PHYLOGENETIC STUDIES OF EUCALYPTS: FOSSILS, MORPHOLOGY AND GENOMES

MICHAEL J. BAYLY
School of BioSciences, The University of Melbourne, Victoria 3010, Australia
Correspondence: Michael Bayly, mbayly@unimelb.edu.au

ABSTRACT: The eucalypt group includes seven genera: Eucalyptus, Corymbia, Angophora, Eucalyptopsis, Stockwellia, Allosyncarpia and Arillastrum. Knowledge of eucalypt phylogeny underpins classification of the group, and facilitates understanding of their ecology, conservation and economic use, as well as providing insight into the history of Australia’s flora. Studies of fossils and phylogenetic analyses of morphological and molecular data have made substantial contributions to understanding of eucalypt relationships and biogeography, but relationships among some genera are still uncertain, and there is controversy about generic circumscription of the bloodwood eucalypts (genus Corymbia). Relationships at lower taxonomic levels, e.g. among sections and series of Eucalyptus, are also not well resolved. Recent advances in DNA sequencing methods offer the ability to obtain large genomic datasets that will enable improved understanding of eucalypt evolution.

Keywords: Angophora, Corymbia, Eucalyptus, fossils, molecular phylogenetics

INTRODUCTION
Eucalypts are quintessential Australian plants. They make a significant contribution to Australian plant biodiversity and, because they dominate most terrestrial environments except extreme arid, alpine or saline habitats, they are integral to a broad range of plant communities (Groves 1994) and provide key resources for a diverse range of animals (Majer et al. 1997; Woinarski et al. 1997) and fungi (May & Simpson 1997). They also have important human uses, including paper, timber, honey and oil production.

There has been longstanding interest in the evolutionary relationships of eucalypts (e.g. Carr & Carr 1962; Johnson 1972). Knowledge of eucalypt relationships underpins modern classification schemes (e.g. Pryor & Johnson 1971; Hill & Johnson 1995; Brooker 2000; Ladiges & Udovicic 2000) and supports a broad range of other studies into eucalypts and the Australian biota more generally. Such studies include those of eucalypt metabolites (e.g. Merchant et al. 2006), plant community assembly (Pollock et al. 2015a), historical biogeography (e.g. Ladiges et al. 2003, 2011), molecular dating and biome evolution (Crisp et al. 2004; Ladiges & Udovicic 2005; Crisp et al. 2005; Thornhill et al. 2015), phylogenetic diversity and endemism (e.g. González Orozco et al. 2015), pathogen resistance and risk assessment for biosecurity (e.g. Potts et al. 2016), climate adaptation (e.g. Steane et al. 2014; Prober et al. 2016), hybridisation and eucalypt breeding (e.g. Griffin 1988; Grattagaglia & Kirst 2008; Dickinson et al. 2012; Larcombe et al. 2015) and diversity and evolution of eucalypt-associated scale insects (e.g. Cook 2001; Mills et al. 2016)

As an example of a relatively new application of phylogenetic information, McCarthy and colleagues incorporated phylogenetic diversity of eucalypts into spatial analysis for planning and prioritisation of conservation reserves, especially in Victoria (Pollock et al. 2015b; McCarthy & Pollock 2016). The aim of that work is to consider not just the numbers of species protected in conservation reserves, but also the diversity of phylogenetic lineages those species represent. Such applications of phylogenetic data highlight the need for greater resolution of eucalypt phylogenies.

This paper gives an overview of eucalypt diversity, classification, fossil history, and current knowledge of eucalypt phylogeny. It also considers the prospects of using new DNA sequencing technology to improve understanding of eucalypt relationships.

DIVERSITY AND TAXONOMY
Eucalypts are classified in family Myrtaceae and comprise the tribe Eucalypteae sensu Wilson et al. (2005). This tribe includes seven genera — Allosyncarpia S.T.Blake, Stockwellia Carr, Carr & Hyland, Eucalyptopsis C.T.White, Arillastrum Pancher ex Baill., Eucalyptus L’Hér. sensu strict., Angophora Cav. and Corymbia K.D.Hill & L.A.S.Johnson — and together are called the eucalypt group (Table 1, Figure 1). The first four genera are rainforest trees, including only five species, while the other three genera are sclerophylls including > 830 species (CHAH 2016). Eucalyptus is likely sister to the lineage of Angophora + Corymbia (Figure 2). Three of the rainforest genera form a monophyletic group (Figure
Table 1: Genera of the eucalypt group.

