeal, 1998; Hillis, 1998; Mitchell et al., 2000; Nabhan and Sarkar, 2012; Wiens and Tiu, 2012). Although it has been claimed that increased resolution and node support can be equally well achieved by increasing the number of taxa sampled as by increasing the number of characters (Rosenberg and Kumar, 2001, 2003), there is evidence to suggest that in more closely related species, such as within a single genus, increasing the number of characters is more beneficial to resolving a tree (Hillis et al., 2003; Zwickl and Hillis, 2002).

High-throughput sequencing has significantly increased the efficiency of phylogenetic studies, in particular by enabling whole genome (typically organelle) sequencing of non-

63 model species, resulting in a vast increase in the data available for phylogenetic tree
64 construction (Bayly et al., 2013; Huang et al., 2014a; Huang et al., 2014b; Lin et al., 2010;
65 Parks et al., 2009; Zhang et al., 2011). The overall genetic resources are thus increasingly
66 consisting of both multi-locus amplicon sequences, and also whole genome data for a small
67 number of species. While the production of many genomic sequences remains an ongoing
68 process, the integration of a small number of genomic sequences with a large number of
69 small amplicon sequences has the potential to allow a transition towards the more
70 commonplace use of whole genome sequences.

71 The issue of combining datasets with vastly different numbers of characters was first
72 addressed in the context of integrating morphological data, particularly fossil data, with
73 molecular data (Huelsenbeck, 1991; Wiens, 2003a, b, 2005; Wiens et al., 2010), and many of
74 the same principles apply to the integration of genomic (whole genome) data with small
75 amplicon sequences (Roure et al., 2013; Sanderson et al., 2010). Responses to the integration
76 of genomic and amplicon data have varied with some studies indicating that it is the number
77 of characters available rather than the number of characters missing that is the key influence
78 on phylogenetic accuracy (Driskell et al., 2004; Roure et al., 2013; Wiens, 2003a, b; Wiens
79 and Moen, 2008), while other studies suggest that the absence of large amounts of data has
80 significant negative impacts on accuracy (Lemmon et al., 2009). While these findings have
81 been shown in simulated datasets, few empirical studies have attempted the integration of
82 genomic and amplicon sequences.

83 A good test of the potential for genomic and amplicon data integration is in the
84 phylogenetic analysis of the plant genus *Acacia* Mill., which is the most speciose genus in the
85 Mimosoideae subfamily and Leguminosae family. The genus is predominantly found
86 throughout Australia, with only a few species native to Southeast Asia, Hawaii and
87 Madagascar (Brown et al., 2012; González-Orozco et al., 2011; Maslin et al., 2003). *Acacia*

88 has the largest number of species of any angiosperm genus in Australia (over 1,000; Council
89 of Heads of Australasian Herbaria, 2012), and *Acacia* woodlands and shrublands make up
90 approximately 24% of Australia's total vegetation (Beeton et al., 2006). These species are not
91 only of ecological significance, but also play a key role in agroforestry (Brockwell et al.,
92 2005; Midgley and Turnbull, 2003; Thomson et al., 1994), and internationally as invasive
93 species, with 23 species of *Acacia* currently listed as invasive species across 12 different
94 geographical regions (Richardson and Rejmánek, 2011). Consequently, understanding the
95 phylogenetic relationships between these species is vital for informing conservation,
96 agroforestry and invasive species management.

97 Substantial incremental knowledge of *Acacia* phylogenetics has been gained over the
98 past two decades through amplicon sequences of nuclear ribosomal (ITS and ETS) and
99 selected plastid loci (e.g. *psbA-trnH, trnL-trnF, rpl32-trnL, matK*) (Miller et al., 2003; Miller
100 and Bayer, 2001, 2003; Murphy et al., 2010; Murphy et al., 2003; Murphy et al., 2000),
101 leading to a phylogeny containing over 500 species terminals (Mishler et al., 2014). These
102 studies have identified well-supported major clades similar to those identified by Murphy et
103 al. (2010), and have provided strong support for many relationships near the tips of the tree
104 and other internal nodes; however, the backbone nodes remain poorly supported with less
105 than 20% of nodes showing bootstrapping support greater than 0.95. Thus, additional taxa
106 and/or character data are necessary to understand the evolutionary relationships of *Acacia*.

107 Here we demonstrate the feasibility and effectiveness of incorporating whole chloroplast
108 genome sequences with small amplicon sequences from a limited number of loci produced in
109 previous phylogenetic analysis of *Acacia*. In this study we sequence the chloroplast genomes
110 for 65 *Acacia* species from across the most recent phylogeny (Mishler et al., 2014). We firstly
111 identify whether increasing the number of characters or taxa has the greatest influence on
112 phylogenetic resolution and support in *Acacia*, then combine our data with the 510 specimens

of Mishler et al. (2014), using both maximum likelihood bootstrapping (with and without constraints) and Bayesian methods to identify the best method of data integration.

2. Materials and Methods

2.1. Sampling

This dataset consisted of 65 *Acacia* species (a total of 94 individuals), and two outgroups, *Pararchidendron pruinosum* and *Paraserianthes lophantha* subsp. *lophantha*. Phyllode material was collected from 77 individuals from native populations, eight individuals from within Kings Park and Botanic Garden (West Perth, Western Australia) and from nine specimens held at the Western Australian Herbarium (Kensington, Western Australia; see Appendix A for all specimen details and herbarium voucher numbers).

2.2. DNA Sequencing

Total genomic DNA was extracted from either fresh or dried phyllode material using the methods of Jobes et al. (1995) or Butcher et al. (1998). DNA quality and quantity were assessed using a NanoDrop spectrophotometer (ND-1000; Thermo Fisher Scientific, USA), and via agarose gel electrophoresis. Individual genome library preparations were performed using a Nextera DNA Sample Preparation Kit (Illumina, USA), following the manufacturer's instructions. Libraries were then prepared for sequencing using the cBot cluster generation and PE V3 flow cell and cluster generation (Illumina, USA). The libraries were sequenced on a single lane in paired end mode using the HiSeq2000 platform and V3 SBS kit (Illumina, USA). Library preparations and sequencing were both performed at the Ramaciotii Centre for Gene Function Analysis (Sydney, Australia; http://devspace.ddtoo.com).

136

137 *2.3. Sequence Assembly*

138 For each specimen, overlapping paired-end reads were merged using the software
139 FLASH (version 1.2.7; Magoc and Salzberg, 2011). Merged reads were assembled using
140 Velvet (version 1,2,08; Zerbino and Birney, 2008) with k-mer values of 31, 41, 51 and 61,
141 and coverage cut-off of 10. For each assembly, MUMmer (version 3.0; Kurtz et al., 2004)
142 was used to compare the assembled chloroplast contigs with the closest related complete
143 chloroplast genome sequence available, *Acacia ligulata* Benth. (Leguminosae; EMBL
144 accession number LN555649). Contigs were then merged to produce a single draft genome.
145 Assemblies were refined by repeatedly mapping raw reads to the draft sequence using
146 Geneious (version 6.1.8; Drummond et al., 2011) and adjusting as necessary. Draft genomes
147 were annotated by direct comparison with the *A. ligulata* genome and sequences were
148 deposited into EMBL (accession numbers are available in Appendix B). Raw reads were also
149 mapped to ITS and ETS sequences from *Acacia anthochaera* Maslin (Genbank accessions
150 DQ029243 and DQ029284) using Geneious (Drummond et al., 2011). All 95 draft genomes
151 and the *A. ligulata* reference genome were aligned using MAFFT (Katoh et al., 2002) in
152 Geneious (Drummond et al., 2011). Due to variation in inverted repeat sizes, particularly
153 relative to the outgroups, only one IR copy was included in the alignment. Separate ITS and
154 ETS alignments were also developed for all 96 specimens using MAFFT (Katoh et al., 2002).

155

156 *2.4. Phylogenetic Analyses*

157 *2.4.1. Effect of character number of phylogenetic accuracy*

158 In order to compare the influence of increased characters on phylogenetic accuracy, a
159 subset of taxa was taken separately from both our dataset and that of Mishler et al. (2014),

which included only those taxa present in both datasets. For both subsets, Bayesian analyses were conducted using the program ExaBayes (version 1.4.1; Aberer et al., 2014) on the Magnus supercomputer (located at the Pawsey Centre, Kensington, Western Australia). Analyses were run for 10 million generations with sampling every 500 generations. Each analysis consisted of four independent runs, each utilising four chains. Convergence between runs was monitored by finding a plateau in the likelihood scores (standard deviation of split frequencies < 0.0015). Convergence of additional parameters was also checked during post-processing, with all ESS vales above 200. The first 25% of each run was discarded as burn-in for the estimation of a majority rule consensus topology and posterior probability for each node.

2.4.2. Effect of increased taxa on phylogenetic accuracy

Our second analysis was designed to provide a baseline for comparing the effect of additional taxa on the integrated dataset. This was achieved by constructing a phylogenetic tree using specimens from both datasets but only at the six loci used by Mishler et al. (2014). Each chloroplast locus was extracted from the whole genome alignment, and individual loci (including ITS and ETS) were aligned with their corresponding alignment in the Mishler et al. (2014) datasets using the MAFFT consensus alignment in Geneious (Drummond et al., 2011; Katoh et al., 2002). All six loci were then concatenated to form a complete dataset for all 606 specimens. The alignment was then used in a maximum likelihood bootstrapping analysis with RAxML (version 8.1.11; Stamatakis, 2014) on the CIPRES Science Gateway server (Miller et al., 2010).

2.4.3. Super matrix integration of increased taxa and characters

184 The integration of the datasets was firstly performed by simply combining the 510
185 specimens of Mishler et al. (2014) with the 96 genomes from this study into a single
186 alignment using the MAFFT consensus alignment in Geneious (Drummond et al., 2011). The
187 resulting alignment was analysed using the RAxML method (above), and then using the
188 ExaBayes method (above), with the analysis taking approximately 14 days of walltime (4
189 years 275 days of CPU time).

