**Using a low-copy nuclear gene (phosphoglycerate kinase; PGK) to explore the phylogeny of the aquatic plant family Hydatellaceae (Nymphaeales)**

by

Qianshi Lin

B.Sc., Fudan University, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Botany)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

September 2014

## ABSTRACT

Hydatellaceae are a small aquatic family of 12 species related to water lilies, part of the ANITA grade of angiosperms. Our current understanding of phylogenetic relationships in the family comes from several plastid genes and nuclear ITS data. These data sets are generally highly congruent, and lend support to the monophyly of multiple species. However, the published nuclear ITS tree was unrooted (outgroups were too distant to align), and there were several minor phylogenetic conflicts between plastid and ITS gene trees for three closely related species, *Trithuria bibracteata*, *T. occidentalis*, and *T. submersa*; two of these species were also not reciprocally monophyletic in individual gene trees. The position of *T. occidentalis* was also based on very limited plastid data, and there was no molecular evidence to link staminate and pistillate individuals in this species. To further clarify phylogenetic relationships and species boundaries, I recovered two copies of nuclear-encoded phosphoglycerate kinase (*PGK*) gene from taxa in Hydatellaceae and several water lilies. I reconstructed the history of the *PGK* duplication in angiosperms as a whole. I also added plastid data from additional populations of several species, and estimated the dated species tree using a Bayesian multispecies coalescent approach to reconcile different gene trees. The angiosperm-level *PGK* gene tree indicated that the duplication of *PGK* gene may have happened around the origin of angiosperms. The root of Hydatellaceae implied by concatenated nuclear *PGK* matches that inferred from plastid data. *Trithuria occidentalis* is clearly placed in sect. *Trithuria*, and staminate and pistillate individuals of this species are linked together using new evidence from the plastid and *PGK* genes. Phylogenetic relationships inferred using each *PGK* copy are consistent with the sectional relationships inferred using plastid and ITS data, with less sharply defined species boundaries. I also explore the possibility here that some of the incongruence that I observed between individual genes trees and in inferred species trees is a consequence of additional minor gene duplications or polyploidization/introgression events.

## PREFACE

This project included assistance from Will Iles and Isabel Marques in the Graham lab at UBC; Marques also contributed new material of several species of Hydatellaceae. Sequences of Hydatellaceae are mainly from extractions of Will Iles, except for *Trithuria australis* (*MARQ13001 & Macfarlane, MARQ13002 & Macfarlane*), *Trithuria bibracteata* (*MARQ13003 & Macfarlane, MARQ13004 & Macfarlane*), *Trithuria inconspicua* (*MARQ13005 & Macfarlane, MARQ13006 & Macfarlane*), *Trithuria submersa* (*MARQ13007 & Macfarlane, MARQ13008 & Macfarlane*), which come from collections of Isabel Marques and collaborators from Australia. I used alignments of four plastid genes from Will Iles (2012, 2014) in phylogenetic analysis of plastid genes. Will Iles also helped me to perform the dating analysis and species-tree inference using *BEAST.

# TABLE OF CONTENTS

# LIST OF TABLES

## LIST OF FIGURES

# ACKNOWLEDGEMENTS

I offer my enduring gratitude to my supervisor, Sean Graham; other committee members: Jeannette Whitton and Wayne Maddison also guided me to finish my program successfully. I also thank my lab members: Will Iles, Isabel Marques, Vivienne Lam, Gregory Ross, Marybel Soto Gomez, David Bell, Hayley Darby and Erin Fenneman, who helped me a lot in the lab and provided coherent answers to my endless questions. Thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC) who helped to fund my research through a Discovery grant to Sean Graham.

Special thanks are owed to my parents, who have supported me throughout the years of my Master's degree.

# 1. INTRODUCTION

Hydatellaceae are a small aquatic family of 12 species related to water lilies, part of the ANITA grade (the grade of five families within which all other angiosperms are nested) of angiosperms (Saarela et al., 2007; Sokoloff et al., 2008). The family is found in Australia, New Zealand, and India, and has primary species diversity in Australia (Yadav and Janarthanam, 1994; Sokoloff et al., 2008; Iles et al., 2014). Traditionally it was placed with Centrolepidaceae, which is a reduced grass-like monocot family related to the southern rushes (Hieronymus, 1888; Gilg-Benedict, 1930). However, Saarela et al. (2007) demonstrated that it should be removed from monocots to the ANITA grade of angiosperms, based on plastid, nuclear and morphological data. Subsequent research confirmed this new placement using additional plastid and mitochondrial data (Graham and Iles, 2009; Qiu et al., 2010; Moore et al., 2011; Soltis et al., 2011). Hydatellaceae are now placed in an expanded order Nymphaeales, as the sister group of Cabombaceae and Nymphaeaceae (Rudall et al., 2007; APG, 2009). This new placement defines a very early split in angiosperm phylogeny (Iles et al., 2014), and has sparked renewed interest in and study of the family's morphology, reproduction, ecology, evolution and systematics (e.g., Rudall et al., 2007, 2009a,b; Friedman, 2008; Remizowa et al., 2008; Sokoloff et al., 2008a,b, 2011, 2013, 2014; Taylor et al., 2010, 2012; Iles et al., 2012, 2014; Costa et al., 2013; Friedman et al., 2012, 2013).