<table>
<thead>
<tr>
<th>Genus</th>
<th>No. of species</th>
<th>Distribution</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corymbia</td>
<td>c. 100</td>
<td>Australia (especially northern), New Guinea</td>
<td>Slee et al. (2006), CHAH (2016)</td>
</tr>
<tr>
<td>Angophora</td>
<td>9–10</td>
<td>Eastern Australia</td>
<td>Slee et al. (2006), CHAH (2016)</td>
</tr>
<tr>
<td>Alloosyncarpia</td>
<td>1</td>
<td>“Top End” of Northern Territory</td>
<td>Blake (1977)</td>
</tr>
<tr>
<td>Stockwellia</td>
<td>1</td>
<td>North Queensland: Atherton Tableland; Bellenden Ker Range</td>
<td>Carr et al. (2002)</td>
</tr>
<tr>
<td>Eucalyptopsis</td>
<td>2</td>
<td>Moluccas, New Guinea, Woodlark Island</td>
<td>White (1952), Craven (1990)</td>
</tr>
<tr>
<td>Arillastrum</td>
<td>1</td>
<td>New Caledonia</td>
<td>Dawson (1970)</td>
</tr>
</tbody>
</table>

*Upper end of range is based on CHAH (2016) list for Australia, but it does not include 33 species accepted in that list that are currently undescribed and listed only by informal phrase names.

Classification of *Eucalyptus*, *Corymbia* and *Angophora* has been a matter of contention in recent decades. The bloodwood eucalypts, genus *Corymbia*, were taxonomically split from *Eucalyptus* by Hill & Johnson (1995), on the grounds that bloodwoods are supported as more closely related to *Angophora* than to other members of *Eucalyptus* (e.g. Figure 2). Subsequently, Brooker (2000) rejected that split and used a much broader circumscription of *Eucalyptus*, sinking not only *Corymbia* but also *Angophora* to the rank of subgenera within it. That broad circumscription has not been followed in subsequent works (e.g. Ladiges & Udovicic 2000; Slee et al. 2006; CHAH 2016). However, other aspects of the classification of Brooker (2000), including circumscriptions of subgenera, sections and series, have been widely applied in a range of works that accept *Angophora*, *Corymbia* and *Eucalyptus* as separate genera (e.g. Steane et al. 2002; Slee et al. 2006; Parra-O et al. 2009).

As currently defined, *Eucalyptus* is by far the largest genus of the eucalypt group (Table 1) and includes ten subgenera (Slee et al. 2006). The three largest subgenera are *Symphyomyrtus*, *Eucalyptus* and *Eudesmia*. The other subgenera are much smaller, including only 1–4 species each, as follows: *Acerosae* (*E. curtisii*), *Cuboidea* (*E. temuipes*), *Idiogenes* (*E. cloeziana*), *Alveolata* (*E. microcorys*), *Cruciformes* (*E. guilfoylei*), *Minutifructus* (four species of tropical boxes, nested within *Symphyomyrtus*; Whittock et al. 2003) and *Primitiva* (*E. rubiginosa*, nested within subgenus *Eucalyptus* — Ladiges et al. 2010).

*Eucalyptus* and *Corymbia* are predominantly Australian genera, but a number of species naturally occur outside, or extend beyond, the Australian continent. *Eucalyptus deglupta* is endemic to Mindanao, Sulawesi, Ceram, New Guinea and New Britain. Three closely related species, *E. orophila*, *E. urophylla* and *E. wetarensis* (Pryor et al. 1995), sometimes treated as conspecific, under the name *E. urophylla* (Ladiges et al. 2003; Payn et al. 2007) are endemic to the Lesser Sunda Islands. Four species of *Eucalyptus* and five of *Corymbia* also have ranges that extend from northern Australia to New Guinea and, in the case of *E. alba*, also to the Lesser Sunda Islands (Chippendale 1988; Ladiges et al. 2003; Slee et al. 2006; Franklin & Preece 2014).