190

191 *2.4.4. Constraint analysis integration of increased taxa and characters*

192 Finally, in order to remove potential bias caused by the presence of missing data, and
193 also to incorporate information present in the genomic sequences, we used the ExaBayes
194 method to produce a phylogenetic tree based solely on the 96 chloroplast genomes. Bayesian
195 analysis of the full chloroplast genome alignment took approximately 12 hours of walltime
196 (4,486 hours of CPU time). The RAxML method was then used to analyse all sequences at
197 the six loci of Mishler et al. (2014), using the whole genome phylogenetic topology as a
198 constraint. The differences between all our integrated trees were determined using the
199 program HashRF (version 6.0.1; Sul and Williams, 2007; Sul and Williams, 2008), which
200 computes the Robinson-Foulds (RF) distance between pairs of trees.

201

202 **3. Results**

203 *3.1. Chloroplast Assembly*

204 Illumina sequencing of libraries prepared from total DNA produced between 405,245
205 and 4,041,457 paired-end reads with a length of 100 nt. For each specimen, approximately
206 5% of reads was assembled into contigs that were homologous to the *A. ligulata* reference

chloroplast. Annotation of the draft genomes confirmed the presence of 76 unique protein coding genes, 4 rRNA genes and 30 tRNA genes, in each individual, indicating that there had been no loss of genes or introns relative to *A. ligulata*. All genes were fully assembled for all 95 individuals, with the exception of the *accD* gene, which displayed a several 100 bp repeat region which could not be accurately assembled, and the *trnS-GCU* gene which could be only partially assembled in six individuals. Of a total of 109 intergenic spacer regions, 21 could not be fully assembled. Following removal of unassembled regions, specimens maintained between 78.1 and 98.5% identity with the *A. ligulata* reference (Appendix B). Key differences between species included inversion of the region between *ndhC* and *trnV-UAC* in *A. exocarpoides, A. erinacea,* of the region between *psbE* and *trnV-UAC* in *A. acanthoclada* subsp. *glaucescens, A. scalene* and *A. acuaria* and of the region between *psaI* and *ycf4* in *A. cerastes, A. restiacea, A. scleroclada* and *A. woodmaniorum*. These inversions were reverted in later analyses in order to facilitate alignment of genes.

3.2. Is Increased Resolution Caused by the Addition of Characters or Taxa?

3.2.1. Effect of character number of phylogenetic accuracy

In order to test whether the addition of characters or taxa was responsible for any changes observed in support and resolution of the integrated phylogenies, we firstly created separate phylogenetic trees from both our dataset and that of Mishler et al. (2014) using only the taxa in common to both. Each alignment consisted of 41 *Acacia* species and two outgroups. Bayesian analysis of the Mishler et al. (2014) subset created a phylogeny with 61.0% of nodes displaying a high level of support (posterior probabilities of 0.95 or more; Fig. 1a). In contrast, the whole genome phylogeny was highly supported in 94.9% of nodes (Fig. 1b). Key clades were compared between these two trees (clades A-Q; Fig. 1). The most

231 important differences seen in the phylogeny created from the Mishler et al. (2014) data
232 included clade A forming a sister group to clades N-Q (PP = 0.9), clade C forming a sister
233 group to clade D (PP=0.94) rather than basal to clades D-M, and clade E forming a sister
234 group to clade G (PP = 0.74) rather than basal to clades G-M. Clade N also formed a sister
235 group with clade P (PP = 0.83) rather than clade O. A number of species also appeared within
236 different clades in each tree, for example, *A. andrewsii*, *A. obtecta*, *A. hemiteles*, *A. acuaria*
237 and *A. stanleyi*.

238

239 *3.2.2. Effect of increased taxa on phylogenetic accuracy*

240 In order to compare the influence of increased taxa on the *Acacia* phylogeny, we
241 followed the method of Mishler et al. (2014) to create a tree using only the six loci from both
242 datasets. Combining the loci in common to both datasets resulted in an alignment of 3,956 bp.
243 In total, this combined dataset consisted of 602 *Acacia* specimens (534 species) and four
244 outgroups (2 species). Overall support for this tree was low with only 18.3% of nodes
245 showing bootstrap values of 95% of more (Appendix C). The major clades previously
246 identified by Murphy et al. (2010) were all present within this phylogeny (Fig. 2), although
247 the presence of another clade (also observed in the Mishler et al. (2014) phylogeny; hereafter
248 referred to as the *A. longispinea* clade) was evident. Support for these clades was highest in
249 the *A. victoriae* / *A. pyrifolia* clade (BS = 100%). The other clades were far less well-
250 supported with 52% for the *A. longispinea*, 78% for the *A. murrayana* clade, 7% for the p.u.b.
251 clade, 30% for the *Pulchelloidea* clade and 67% for the Botrycephalae subclade. Smaller
252 clades (A-Q) were identified in order to more closely compare trees. These clades all showed
253 low support, with bootstrap support values between 1% and 78% (Fig. 3a). Of the 41 species
254 present in both datasets, 22 occurred within the same clade and a further 17 formed
255 monophyletic clades with conspecific individuals (Appendix C).

256

257 *3.3. Integration of Genomic and Amplicon Sequences*

258 *3.3.1. Super matrix analysis (RAxML)*

259 All 606 specimens were used to create a phylogeny using any available data for the
260 given individual, i.e. approximately 4,000 bp for 510 specimens and approximately 141,000
261 bp for 96 specimens. This meant that the overall alignment contained a large proportion of
262 missing data. Overall, this tree displayed low support (18.6% of nodes showed high support;
263 Appendix D). The major clades were all present within this tree with high support observed
264 in the *A. victoriae* / *A. pyrifolia* clade (100% support). The *A. longispinea* and *A. murrayana*
265 clades displayed 62% and 64$%, respectively, while p.u.b. clade (BS = 12%), Pulchelloidea
266 clade (BS = 7%) and Botrycephalae subclade (BS = 27%) all showed much lower support. Of
267 the smaller clades, all were present but none displayed a high level of support, with clades
268 displaying between 0% and 64% bootstrap support (Fig. 3). Of the 41 species present in both
269 datasets, 22 occurred in the same clade and a further 16 were monophyletic with conspecific
270 individuals (Appendix D).

271

272 *3.3.2. Super matrix analysis (ExaBayes)*

273 The super matrix analysis using ExaBayes produced the tree with the most variation
274 from the other combined trees (RF = 200-217; Table 1), and overall support was still low at
275 42.0% (Appendix E). The major clades were all present and the *A. longispinea* and *A.*
276 *victoriae* / *A. pyrifolia* clades showed posterior probabilities of 0.95 or more (Fig. 2). The
277 smaller clades were also highly supported in four out of the seventeen clades (A, C, D, K;
278 Fig. 3c). Seventeen of the species present in both datasets formed monophyletic groups, while

nineteen others were present in the same larger clades as other conspecific individuals (Appendix E).

3.3.3. Constraint analysis

In order to incorporate the genomic data while also avoiding large proportions of missing data within the overall dataset, the 3, 956 bp alignment was analysed using RAxML, with a topology of the 96 genomes analysis as a constraint. To develop the constraint, we analysed all 96 whole plastid genomes separately. The complete MAFFT alignment of all 96 genomes resulted in an aligned length of 187,573 bp. This tree was highly supported in 96.8% of nodes (Appendix F). Sixteen out of the seventeen smaller clades showed a high level of support for their topology with the lowest posterior probability observed in the tree being only 0.62 (Fig. 3e). Given the high support for this tree, we were confident that this topology provided a good constraint for the backbone of the larger dataset. Using this tree as a constraint on the 3,956 bp alignment produced an identical topology to the whole genome tree, but with far lower overall support (20.0% of nodes were highly supported; Appendix G). This tree showed the greatest similarity to the small amplicon sequence tree (RF = 144; Table 1), with the major clades again showing high support in the *A. victoriae* / *A. pyrifolia* clade (bootstrapping support of 100%; Fig. 2), while the smaller clades had lower support ranging from 1% to 100% (Fig. 3d). Of the 41 specimens present in both datasets, 17 formed monophyletic clades and 21 others were present within the same clade as conspecific individuals (Appendix G).

4. Discussion

Of key interest to this study is the extent to which using this genomic data increased the support of the *Acacia* phylogeny. In order to determine whether increase in characters or taxa was responsible for any perceived increase in support and resolution, we firstly compare the support and resolution of two trees that differed only in the number of characters used to build them (Fig. 1). Our results clearly showed that, with 94.5% of nodes showing a posterior probability of more than 0.95 (Fig. 1), the use of a much larger volume of data produced 1.5X the number of highly supported nodes compared to when only six loci were used (where only 61.0% of nodes were highly supported; Fig. 1). This result was consistent with previous findings in which a much higher level of support was observed in a genomic phylogeny of *Pinus* species (Parks et al., 2009), than when small amplicon sequences were used (Gernandt et al., 2005; Liston et al., 2007; Syring et al., 2007; Wang et al., 1999). Similar results have also been observed from the whole chloroplast genome analysis of apple species (Nikiforova et al., 2013), rice species (Waters et al., 2012) and *Araucaria* species (Ruhsam et al., 2015).

Our analysis of small amplicon sequences supported our hypothesis that the number of characters had a greater influence on the support and resolution of the *Acacia* phylogeny. In this analysis, the two datasets were combined but only analysed using the six loci in common to all 606 specimens. Although this tree was slightly different to the phylogeny of Mishler et al. (2014), in particular clade O becoming a sister group to clades N+P+Q, the addition of taxa failed to improve the overall support of the tree which remained at only 18%. This result confirmed that the addition of further taxa was insufficient to produce a more well-supported phylogeny, and indicated that any increase in support observed in subsequent integrated trees was most likely caused by the increased number of characters. This result was consistent with the findings of Rokas and Carroll (2005), who also identified increased characters rather than increased taxa as being the key influence on phylogenetic accuracy in yeast.

4.1. Integration of Genomic and Amplicon Sequences

Although our initial results using a reduced number of taxa clearly showed that the use of whole genome sequences has the potential to increase phylogenetic support and resolution, the challenge remains in finding the best method of data integration. The phylogeny developed by Mishler et al. (2014), while showing strong support for the major clades, including *A. victorae* / *A. pyrifolia*, *A. murrayana* and *A. longispinea* clades, was less well resolved in the p.u.b. and Pulchelloidea clades and Botrycephalae subclade, and among the minor clades only showed high support for clades B, C, N, P and Q.