FOSSILS AND AGE OF THE EUCALYPT GROUP

Fossils of eucalypts have been reviewed or discussed by Rozefelds (1996), Ladiges (1997), Pole (1989, 1993), Byrne (2008), Gandolfo et al. (2011), Hermens et al. (2012) and Thornhill and Macphail (2012). Macrofossils of *Eucalyptus* extend their range historically to South America (Early Eocene — Gandolfo et al. 2011; Hermens et al. 2012) and New Zealand (Early Miocene — Pole 1989) reflecting a Gondwanan distribution.

Ladiges et al. (2003) and Ladiges & Cantrill (2007) argued, based on fossil evidence and the distributions of extant taxa, that the eucalypt group has origins in the Late Cretaceous, whereas Thornhill et al. (2015) applied molecular dating and estimated the group (all genera) is younger, dating to the Early Eocene. That molecular age estimate hinges on placement of the South American fossils as a calibration point. Exactly how those fossils (c. 52 million years old) were used in the analyses of Thornhill et al. (2015) is not clear because of inconsistencies in the text, i.e. their Figure 2 and Table S2 suggest the fossils were used as an age constraint for tribe Eucalypteae (the whole eucalypt group), whereas the placement of an asterisk in their Table 2 suggests the fossils were used to calibrate to the crown of a smaller group including *Eucalyptus*, *Arillastrum*, *Angophora* and *Corymbia* (i.e. excluding *Alloosyncarpia*, *Eucalyptopsis* and *Stockwellia*). Gandolfo
et al. (2011) considered the South American fossils were allied to *Eucalyptus* subgenus *Symphyomyrtus*, and Hermsen et al. (2012), although noting that ‘morphology of the fossils is also quite similar to some groups within extant *Corymbia* subgenus *Blakella’*, considered the evidence sufficient to formally name them as members of the genus *Eucalyptus*. If the fossils are members of *Eucalyptus* (or even *Corymbia*), then ages in the eucalypt group estimated by Thornhill et al. (2015) are underestimates, regardless of exactly how they applied the calibrations.

**MORPHOLOGICAL AND MOLECULAR PHYLOGENIES**


From the early 1990s, phylogenetic analyses have used DNA markers including chloroplast DNA restriction fragments (Sale et al. 1993) and 5S (Udovicic et al. 1995; Ladiges et al. 1995) and ITS (Steane et al. 1999, 2002) ribosomal DNA sequences. More recent works have employed a greater range of chloroplast and nuclear DNA sequences (Udovicic & Ladiges 2000; Poke et al. 2003, 2006; Whittock et al. 2003; Bayly & Ladiges 2007; Ochieng et al. 2007a; Bayly et al. 2008; Gibbs et al. 2009; Parra-O. et al. 2006, 2009), microsatellites (Steane et al. 2005; Ochieng et al. 2007b), AFLPs (McKinnon et al. 2008), and Diversity Array Technology (DArT) markers (Steane et al. 2011; Woodhams et al. 2013; Hudson et al. 2015; Rutherford et al. 2015).

Despite this progress, relationships among some eucalypt genera are still not clear, or robustly supported (Figure 2), and there is still controversy about generic circumscription of the bloodwood eucalypts. Since the genus *Corymbia* (the bloodwoods) was split from *Eucalyptus* (Hill & Johnson 1995), phylogenetic analyses have been contradictory as to whether *Corymbia* is monophyletic or paraphyletic (Udovicic et al. 1995; Steane et al. 1999, 2002; Udovicic & Ladiges 2000; Whittock et al. 2003; Parra-O. et al. 2006, 2009; Ochieng et al. 2007a,b). Furthermore, determining genetic relationships at the level of sections, series and species within clades has been challenging.

Obtaining molecular data from eucalypt genomes for phylogenetic study has been limited by the cost and effort involved in DNA sequencing, meaning that most studies have been based on very small amounts of data from DNA sequences or other molecular markers. The situation has changed radically in the last few years with the advent of ‘next-generation’ sequencing (NGS) technology. This offers unprecedented ability to obtain large genomic datasets at greatly reduced cost (e.g. Harrison & Kidner 2011; Straub et al. 2012; Hörandl & Appelhans 2015). Not only does this contribute to phylogenetic analysis but it will also allow discovery and interpretation of functional elements encoded within sequences, providing a basis for understanding key evolutionary changes that correlate with the diversification and adaptation of clades.