The addition of full genomic sequences to the dataset showed a clear change in the relationships among the clades compared to what was seen in both the Mishler et al. (2014) tree and the small amplicon sequence tree. The super matrix analysis tree clearly showed clade A as sister to clades B-Q, and clade H as a sister group to clades F+G. Additionally, clades L+M became sister to clade K, and clades N+O sister to clade P. Despite the change in tree topology, the RAxML tree did not show any more significant support than was seen in the small amplicon sequence tree.

The ExaBayes super matrix analysis revealed an identical topology to the RAxML analysis with regards to the small clades (Fig. 3c), however the RF calculation clearly showed that the position of the tips within those clades was quite different (RF = 200; Table 1). The ExaBayes tree showed generally better support for the major clades and for the positions of many of the minor clades, with clades A, C, D, and K all showing posterior probabilities of greater than 0.95 (Fig. 3), suggesting that this tree was a better phylogenetic reconstruction than the RAxML tree. It should be noted however, that some of this may potentially have been an artefact of the Bayesian method, which has previously been identified as exhibiting higher support values than when using maximum likelihood methods (Douady et al., 2003;

Simmons et al., 2004). Compared to the phylogeny of Mishler et al. (2014), there remained a number of differences, including clade A becoming a sister group to clades B-P, clade F becoming a sister group to clade G and clade O becoming a sister group to clade N (Fig. 3c).

As expected, from the phylogenies based only on the taxa held in common to both datasets, the whole genome tree showed the greatest support of any of the trees. This phylogeny showed high support for sixteen out of the seventeen minor clades (Fig. 3). By using this tree as a constraint on the amplicon sequence data, we were able to remove any error caused by large proportions of missing data, while also maintaining the highly supported backbone identified in the whole genome phylogeny (Appendix F). The relationships between the minor clades were very similar to that seen in the super matrix analyses with the exception of clade C which became sister to clades D-M (PP = 0.62). The topology of the highly supported whole genome phylogeny was reflected in the constraint tree; however the constraint lacked the high support values found in the whole genome tree due to our reliance on a subset of the sequence length used in the whole genome tree. However, given that the topology of the minor clades was highly supported in the whole genome tree, we conclude that the constraint tree enabled the best integration of genomic and small amplicon sequence data.

5. Conclusions

Our study shows that the use of whole chloroplast genome data for phylogenetics provides a far greater support and resolution than can be achieved using a small number of amplicon sequences. The results of our analyses suggest that the whole genome sequences play an important role in identifying highly supported nodes in the backbone of large phylogenies. The integration of data types showed typically low support, however higher

support was seen using Bayesian methods, and the best supported topology was achieved by using genomic sequences to build a highly supported backbone, upon which a large number of small amplicon sequences can be constrained. Our analyses have clearly shown the potential of genomic and amplicon data integration in phylogenetic studies of large genera, however this method is likely to also improve resolution and support of phylogenies displaying weak backbone support and where closely related species require additional characters to fully understand the phylogenetic relationships between them. We believe that the integration of genomic and amplicon sequences provides a practical means of bridging the gap between the large number of amplicon sequences currently available and the ever-increasing number of genomic sequences that continue to be created.

Acknowledgements

This works was supported by an Australian Postgraduate Award to AVW. Additional funds were provided in kind by Bioplatforms Australia and by Karara Mining Ltd. This work was also supported by resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia. We would like to thank Bruce Maslin and Ladislav Mucina for their aid in specimen identification and storage, and Karina Knight for her assistance in obtaining specimens from the Western Australian Herbarium. This manuscript includes work done by JTM while serving at the National Science Foundation. The views expressed in this paper do not necessarily reflect those of the National Science Foundation or the United States Government.

References

399 Aberer, A.J., Kobert, K., Stamatakis, A., 2014. ExaBayes: massively parallel Bayesian
400 tree inference for the whole-genome era. Molecular Biology and Evolution 31, 2553-2556.

401 Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., Doolittle, W.F., 2000. A Kingdom-Level
402 Phylogeny of Eukaryotes Based on Combined Protein Data. Science 290, 972-977.

403 Bayly, M.J., Rigault, P., Spokevicius, A., Ladiges, P.Y., Ades, P.K., Anderson, C.,
404 Bossinger, G., Merchant, A., Udovicic, F., Woodrow, I.E., 2013. Chloroplast genome
405 analysis of Australian eucalypts - *Eucalyptus*, *Corymbia*, *Angophora*, *Allosyncarpia* and
406 *Stockwellia* (Myrtaceae). Molecular Phylogenetics and Evolution 69, 704-716.

407 Beeton, R., Buckley, K.I., Jones, G.J., Morgan, D., Reichelt, R.E., Trewin, D., 2006.
408 Independent report to the Australian Government Minister for the Environment and Heritage.
409 In: Environment, D.o. (Ed.). 2006 Australian State of the Environment Committee.

410 Brockwell, J., Searle, S.D., Jeavons, A.C., Waayers, M., 2005. Nitrogen fixation in
411 acacias: an untapped resource for sustainable plantations, farm forestry and land reclamation.
412 Australian Centre for International Agricultural Research (ACIAR).

413 Brown, G.K., Murphy, D.J., Kidman, J., Ladiges, P.Y., 2012. Phylogenetic connections
414 of phyllodinous species of *Acacia* outside Australia are explained by geological history and
415 human-mediated dispersal. Australian Systematic Botany 25, 390-403.

416 Butcher, P.A., Moran, G.F., Perkins, H.D., 1998. RFLP diversity in the nuclear genome
417 of *Acacia mangium*. Heredity 81, 205-213.

418 Council of Heads of Australasian Herbaria, 2012. Australian Plant Census.

419 Douady, C.J., Delsuc, F., Boucher, Y., Doolittle, W.F., Douzery, E.J.P., 2003.
420 Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic
421 reliability. Molecular Biology and Evolution 20, 248-254.

422 Driskell, A.C., Ané, C., Burleigh, J.G., McMahon, M.M., O'Meara, B.C., Sanderson,
423 M.J., 2004. Prospects for building the tree of life from large sequence databases. Science 306,
424 1172-1174.

425 Drummond, A.J., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., Field, M.,
426 Heled, J., Kearse, M., Markowitz, S., Moir, R., Stones-Havas, S., Sturrock, S., Thierer, T.,
427 Wilson, A., 2011. Geneious v. 5.4.

428 Gernandt, D.S., López, G.G., García, S.O., Liston, A., 2005. Phylogeny and
429 classification of *Pinus*. Taxon 54, 29-42.

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

432 González-Orozco, C.E., Laffan, S.W., Miller, J.T., 2011. Spatial distribution of species
433 richness and endemism of the genus *Acacia* in Australia. Australian Journal of Botany 59,
434 601-609.

435 Graybeal, A., 1998. Is it better to add taxa or characters to a difficult phylogenetic
436 problem? Systematic Biology 47, 9-17.

437 Hillis, D.M., 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias.
438 Systematic Biology, 3-8.

Hillis, D.M., Pollock, D.D., McGuire, J.A., Zwickl, D.J., 2003. Is sparse taxon sampling a problem for phylogenetic inference? Systematic Biology 52, 124.

Huang, D., Huang, C., Hefer, N., Kolosova, C., Douglas, Q.C.B., Cronk, 2014a. Whole plastome sequencing reveals deep plastid divergence and cytonuclear discordance between closely related balsam poplars, *Populus balsamifera* and *P. trichocarpa* (Salicaceae). New Phytologist 204, 693-703.

Huang, H., Shi, C., Lui, Y., Mao, S.-Y., Gao, L.-Z., 2014b. Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. BMC Evolutionary Biology 14.

Huelsenbeck, J.P., 1991. When are fossils better than extant taxa in phylogenetic analysis? Systematic Biology 40, 458-469.

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., dePamphilis, C.W., Leebens-Mack, J., Müller, K.F., Guisinger-Bellian, M., Haberle, R.C., Hansen, A.K., Chumley, T.W., Lee, S.-B., Rhiannon, P., McNeal, J.R., Kuehl, J.V., Boore, J.L., 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proceedings of the National Academy of Sciences of the United States of America 104, 19369-19374.

Jobes, D.V., Hurley, D.L., Thien, L.B., 1995. Plant DNA isolation: a method to efficiently remove polyphenolics, polysaccharides, and RNA. Taxon 44, 379-386.

Kane, N., Sveinsson, S., Dempewolf, H., Yang, J.Y., Zhang, D., Engels, J.M.M., Cronk, Q., 2012. Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. American Journal of Botany 99, 320-329.

Katoh, K., Misawa, K., Kuma, K.i., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30, 3059-3066.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S.L., 2004. Versatile and open software for comparing large genomes. Genome Biology 5, R12.

Lemmon, A.R., Brown, J.M., Stanger-Hall, K., Lemmon, E.M., 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Systematic Biology 58, 130-145.

Lin, C.-P., Huang, J.-P., Wu, C.-S., Hsu, C.-Y., Chaw, S.-M., 2010. Comparative chloroplast genomics reveals the evolution of Pinaceae genera and subfamilies. Genome Biology and Evolution 2, 504-517.

Liston, A., Parker-Defeniks, M., Syring, J.V., Willyard, A., Cronn, R., 2007. Interspecific phylogenetic analysis enhances intraspecific phylogeographical inference: a case study in *Pinus lambertiana*. Molecular Ecology 16, 3926-3937.

Magoc, T., Salzberg, S., 2011. FLASH: Fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27, 2957-2963.

Maslin, B.R., Miller, J.T., Seigler, D.S., 2003. Overview of the generic status of *Acacia* (Leguminosae: Mimosoideae). Australian Systematic Botany 16, 1-18.

Midgley, S.J., Turnbull, J.W., 2003. Domestication and use of Australian acacias: case studies of five important species. Australian Systematic Botany 16, 89-102.

Miller, J.T., Andrew, R., Bayer, R.J., 2003. Molecular phylogenetics of the Australian acacias of subg. *Phyllodineae* (Fabaceae: Mimosoideae) based on the *trnK* intron. Australian Journal of Botany 51, 167-177.

Miller, J.T., Bayer, R.J., 2001. Molecular phylogenetics of *Acacia* (Fabaceae: Mimosoideae) based on the chloroplast *matK* coding sequence and flanking *trnK* intron spacer regions. American Journal of Botany 88, 697-705.

Miller, J.T., Bayer, R.J., 2003. Molecular phylogenetics of *Acacia* subgenera *Acacia* and *Aculeiferum* (Fabaceae: Mimosoideae), based on the chloroplast *matK* coding sequence and flanking *trnK* intron spacer regions. Australian Systematic Botany 16, 27-33.

Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Gateway Computing Environments Workshop (GCE), 2010. IEEE, pp. 1-8.

Mishler, B.D., Knerr, N., Gonzalez Orozco, C.E., Thornhill, A.H., Laffan, S.W., Miller, J.T., 2014. Phylogenetic measures of biodiversity and neo- and paleo-endemism in Australian *Acacia*. Nature Communications.

Mitchell, A., Mitter, C., Regier, J.C., 2000. More taxa or more characters revisited: combining data from nuclear protein-encoding genes for phylogenetic analyses of Noctuoidea (Insecta: Lepidoptera). Systematic Biology 49, 202-224.

Moncalvo, J.-M., Vilgalys, R., Redhead, S.A., Johnson, J.E., James, T.Y., Catherine Aime, M., Hofstetter, V., Verduin, S.J.W., Larsson, E., Baroni, T.J., Greg Thorn, R., Jacobsson, S., Clémençon, H., Miller Jr, O.K., 2002. One hundred and seventeen clades of euagarics. Molecular Phylogenetics and Evolution 23, 357-400.

Murphy, D.J., Brown, G.K., Miller, J.T., Ladiges, P.Y., 2010. Molecular phylogeny of *Acacia* Mill.(Mimosoideae: Leguminosae): evidence for major clades and informal classification. Taxon, 7-19.

Murphy, D.J., Miller, J.T., Bayer, R.J., Ladiges, P.Y., 2003. Molecular phylogeny of *Acacia* subgenus *Phyllodineae* (Mimosoideae: Leguminosae) based on DNA sequences of the internal transcribed spacer region. Australian Systematic Botany 16, 19-26.

Murphy, D.J., Udovicic, F., Ladiges, P.Y., 2000. Phylogenetic analysis of Australian *Acacia* (Leguminosae: Mimosoideae) by using sequence variations of an intron and two intergenic spacers of chloroplast DNA. Australian Systematic Botany 13, 745-754.

Nabhan, A.R., Sarkar, I.N., 2012. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. Briefings in Bioinformatics 13, 122-134.

Nikiforova, S.V., Cavalieri, D., Velasco, R., Goremykin, V., 2013. Phylogenetic analysis of 47 chloroplast genomes clarifies the contribution of wild species to the domesticated apple maternal line. Molecular Biology and Evolution 30, 1751-1760.

Parks, M., Cronn, R., Liston, A., 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biology 7, 84.

Peterson, K.J., Eernisse, D.J., 2001. Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences. Evolution & Development 3, 170-205.

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T.J., Manuel, M., Wörheide, G., Baurain, D., 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol 9, e1000602.

Richardson, D.M., Rejmánek, M., 2011. Trees and shrubs as invasive alien species – a global review. Diversity and Distributions 17, 788-809.

Rokas, A., Carroll, S.B., 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.

Rosenberg, M.S., Kumar, S., 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proceedings of the National Academy of Sciences of the United States of America 98, 10751-10756.

Rosenberg, M.S., Kumar, S., 2003. Taxon sampling, bioinformatics, and phylogenomics. Systematic Biology 52, 119-124.

Roure, B., Baurain, D., Philippe, H., 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Molecular biology and evolution 30, 197-214.

Ruhsam, M., Rai, H.S., Mathews, S., Ross, T.G., Graham, S.W., Raubeson, L.A., Mei, W., Thomas, P.I., Gardner, M.F., Ennos, R.A., 2015. Does complete plastid genome sequencing improve species discrimination and phylogenetic resolution in Araucaria? Molecular Ecology Resources.

Sanderson, M.J., McMahon, M.M., Steel, M., 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. BMC Evolutionary Biology 10, 155.

Simmons, M.P., Pickett, K.M., Miya, M., 2004. How meaningful are Bayesian support values? Molecular Biology and Evolution 21, 188-199.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312-1313.

Sul, S.-J., Williams, T.L., 2007. A randomized algorithm for comparing sets of phylogenetic trees. APBC, pp. 121-130.

Sul, S.-J., Williams, T.L., 2008. An experimental analysis of robinson-foulds distance matrix algorithms. Algorithms-ESA 2008. Springer, pp. 793-804.

Syring, J., Farrell, K., Businský, R., Cronn, R., Liston, A., 2007. Widespread genealogical nonmonophyly in species of *Pinus* subgenus *Strobus*. Systematic Biology 56, 163-181.

Taberlet, P., 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.

Thomson, L.A.J., Turnbull, J.W., Maslin, B.R., 1994. The utilization of Australian species of *Acacia*, with particular reference to those of the subtropical dry zone. Journal of Arid Environments 27, 279-295.

562 Wang, X.-R., Tsumura, Y., Yoshimaru, H., Nagasaka, K., Szmidt, A.E., 1999.
563 Phylogenetic relationships of Eurasian pines (*Pinus*, Pinaceae) based on chloroplast *rbcL,*
564 *matK, rpl20-rps18* spacer, and *trnV* intron sequences. American Journal of Botany 86, 1742-
565 1753.

566 Waters, D.L.E., Nock, C.J., Ishikawa, R., Rice, N., Henry, R.J., 2012. Chloroplast
567 genome sequence confirms distinctness of Australian and Asian wild rice. Ecology and
568 Evolution 2, 211-217.

569 Whittall, J.B., Syring, J., Parks, M., Buenrostro, J., Dick, C., Liston, A., Cronn, R.,
570 2010. Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare
571 and widespread pines. Molecular Ecology 19, 100-114.

572 Wiens, J.J., 2003a. Incomplete taxa, incomplete characters, and phylogenetic accuracy:
573 is there a missing data problem? Journal of Vertebrate Paleontology 23, 297-310.

574 Wiens, J.J., 2003b. Missing data, incomplete taxa, and phylogenetic accuracy.
575 Systematic Biology 52, 528-538.

576 Wiens, J.J., 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch
577 attraction? Systematic Biology 54, 731-742.

578 Wiens, J.J., Kuczynski, C.A., Townsend, T., Reeder, T.W., Mulcahy, D.G., Jr, J.W.S.,
579 2010. Combining phylogenomics and fossils in higher-level squamate reptile phylogeny:
580 molecular data change the placement of fossil taxa. Systematic Biology 59, 674-688.

581 Wiens, J.J., Moen, D.S., 2008. Missing data and the accuracy of Bayesian
582 phylogenetics. Journal of Systematics and Evolution 46, 307-314.

583 Wiens, J.J., Tiu, J., 2012. Highly incomplete taxa can rescue phylogenetic analyses from
584 the negative impacts of limited taxon sampling. PLoS ONE 7, e42925.

585 Yang, J.-B., Tang, M., Li, H.-T., Zhang, Z.-R., Li, D.-Z., 2013. Complete chloroplast
586 genome of the genus *Cymbidium*: lights into the species identification, phylogenetic
587 implications and population genetic analyses. BMC evolutionary biology 13, 84.

588 Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for *de novo* short read assembly
589 using de Brujin graphs. Genome Research 18, 821-829.

590 Zhang, Y.-J., Ma, P.-F., Li, D.-Z., 2011. High-throughput sequencing of six bamboo
591 chloroplast genomes: phylogenetic implication for temperate woody bamboos (Poaceae:
592 Bambusoideae). PLoS ONE 6, e20596.

593 Zwickl, D.J., Hillis, D.M., 2002. Increased taxon sampling greatly reduces phylogenetic
594 error. Systematic Biology 51, 588-598.

595

596

597 Figures

598 Figure 1: Bayesian phylogenetic reconstruction using (a) the alignment of Mishler *et al*.
599 (2014) and (b) the whole chloroplast genome alignments, of only the taxa present in both
600 studies. Numbers at nodes indicate posterior probabilities.

601

602 Figure 2: Phylogenetic trees of all 606 integrated specimens across (a) four chloroplast
603 loci and two nuclear ribosomal loci using RAxML (small amplicon sequence analysis); whole
604 chloroplast genomes for 96 individuals, and four chloroplast loci and two nuclear ribosomal
605 loci for 510 individuals analysed in a super matrix analysis using (b) RAxML or (c)
606 ExaBayes; and (d) four chloroplast loci and two nuclear ribosomal loci using the whole
607 chloroplast genome phylogeny (Appendix F) as a constraint.

608

609 Figure 3: Positions of the 15 minor clades within each of the integrated analyses,
610 including (a) the six locus small amplicon sequence tree; (b) the RAxML super matrix
611 analysis; (c) the ExaBayes super matrix analysis; and (d) the constraint tree, as well as (e) the
612 whole genome phylogeny. Values at nodes represent posterior probabilities in (c) and (e), and
613 maximum likelihood bootstrapping values in (a), (b) and (d). Solid lines indicate branches
614 with high (above 95% support) while dotted lines indicate lower support.

615

Table 1: Robinson-Foulds distances between each of the combined phylogenies calculated using HashRF.