**GENOMIC ANALYSIS**

The structure and composition of eucalypt genomes has been of longstanding interest (Poke et al. 2005; Steane 2005; Byrne 2008; Grattapaglia & Kirst 2008; Grattapaglia et al. 2012) and research has progressed substantially with
Figure 3: Phylogenetic tree (maximum parsimony bootstrap consensus) produced by analysis of whole chloroplast genomes (adapted from Bayly et al. 2013). *Corymbia* is shown as paraphyletic, with members of subg. *Blakella* more closely related to *Angophora* than to subg. *Corymbia*. Members of the Symphyomyrt, Monocalypt and Eudesmid clades are members of subg. *Symphyomyrtus*, subg. *Eucalyptus* or subg. *Eudesmia* (Brooker et al. 2000), respectively, except where indicated. Parsimony bootstrap support values are given above branches.
development of NGS methods. A draft nuclear genome for *Eucalyptus grandis* W.Hill published by Myburg et al. (2014) was only the second published genome for any forest tree, after black cottonwood (*Populus trichocarpa* — Tuskan et al. 2006). Knowledge of the *E. grandis* genome has already paved the way for detailed studies of genes associated with wood (Carocha et al. 2015) and oil (Külheim et al. 2015) production, and for development of a new system for genotyping eucalypt individuals using genome-wide markers (Silva-Junior et al. 2015) that could be widely applied to population studies and tree breeding programs. These are just some examples of opportunities that come from having such genomic resources (Strauss & Myburg 2015).

In terms of phylogenetic studies of eucalypts, the largest published genomic dataset is that of Bayly et al. (2013). That analysed 41 complete chloroplast genomes, representing 39 different species from across the eucalypt group, including members of *Eucalyptus*, *Angophora*, *Corymbia*, *Allosyncarpia* and *Stockwellia*. That study was, at the time, the largest chloroplast dataset for any single plant genus (*Eucalyptus*). The phylogeny it resolved (Figure 3) broadly agreed with previous molecular phylogenetic studies but, being based on a dataset of >7000 variable characters, it provided better resolution of relationships and stronger support for nodes than those recovered from analyses of smaller molecular datasets. This phylogeny (Figure 3) showed *Corymbia* as paraphyletic, with subg. *Blakella* more closely related to *Angophora* than to the red bloodwoods classified in subg. *Corymbia* (represented only by *C. gummifera* in this study). This result has potential implications for generic classification of eucalypts, but it should be tested with further taxon sampling and sequence data before any formal changes to classification are considered; such further work is currently being supported by a research grant from the Dahl Trust.

There is good potential for further use of chloroplast genome sequences for resolving relationships of major clades of eucalypts, and even patterns of geographic variation within species, given the way in which chloroplast markers are already widely used to assess variation within or among closely related species (e.g. Byrne & Moran 1994; Byrne & Macdonald 2000; Freeman et al. 2001; Byrne & Hines 2004; Wheeler & Byrne 2006; Payn et al. 2007; Rathbone et al. 2007; Byrne & Hopper 2008; Nevill et al. 2010; Bloomfield et al. 2011; Gauli et al. 2014). Although the phylogenetic value of chloroplasts is limited by genome size (~160,000 base pairs; Bayly et al. 2013), a relatively slow mutation rate (Hörandl & Appelhans 2015), uniparental inheritance (maternal in eucalypts — Byrne et al. 1993; McKinnon et al. 2001a), and exchange between species through hybridisation and introgression (e.g. McKinnon et al. 2001b), they nonetheless provide useful insight into evolutionary history at a range of taxonomic levels, and their genomes are currently more readily sequenced than those of other organelles. Because chloroplast genomes in plant cells are present in high numbers when compared with nuclei (multiple chloroplasts to each nucleus, and each chloroplast containing multiple copies of the genome), relatively shallow sequencing of genomic DNA (which includes DNA from chloroplasts, nuclei and mitochondria) can be sufficient to assemble complete chloroplast genomes (e.g. Nock et al. 2011; Bock et al. 2014; Turner et al. 2016). This means that large numbers of chloroplast genomes can potentially be sequenced at relatively low cost (~$50–$80 per sample using current protocols — pers. obs.).