	Amplicon	**Super (RAxML)**	**Super (ExaBayes)**	**Constraint**
Amplicon	0			
Super (RAxML)	166	0		
Super (ExaBayes)	216	200	0	
Constraint	144	150	217	0

Figure 1

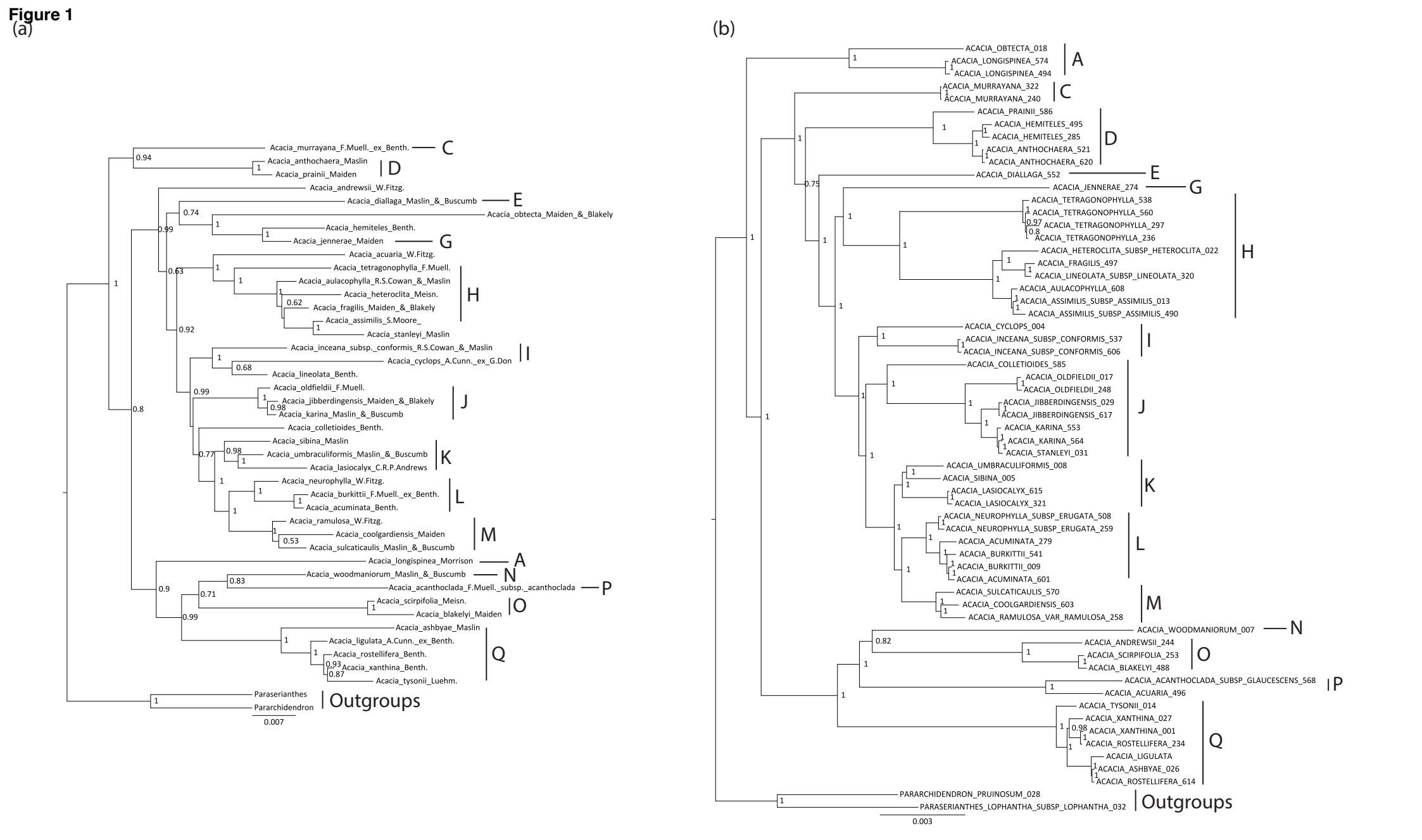

(a)
Acacia_murrayana_F.Muell._ex_Benth.
C
Acacia_anthochaera_Maslin
Acacia_prainii_Maiden
D
Acacia_andrewsii_W.Fitzg.
Acacia_diallaga_Maslin_&_Buscumb
E
Acacia_obtecta_Maiden_&_Blakely
Acacia_hemiteles_Benth.
Acacia_jennerae_Maiden
G
Acacia_acuaria_W.Fitzg.
Acacia_tetragonophylla_F.Muell.
Acacia_aulacophylla_R.S.Cowan_&_Maslin
Acacia_heteroclita_Meisn.
Acacia_fragilis_Maiden_&_Blakely
Acacia_assimilis_S.Moore_
Acacia_stanleyi_Maslin
H
Acacia_inceana_subsp._conformis_R.S.Cowan_&_Maslin
Acacia_cyclops_A.Cunn._ex_G.Don
Acacia_lineolata_Benth.
I
Acacia_oldfieldii_F.Muell.
Acacia_jibberdingensis_Maiden_&_Blakely
Acacia_karina_Maslin_&_Buscumb
J
Acacia_colletioides_Benth.
Acacia_sibina_Maslin
Acacia_umbraculiformis_Maslin_&_Buscumb
Acacia_lasiocalyx_C.R.P.Andrews
K
Acacia_neurophylla_W.Fitzg.
Acacia_burkittii_F.Muell._ex_Benth.
Acacia_acuminata_Benth.
L
Acacia_ramulosa_W.Fitzg.
Acacia_coolgardiensis_Maiden
Acacia_sulcaticaulis_Maslin_&_Buscumb
M
Acacia_longispinea_Morrison
A
Acacia_woodmaniorum_Maslin_&_Buscumb
N
Acacia_acanthoclada_F.Muell._subsp._acanthoclada
P
Acacia_scirpifolia_Meisn.
Acacia_blakelyi_Maiden
O
Acacia_ashbyae_Maslin
Acacia_ligulata_A.Cunn._ex_Benth.
Acacia_rostellifera_Benth.
Acacia_xanthina_Benth.
Acacia_tysonii_Luehm.
Q
Paraserianthes
Pararchidendron
Outgroups
0.007
(b)
ACACIA_OBTECTA_018
ACACIA_LONGISPINEA_574
ACACIA_LONGISPINEA_494
A
ACACIA_MURRAYANA_322
ACACIA_MURRAYANA_240
C
ACACIA_PRAINII_586
ACACIA_HEMITELES_495
ACACIA_HEMITELES_285
ACACIA_ANTHOCHAERA_521
ACACIA_ANTHOCHAERA_620
D
ACACIA_DIALLAGA_552
E
ACACIA_JENNERAE_274
G
ACACIA_TETRAGONOPHYLLA_538
ACACIA_TETRAGONOPHYLLA_560
ACACIA_TETRAGONOPHYLLA_297
ACACIA_TETRAGONOPHYLLA_236
ACACIA_HETEROCLITA_SUBSP_HETEROCLITA_022
ACACIA_FRAGILIS_497
ACACIA_LINEOLATA_SUBSP_LINEOLATA_320
ACACIA_AULACOPHYLLA_608
ACACIA_ASSIMILIS_SUBSP_ASSIMILIS_013
ACACIA_ASSIMILIS_SUBSP_ASSIMILIS_490
H
ACACIA_CYCLOPS_004
ACACIA_INCEANA_SUBSP_CONFORMIS_537
ACACIA_INCEANA_SUBSP_CONFORMIS_606
I
ACACIA_COLLETIOIDES_585
ACACIA_OLDFIELDII_017
ACACIA_OLDFIELDII_248
ACACIA_JIBBERDINGENSIS_029
ACACIA_JIBBERDINGENSIS_617
ACACIA_KARINA_553
ACACIA_KARINA_564
ACACIA_STANLEYI_031
J
ACACIA_UMBRACULIFORMIS_008
ACACIA_SIBINA_005
ACACIA_LASIOCALYX_615
ACACIA_LASIOCALYX_321
K
ACACIA_NEUROPHYLLA_SUBSP_ERUGATA_508
ACACIA_NEUROPHYLLA_SUBSP_ERUGATA_259
ACACIA_ACUMINATA_279
ACACIA_BURKITTII_541
ACACIA_BURKITTII_009
ACACIA_ACUMINATA_601
L
ACACIA_SULCATICAULIS_570
ACACIA_COOLGARDIENSIS_603
ACACIA_RAMULOSA_VAR_RAMULOSA_258
M
ACACIA_WOODMANIORUM_007
N
ACACIA_ANDREWSII_244
ACACIA_SCIRPIFOLIA_253
ACACIA_BLAKELYI_488
O
ACACIA_ACANTHOCLADA_SUBSP_GLAUCESCENS_568
ACACIA_ACUARIA_496
P
ACACIA_TYSONII_014
ACACIA_XANTHINA_027
ACACIA_XANTHINA_001
ACACIA_ROSTELLIFERA_234
ACACIA_LIGULATA
ACACIA_ASHBYAE_026
ACACIA_ROSTELLIFERA_614
Q
PARARCHIDENDRON_PRUINOSUM_028
PARASERIANTHES_LOPHANTHA_SUBSP_LOPHANTHA_032
Outgroups
0.003

Figure 2

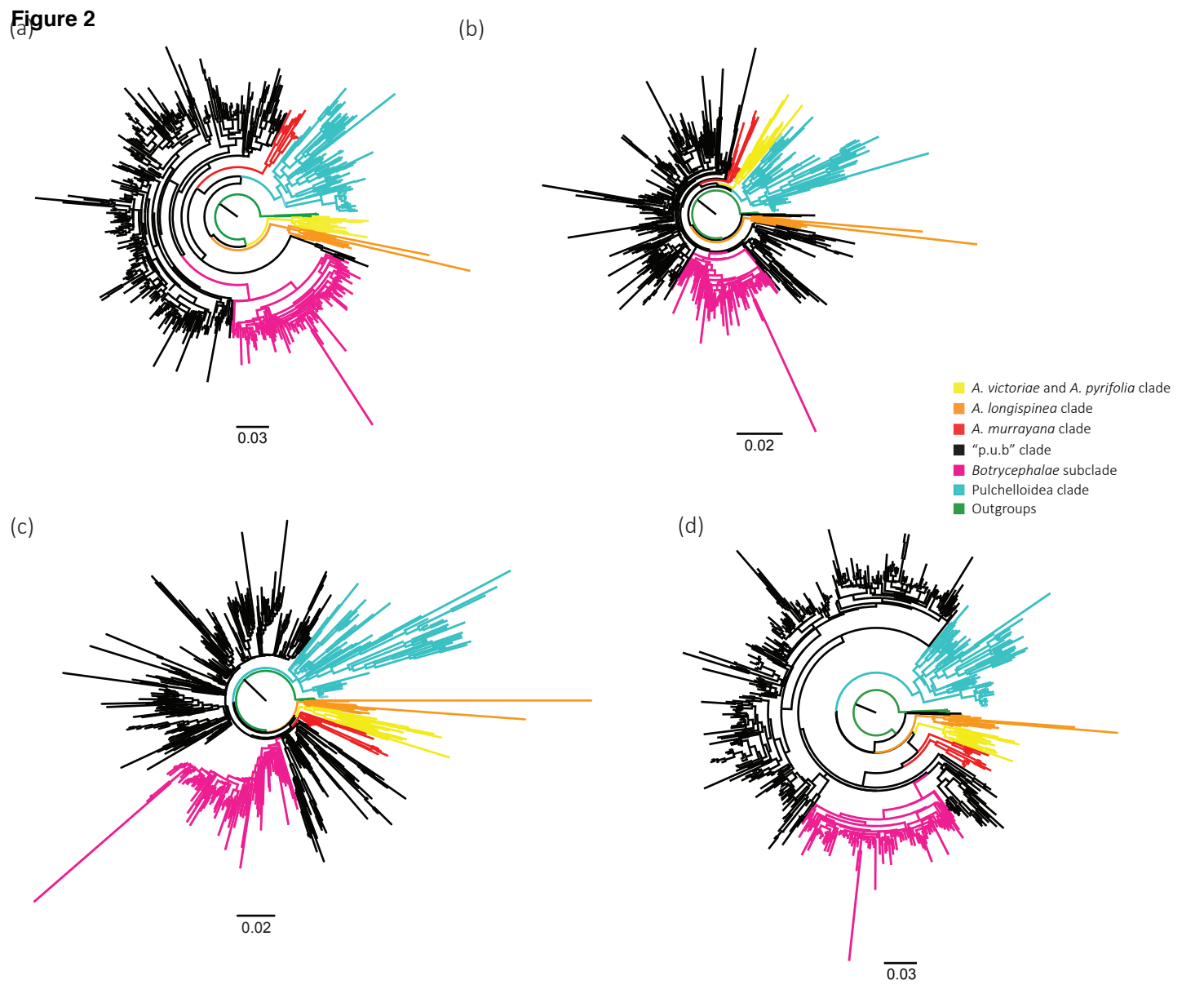
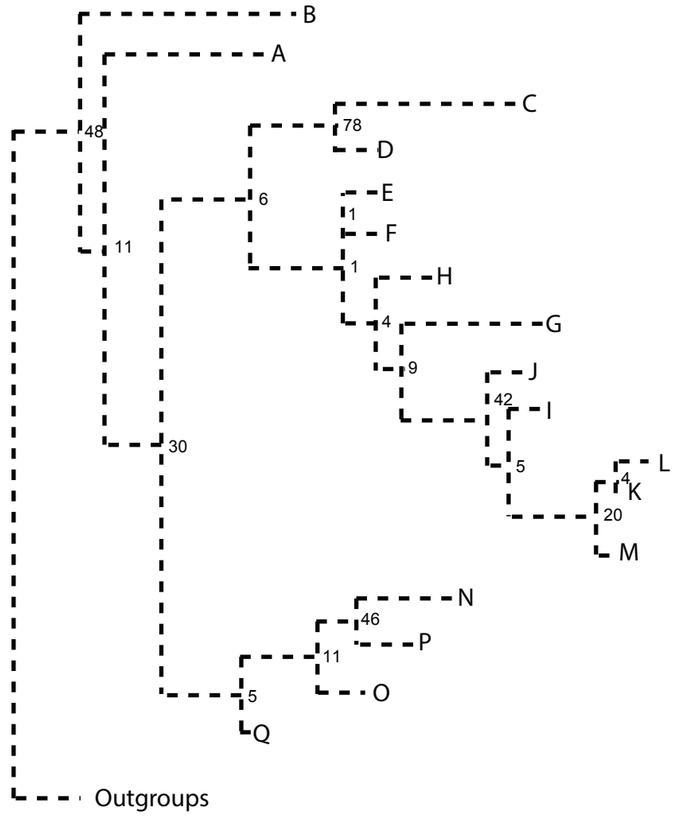
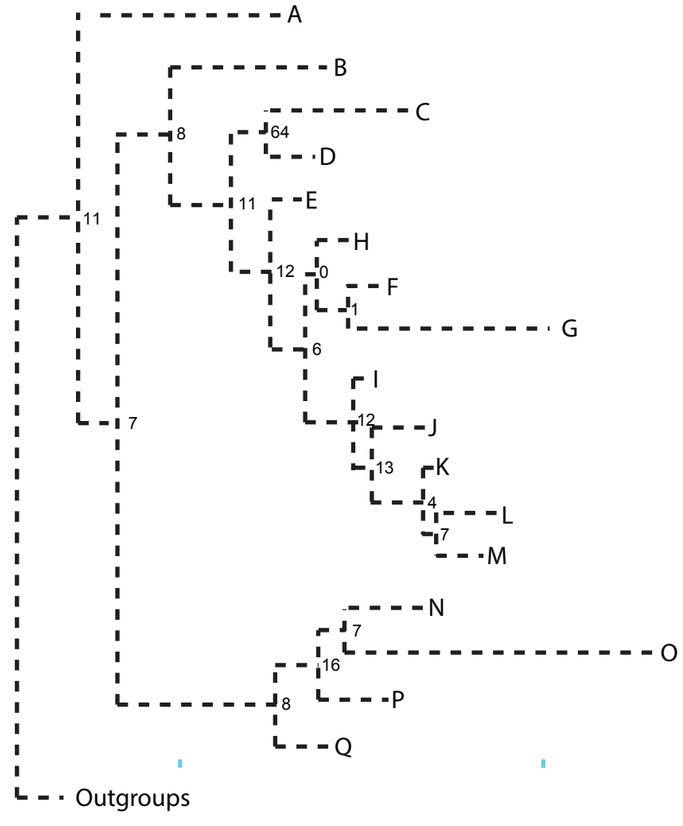

(a)
(b)
(c)
(d)
0.03
0.02
0.02
0.03
A. victoriae and A. pyrifolia clade
A. longispinea clade
A. murrayana clade
"p.u.b" clade
Botrycephalae subclade
Pulchelloidea clade
Outgroups

Figure 3

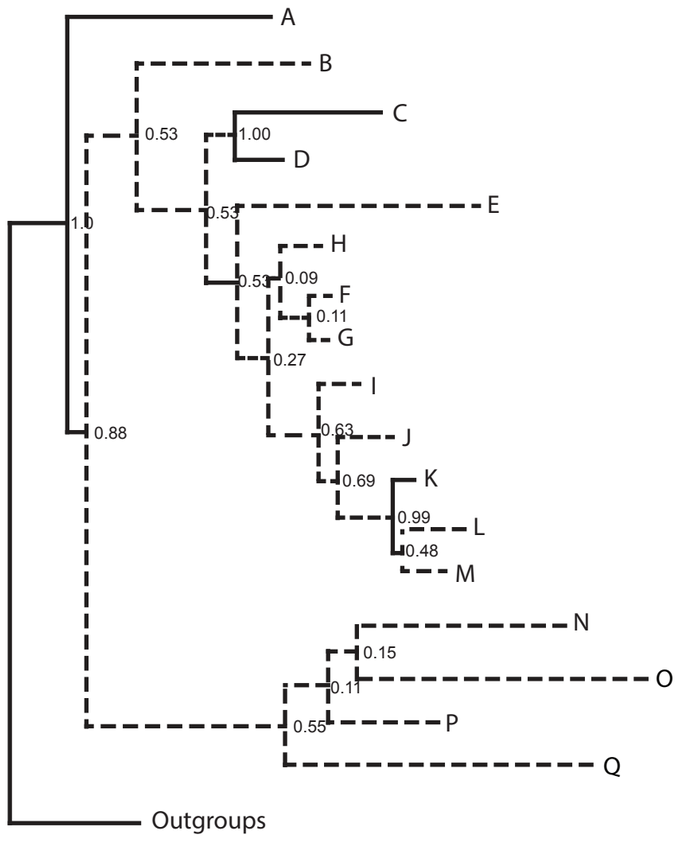
(a) Six gene small amplicon sequence tree



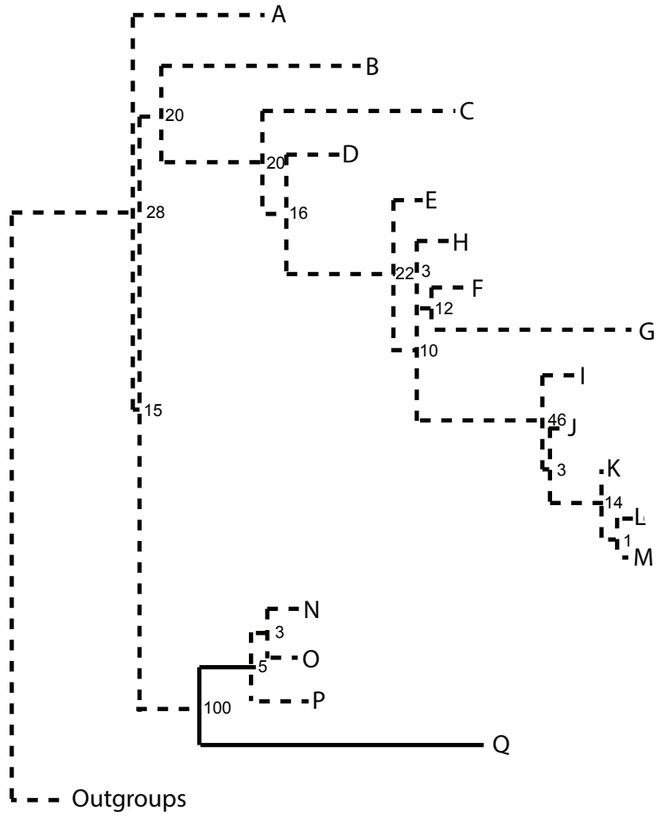
(b) Super matrix analysis (RAxML)



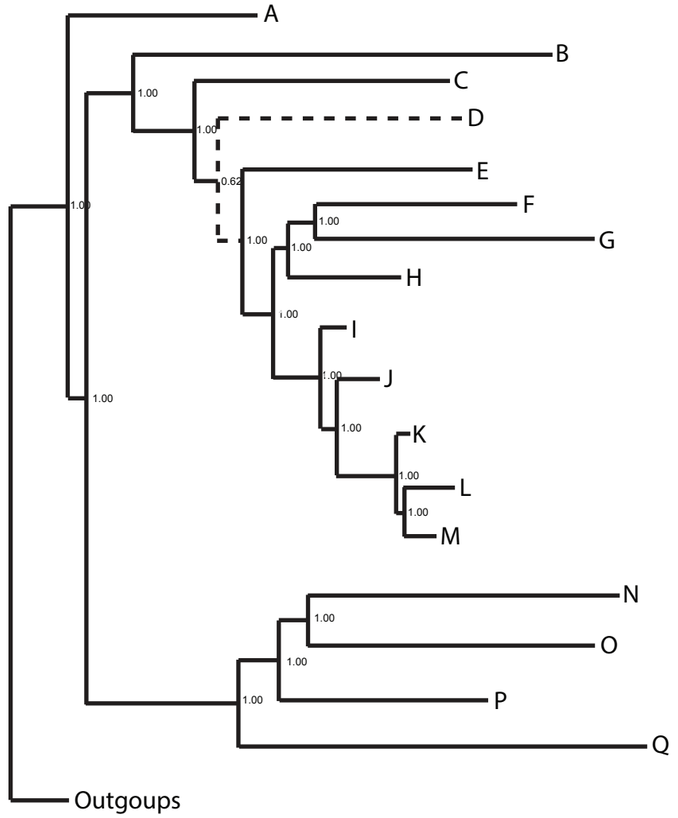
(c) Super matrix analysis (ExaBayes)



(d) Constraint tree



(e) Whole chloroplast genome tree



Appendix A: Specimens used in this study including collection locations and herbarium voucher numbers. Vouchers marked "PERTH" are held at the Western Australian Herbarium (Kensington, Western Australia) while all others are held at The University of Western Australia Herbarium (Crawley, Western Australia).