Making greater use of NGS technologies for phylogenetic studies of eucalypts will necessarily also involve development of larger datasets based on mitochondrial and nuclear DNA sequences. In general, mitochondrial genomes have been under-utilised in phylogenetic studies of plants, when compared with those of animals (e.g. Schaal et al. 1998). A range of factors has contributed to this, including the larger size of plant mitochondria, a propensity for common structural rearrangements in plant mitochondria (making genome assembly from ‘shotgun’ sequencing difficult in the absence of closely related reference genomes) and a relatively low mutation rate in plant mitochondrial genes (Palmer & Herbon 1988; Sloan et al. 2012; Christensen 2013; Wicke & Schneeweiss 2015). No assemblies of eucalypt mitochondria have yet been published, but some are in progress (J. Tibbits & P. Rigault pers. comm.) and should provide reference sequences that will allow more widespread use of mitochondrial data, acquired by relatively shallow sequencing of genomic DNA, in phylogenetic studies.

Nuclear genomes are by far the largest untapped resource for molecular phylogenetic studies of eucalypts and for plants in general. In eucalypts, haploid genomes range in size from 370 (*Corymbia citriodora*) to 710 (*Eucalyptus saligna*) million base pairs (Grattapaglia & Bradshaw 1994), providing a large number of potential sequence comparisons for phylogenetic analyses. NGS approaches reduce some previous impediments to utility of such data, e.g. the lack of universal primers for amplifying specific gene regions, difficulties of identifying sequences that are equivalent (homologous) within multigene families, and some difficulties in separating allelic variants of genes. However, assembly of whole genomes, even with a reference for *Eucalyptus grandis* (Myburg et al. 2014) is still an involved task and requires much deeper sequencing (more financial cost) than the ‘skimming’ of high copy chloroplast and mitochondrial data from lower
effort sequencing. A range of methods based on sequencing small proportions of genomes (‘reduced representation libraries’) probably show the greatest immediate promise as sources of nuclear data for phylogenetic analyses. Approaches variously known as Restriction-Associated DNA sequencing (RAD-seq or double digest RAD-seq — Rowe et al. 2011; Petersen et al. 2012), Genotyping By Sequencing (GBS — Elshire et al. 2011) or Diversity Array Technology sequencing (DART-seq — Sansaloni et al. 2011) identify sequence variation in thousands of essentially random loci across the genome at relatively low cost, and are starting to be trialled in eucalypts, albeit in studies of one to a few species at this stage (Grattapaglia et al. 2011; Steane et al. 2014), but have potential to be applied more broadly. The same is true of methods using ‘exome capture’ that deliver sequences of specific or known genes (e.g. Dasgupta et al. 2015).

Further use of nuclear DNA sequences will be critical not only to resolving or testing higher-level phylogenetic relationships among eucalypts (e.g. among currently recognised genera and subgenera), but also for understanding relationships among closely related species. Existing phylogenetic markers are not sufficiently variable to do this, and a growing range of studies, based mostly on chloroplast data, has highlighted complex patterns of hybridisation and introgression between related species that severely limit phylogenetic reconstructions based on uniparentally inherited organellar markers (e.g. Steane et al. 1998; Jackson et al. 1999; McKinnon et al. 1999, 2001b, 2004, 2010; Pollock et al. 2013; Nevill et al. 2014). Use of large numbers of nuclear DNA markers will allow greater insight into the relationships between species and the extent of genetic interaction between them. Such insight will no doubt underpin further refinements to eucalypt classification as well as studies of eucalypt ecology, biogeography, conservation, adaptation and breeding etc. that require knowledge of phylogenetic relationships.

Acknowledgements
I thank the Dahl Trust and Royal Society of Victoria for the invitation to participate in the symposium to which this paper relates. I also thank the Dahl Trust for funding to support research on eucalypt phylogenetics. Pauline Ladiges gave insightful advice on the content of this paper and on the structure of the preceding oral presentation. The text was also improved by constructive advice through anonymous peer review. Andrew Drinnan and Pauline Ladiges also gave permission to use their photographs in Figure 1.

References
Byrne, M. & Hopper, S.D., 2008. Granite outcrops as ancient islands in old landscapes: evidence from the phylogeography and population genetics of Eucalyptus
caesia (Myrtaceae) in Western Australia. *Biological Journal of the Linnean Society* 93: 177–188.


Christensen, A.C., 2013. Plant mitochondrial genome evolution can be explained by DNA repair mechanisms. *Genome Biology and Evolution* 5: 1079–1086.


Wicke, S. & Schneeweiss, G.M., 2015. Next-generation organellar genomics: Potentials and pitfalls of high-