Species	Latitude	Longitude	Voucher number
Acacia acanthoclada subsp. *glaucescens* Maslin	-29.18979	116.95141	Williams 568
Acacia acuaria W.Fitzg.	-30.2853	116.9554	Williams 496
Acacia acuminata Benth.	-31.09257	120.6918	Williams 279
Acacia acuminata Benth.	-29.55589	116.90052	Williams 601
Acacia ampliata R.S.Cowan & Maslin	-28.583333	115.483333	PERTH 07018231
Acacia andrewsii W. Fitzg.	-27.45143	114.69132	Williams 244
Acacia anthochaera Maslin	-30.0688	117.425	Williams 521
Acacia anthochaera Maslin	-31.96458	115.83852	Williams 620
Acacia ashbyae Maslin	-31.95397	115.83672	Williams 026
Acacia assimilis S.Moore subsp. *assimilis*	-29.307028	116.730354	Williams 013
Acacia assimilis S.Moore subsp. *assimilis*	-30.2866	116.5924	Williams 490
Acacia aulacophylla R.S.Cowan & Maslin	-29.50072	116.99813	Williams 608
Acacia blakelyi Maiden	-30.3115	116.4497	Williams 488
Acacia burkittii Benth.	-29.070175	116.814011	Williams 009
Acacia burkittii Benth.	-29.7837	116.7762	Williams 541
Acacia cerastes Maslin	-29.677	117.02599	Williams 592
Acacia colletioides Benth.	-29.61772	116.96724	Williams 585
Acacia coolgardiensis Maiden	-29.51122	116.91828	Williams 603
Acacia cyclops G.Don	-31.99823	115.75253	Williams 004
Acacia daphnifolia Meisn.	-29.879167	116.03	PERTH 05689414
Acacia diallaga Madlin & Buscumb	-29.1497	116.96993	Williams 552
Acacia duriuscula W.Fitzg.	-29.68959	116.91246	Williams 589
Acacia effusifolia Maslin & Buscumb	-29.21036	116.663506	Williams 006
Acacia effusifolia Maslin & Buscumb	-29.196028	116.774028	Williams 030
Acacia eremaea C.R.P.Andrews	-30.367	117.1934	Williams 527
Acacia erinacea Benth.	-30.51345	121.38813	Williams 308
Acacia erinacea Benth.	-29.18909	116.94986	Williams 567
Acacia exocarpoides W.Fitzg.	-29.305696	116.732933	Williams 011
Acacia exocarpoides W.Fitzg	-31.96709	115.83752	Williams 621
Acacia formidabilis R.S.Cowan & Maslin	-29.51794	117.02118	Williams 611
Acacia fragilis Maiden & Blakely	-30.2853	116.9551	Williams 497
Acacia gibbosa R.S.Cowan & Maiden	-30.0973	117.3957	Williams 524
Acacia hemiteles Benth.	-31.10609	120.73764	Williams 285
Acacia hemiteles Benth.	-30.2853	116.9554	Williams 495
Acacia heteroclita Meisn. subsp. *heteroclita*	-32.549722	118.146667	PERTH

			06834914
Acacia inceana subsp. *conformis* R.S.Cowan & Maslin	-29.50682	116.9507	Williams 606
Acacia inceana subsp. *conformis* R.S.Cowan & Maslin	-30.3807	117.4111	Williams 537
Acacia jennerae Maiden	-31.2743	119.81621	Williams 274
Acacia jibberdingensis Maiden & Blakely	-30.0885	117.387222	Williams 029
Acacia jibberdingensis Maiden & Blakely	-31.9641	115.83834	Williams 617
Acacia karina Maslin & Buscumb	-29.14881	116.96901	Williams 553
Acacia karina Maslin & Buscumb	-29.19423	116.97187	Williams 564
Acacia kochii Ewart & Jean White	-29.318333	117.387667	PERTH 07435838
Acacia lasiocalyx C.R.P.Andrews	-31.22075	121.46321	Williams 321
Acacia lasiocalyx C.R.P.Andrews	-31.96373	115.83798	Williams 615
Acacia ligulata Benth.	-26.1445	121.077889	PERTH 07807864
Acacia lineolata Benth. subsp. *lineolata*	-31.2208	121.46406	Williams 320
Acacia longiphyllodinea Maiden	-30.4193	116.962	Williams 505
Acacia longiphyllodinea Maiden	-31.96424	115.83853	Williams 618
Acacia longispinea Morrison	-30.2853	116.9554	Williams 494
Acacia longispinea Morrison	-29.0807	116.90716	Williams 574
Acacia merrallii F.Muell.	-31.267	119.81605	Williams 272
Acacia merrallii F.Muell.	-30.274	116.6684	Williams 510
Acacia murrayana Benth.	-27.82636	115.39928	Williams 240
Acacia murrayana Benth.	-31.00187	121.27076	Williams 322
Acacia neurophylla subsp. *erugata* R.S.Cowan & Maslin	-27.64887	114.45508	Williams 259
Acacia neurophylla subsp. *erugata* R.S.Cowan & Maslin	-30.4285	116.9666	Williams 508
Acacia obtecta Maiden & Blakely	-30.021833	117.438972	PERTH 06876366
Acacia oldfieldii F.Muell.	-27.78858	114.46806	Williams 248
Acacia oldfieldii F.Muell.	-27.789167	114.466944	PERTH 06234194
Acacia prainii Maiden	-29.61753	116.96766	Williams 586
Acacia puncticulata Maslin	-27.75514	114.36212	Williams 256
Acacia ramulosa W.Fitzg. var. *ramulosa*	-27.64912	114.45499	Williams 258
Acacia resinimarginea W.Fitzg.	-31.09191	120.69183	Williams 281
Acacia resinimarginea W.Fitzg.	-30.3723	117.2687	Williams 530
Acacia resinimarginea W.Fitzg.	-29.61439	117.03455	Williams 594
Acacia resinosa R.S.Cowan & Maslin	-30.2853	116.9283	Williams 493
Acacia resinosa R.S.Cowan & Maslin	-29.51634	117.02502	Williams 612
Acacia restiacea Benth.	-30.4198	116.9622	Williams 506
Acacia restiacea Benth.	-31.96441	115.83857	Williams 619
Acacia rostellifera Benth.	-28.49665	114.62603	Williams 234
Acacia rostellifera Benth.	-29.52686	117.02173	Williams 614
Acacia scalena Maslin	-30.4328	116.9617	Williams 507

Acacia scirpifolia Meisn.	-27.74849	114.36269	Williams 253
Acacia scleroclada Maslin	-27.716722	117.089167	PERTH 07769776
Acacia sclerosperma F.Muell. subsp. *scleropsperma*	-27.82822	115.39806	Williams 242
Acacia sclerosperma F.Muell. subsp. *sclerosperma*	-30.2739	116.6684	Williams 509
Acacia sibina Maslin	-29.21036	116.663506	Williams 005
Acacia stanleyi Maslin	-30.088194	117.386056	Williams 031
Acacia stereophylla Meisn. var. *stereophylla*	-30.2854	116.9551	Williams 499
Acacia sulcaticaulis Maslin & Buscumb	-29.18542	116.97486	Williams 570
Acacia tetragonophylla F.Muell.	-28.49693	114.62574	Williams 236
Acacia tetragonophylla F.Muell.	-30.96193	121.1562	Williams 297
Acacia tetragonophylla F.Muell.	-30.4362	117.3859	Williams 538
Acacia tetragonophylla F.Muell.	-29.14643	116.9669	Williams 560
Acacia tysonii Luehm.	-29.260944	116.020167	PERTH 06876358
Acacia umbraculiformis Maslin & Buscumb	-29.188056	116.921056	Williams 008
Acacia uncinella Benth.	-31.0919	120.69177	Williams 280
Acacia websteri Maiden & Blakely	-30.95761	121.02514	Williams 301
Acacia woodmaniorum Maslin & Buscumb	-29.141117	116.883064	Williams 007
Acacia xanthina Benth.	-32.01546	115.76039	Williams 001
Acacia xanthina Benth.	-31.95417	115.83678	Williams 027
Acacia yorkrakinensis subsp. *acrita* R.S.Cowan & Maslin	-31.09177	120.69211	Williams 283
Acacia yorkrakinensis subsp. *acrita* R.S.Cowan & Maslin	-30.9586	117.1154	Williams 543
Pararchidendron pruinosum (Benth.) I.C.Nielsen	-31.955242	115.843003	Williams 028
Paraserianthes lophantha (Willd.) I.C.Nielsen subsp. *lophantha*	-31.917545	115.798813	Williams 032

Appendix B: ID number, species name, ENA accession number, number of reads produced using Illumina HiSeq2000 sequencing, number of contigs generated using Velvet, assembled length of the chloroplast genome and percentage identity with the *Acacia ligulata* reference chloroplast genome for each specimen used in this study.

#	Specimen	ENA accession	Number reads	Contigs	Assembled length (bp)	PI% with *Acacia ligulata*
001	*Acacia xanthina* Benth.	LN885329	3,830,703	40	174,359	98.4
004	*Acacia cyclops* G.Don	LN885258	1,971,156	36	175,320	92.8
005	*Acacia sibina* Maslin	LN885316	1,733,214	34	175,276	92.7
006	*Acacia effusifolia* Maslin & Buscumb	LN885262	1,317,856	43	175,367	92.1
007	*Acacia woodmaniorum* Maslin & Buscumb	LN885328	3,618,885	45	172,588	88.1
008	*Acacia umbraculiformis* Maslin & Buscumb	LN885325	2,400,060	39	175,596	92.6
009	*Acacia burkittii* Benth.	LN885253	2,298,474	34	174,711	91.3
011	*Acacia exocarpoides* W.Fitzg.	LN885267	1,711,739	43	173,733	87
013	*Acacia assimilis* S.Moore subsp. *assimilis*	LN885249	613,200	45	173,316	89.3
014	*Acacia tysonii* Luehm.	LN885324	3,173,818	50	176,254	97.7
015	*Acacia scleroclada* Maslin	LN885313	4,041,457	67	172,875	88.1
017	*Acacia oldfieldii* F.Muell.	LN885297	1,612,955	60	174,937	90.7
018	*Acacia obtecta* Maiden & Blakely	LN885296	694,025	42	175,857	91.1
021	*Acacia kochii* Ewart & Jean White	LN885282	1,386,421	47	173,440	91.6
022	*Acacia heteroclita* Meisn. subsp. *heteroclita*	LN885274	1,389,033	47	173,268	90
023	*Acacia daphnifolia* Meisn.	LN885259	2,895,801	52	174,886	90.5
024	*Acacia ampliata* R.S.Cowan & Maslin	LN885244	2,506,957	31	175,297	93.1
026	*Acacia ashbyae* Maslin	LN885248	2,466,871	39	174,020	98.5
027	*Acacia xanthina* Benth.	LN885330	2,792,320	42	175,889	97.2
028	*Pararchidendron pruinosum* (Benth.) I.C.Nielsen	LN885333	1,424,066	35	158,986	78.1
029	*Acacia jibberdingensis* Maiden & Blakely	LN885278	2,081,415	39	177,334	92
030	*Acacia effusifolia* Maslin & Buscumb	LN885263	2,122,879	30	176,478	92.7
031	*Acacia stanleyi* Maslin	LN885317	1,472,498	18	175,246	90.3
032	*Paraserianthes lophantha* (Willd.) I.C.Nielsen subsp. *lophantha*	LN885334	1,619,793	41	160,052	78.4
234	*Acacia rostellifera* Benth.	LN885309	2,182,983	45	176,285	96.6
236	*Acacia tetragonophylla* F.Muell.	LN885320	1,542,388	36	174,645	89.5
240	*Acacia murrayana* Benth.	LN885292	1,013,600	35	175,408	91.8
242	*Acacia sclerosperma* F.Muell. subsp. *scleropsperma*	LN885314	2,490,236	40	175,243	96.6
244	*Acacia andrewsii* W. Fitzg.	LN885245	1,607,420	35	176,784	92
248	*Acacia oldfieldii* F.Muell.	LN885298	1,695,383	34	174,797	90.2

253	*Acacia scirpifolia* Meisn.	LN885312	2,628,588	36	175,887	90.7
256	*Acacia puncticulata* Maslin	LN885300	1,172,986	25	173,905	88.9
258	*Acacia ramulosa* W.Fitzg. var. *ramulosa*	LN885301	2,578,531	34	175,238	92
259	*Acacia neurophylla* subsp. *erugata* R.S.Cowan & Maslin	LN885294	3,718,413	52	174,628	92.1
272	*Acacia merrallii* F.Muell.	LN885290	662,007	30	174,916	90
274	*Acacia jennerae* Maiden	LN885277	1,398,603	39	173,866	90.2
279	*Acacia acuminata* Benth.	LN885242	1,159,144	12	174,238	89.4
280	*Acacia uncinella* Benth.	LN885326	2,201,447	37	173,482	89.8
281	*Acacia resinimarginea* W.Fitzg.	LN885302	1,941,966	34	174,758	91.5
283	*Acacia yorkrakinensis* subsp. *acrita* R.S.Cowan & Maslin	LN885331	1,065,647	34	175,155	92.5
285	*Acacia hemiteles* Benth.	LN885272	2,322,134	37	175,055	91.6
297	*Acacia tetragonophylla* F.Muell.	LN885321	3,361,288	59	174,115	89.8
301	*Acacia websteri* Maiden & Blakely	LN885327	1,670,247	30	175,163	91.8
308	*Acacia erinacea* Benth.	LN885265	1,879,367	45	175,277	82.9
320	*Acacia lineolata* Benth. subsp. *lineolata*	LN885285	1,290,586	37	174,839	89.3
321	*Acacia lasiocalyx* C.R.P.Andrews	LN885283	1,176,444	37	174,493	91.3
322	*Acacia murrayana* Benth.	LN885293	1,323,170	32	175,712	92.4
488	*Acacia blakelyi* Maiden	LN885252	1,603,436	22	175,441	90.9
490	*Acacia assimilis* S.Moore subsp. *assimilis*	LN885250	978,448	42	175,226	88.9
493	*Acacia resinosa* R.S.Cowan & Maslin	LN885305	3,009,912	34	175,927	92.1
494	*Acacia longispinea* Morrison	LN885288	2,404,180	40	175,221	90.3
495	*Acacia hemiteles* Benth.	LN885273	2,071,909	35	173,964	91.5
496	*Acacia acuaria* W.Fitzg.	LN885241	1,821,446	33	173,782	86.3
497	*Acacia fragilis* Maiden & Blakely	LN885270	2,059,604	46	174,069	90
499	*Acacia stereophylla* Meisn. var. *stereophylla*	LN885318	780,668	35	174,719	91.8
505	*Acacia longiphyllodinea* Maiden	LN885286	2,014,232	39	175,190	91.5
506	*Acacia restiacea* Benth.	LN885307	3,671,253	45	173,222	87.6
507	*Acacia scalena* Maslin	LN885311	2,762,554	45	176,851	85.9
508	*Acacia neurophylla* subsp. *erugata* R.S.Cowan & Maslin	LN885295	1,744,125	38	174,679	91.7
509	*Acacia sclerosperma* F.Muell. subsp. *sclerosperma*	LN885315	1,547,970	22	175,368	96.2
510	*Acacia merrallii* F.Muell.	LN885291	1,094,617	34	174,397	91.9
521	*Acacia anthochaera* Maslin	LN885246	2,910,804	40	173,720	92.3
524	*Acacia gibbosa* R.S.Cowan & Maiden	LN885271	1,640,151	34	177,419	91.9
527	*Acacia eremaea* C.R.P.Andrews	LN885264	1,718,792	33	174,238	91.8
530	*Acacia resinimarginea* W.Fitzg.	LN885303	1,542,336	36	174,871	91.8

537	*Acacia inceana* subsp. *conformis* R.S.Cowan & Maslin	LN885275	405,245	41	175,011	90.6
538	*Acacia tetragonophylla* F.Muell.	LN885322	3,627,752	49	174,410	89.4
541	*Acacia burkittii* Benth.	LN885254	1,819,580	13	173,921	90
543	*Acacia yorkrakinensis* subsp. *acrita* R.S.Cowan & Maslin	LN885332	2,001,053	34	174,876	92.5
552	*Acacia diallaga* Madlin & Buscumb	LN885260	2,767,182	37	176,123	91.9
553	*Acacia karina* Maslin & Buscumb	LN885280	1,092,043	37	176,185	91.1
560	*Acacia tetragonophylla* F.Muell.	LN885323	1,219,136	34	174,985	89.3
564	*Acacia karina* Maslin & Buscumb	LN885281	637,643	37	175,058	90
567	*Acacia erinacea* Benth.	LN885266	990,844	47	174,732	83
568	*Acacia acanthoclada* subsp. *glaucescens* Maslin	LN885240	1,566,088	55	174,749	85.4
570	*Acacia sulcaticaulis* Maslin & Buscumb	LN885319	1,720,677	30	175,136	91.6
574	*Acacia longispinea* Morrison	LN885289	1,909,845	46	175,602	90.7
585	*Acacia colletioides* Benth.	LN885256	2,792,984	34	176,817	92.1
586	*Acacia prainii* Maiden	LN885299	2,579,014	39	175,472	91.7
589	*Acacia duriuscula* W.Fitzg.	LN885261	461,049		175,605	91.6
592	*Acacia cerastes* Maslin	LN885255	1,701,044	65	173,793	87.1
594	*Acacia resinimarginea* W.Fitzg.	LN885304	899,241	36	174,684	91.2
601	*Acacia acuminata* Benth.	LN885243	1,699,840	10	174,282	89.6
603	*Acacia coolgardiensis* Maiden	LN885257	633,187	36	174,741	91.8
606	*Acacia inceana* subsp. *conformis* R.S.Cowan & Maslin	LN885276	716,683	42	175,082	90
608	*Acacia aulacophylla* R.S.Cowan & Maslin	LN885251	2,468,372	43	173,215	88.2
611	*Acacia formidabilis* R.S.Cowan & Maslin	LN885269	1,185,416	37	173,894	90.1
612	*Acacia resinosa* R.S.Cowan & Maslin	LN885306	2,109,674	40	175,046	92
614	*Acacia rostellifera* Benth.	LN885310	1,657,690	41	175,208	96.5
615	*Acacia lasiocalyx* C.R.P.Andrews	LN885284	1,696,568	41	174,833	91.8
617	*Acacia jibberdingensis* Maiden & Blakely	LN885279	1,848,284	41	178,309	91.6
618	*Acacia longiphyllodinea* Maiden	LN885287	1,778,076	36	175,529	91.8
619	*Acacia restiacea* Benth.	LN885308	3,528,837	49	173,695	87.2
620	*Acacia anthochaera* Maslin	LN885247	1,435,546	42	173,093	92.1
621	*Acacia exocarpoides* W.Fitzg	LN885268	1,481,765	44	174,462	71

Appendix C

Click here to download Phylogenetic tree data: Appendix C.nwk

Appendix D

Click here to download Phylogenetic tree data: Appendix D.nwk

Appendix E

Click here to download Phylogenetic tree data: Appendix E.nwk

Appendix F

Click here to download Phylogenetic tree data: Appendix F.nwk

Appendix G

Click here to download Phylogenetic tree data: Appendix G.nwk