Two genera were traditionally recognized in the family: *Trithuria* Hook. f. and *Hydatella* Diels. This division was based on features of the fruit and reproductive units (RUs; a term used for the compacted floral structures in the family because of uncertainty in whether they are flowers, inflorescences or a pre-floral structure). As classically

defined, the genus *Trithuria* had three ribs on fruit (pericarp) and bisexual RUs, whereas *Hydatella* had no pericarp ribs and unisexual RUs (Cooke, 1987; Hamann, 1998). However, in 2008, Sokoloff et al. accepted only one genus in this family: of four new species in the family, two species (*T. cowieana* and *T. polybracteata*) combine characters from both genera, conflicting with a morphological separation of the family into the two genera. They also found that the species *H. dioica* D. A. Cooke (known only from staminate plants) is conspecific with *T. occidentalis* Benth (known only from pistillate plants) based on the observation of a physical connection between a seed with features of *T. occidentalis* and a single plant of *H. dioica*.

Iles et al. (2012; 2014) recently produced a molecular phylogeny of this family based on four plastid genes (*atp*B, *rbc*L, *mat*K and *ndh*F) and the nuclear ITS region (the two internal transcribed spacer regions between nuclear rDNA genes). Most species-level relationships in this family were inferred with high support, and the resulting phylogeny is consistent with a new single-genus, four-section classification (Sokoloff et al., 2008; Iles et al., 2014). This classification reflects a deep geographic split in the family, as two sections (*Altofinia* and *Hamannia*) together represent a tropical clade of the northern Australian and Indian species, and the other two (*Hydatella* and *Trithuria*) include all the species in southern Australia and New Zealand. Iles et al. (2014) related the division of these two clades to a vicariant event associated with the drying out of central Australia in the Miocene. However, they rejected a vicariant explanation for the disjunct placement of the Indian species (*T. konkanensis*) from those in Australia, as this species is closely related to species in northern Australia. They also predicted that long-distance dispersal plays a major role in the distribution and dispersal of the remaining species, based on

additional conflicts between predicted vicariance events and the molecular dates of associated phylogenetic splits.

Several additional problems remain to be addressed. The first is that the existing nuclear-based data provide no information on the phylogenetic root of the family (the nuclear ITS data for the family are not readily alignable to those in outgroup taxa). Additional nuclear data would therefore be useful to address this issue. Iles et al. (2012, 2014) also found relatively minor phylogenetic conflicts between trees inferred from the current plastid and ITS data (Iles et al. 2012). Specifically, gene trees for *T. submersa* and *T. bibracteata* intermingled in different ways in the trees inferred from plastid vs. ITS data. In the plastid tree, some specimens of *T. bibracteata* formed a sister group to other *T. bibracteata* and *T. submersa* sequences. In contrast, in the ITS tree, a conflicting set of *T. bibracteata* and several *T. submersa* sequences were the sister group of the remaining *T. submersa* samples. Both sides of the conflict were well-supported. In the species tree inferred using a Bayesian multispecies coalescent approach, the relationship between these two species was poorly supported (Iles et al., 2012, 2014). This kind of non-monophyly and incongruence may result from incomplete lineage sorting, hybridization or polyploidization (e.g., Maddison, 1997). Because *T. submersa* is likely a polyploid (possibly an allopolyploid; Kynast et al., in press), reticulate evolution may also have occurred in this family. These conflicts may be amenable to study by examining additional gene trees. In addition, as the phylogenetic position of the rare dioecious species *T. occidentalis* is based on only very limited data (a very short sequence of *mat*K), and there is no molecular evidence linking the pistillate and staminate individuals for it

(the staminate plants were previously recognized as *H. dioica*), it would be useful to obtain additional phylogenetic information for this species.

All plastid genes are part of the same genetic linkage group (the plastid genome), and so they do not provide mutually independent evidence of lineage sorting or hybridization events (e.g., Doyle, 1992). Different nuclear genes may provide independent information about these aspects of a species' evolutionary history (e.g., Small et al., 2004; Strand et al., 1997). However, the fluidity of gene copy number in many nuclear genes can also make phylogenetic inferences challenging because of difficulties in orthology assessment. Finding single-copy or low-copy nuclear genes in plants may be useful in this situation (Steele et al. 2008; Duarte et al. 2010; Regier et al. 2010). Low-copy duplicated genes have become a very useful tool for investigating tree rooting issues and to clarify the process of hybridization and polyploidization in plant evolutionary history (e.g. Mathews et al. 1999, 2000; Ness et al., 2011; Adderley et al., 2014). However, we have relatively limited genomic information from ANITA-grade taxa: *Amborella* has a fully sequenced nuclear genome (*Amborella* Genome Project, 2013), but has an estimated evolutionary separation of ~150 Ma or more from Hydatellaceae (Iles et al. 2014), limiting our ability to develop new low-copy nuclear gene markers in ANITA-grade taxa.

I examined several ~10 low-copy nuclear genes as candidate nuclear phylogenetic markers by developing primers for them (including the commonly used markers *PHYA*, *PHYC*, *LEAFY*), but most were not successful in amplification (I developed candidate primers using a *Trithuria* transcriptome, see below, and tested them using *Nuphar polysepala*, Nymphaeaceae, as I had limited amounts of *Trithuria* DNA available for test

reactions). I finally focused on one duplicated locus that amplified well in *Trithuria*, phosphoglycerate kinase (*PGK*), which has also been used for the inference of grass phylogeny (Huang et al., 2002; Chen et al., 2013; Adderley and Sun, 2014). Phosphoglycerate kinase (*PGK*) is a housekeeping enzyme involved in the Calvin cycle and glycolysis. Anderson and Advani (1970) first noted that pea has two functionally divergent copies of the *PGK* protein, plastid and cytosolic isoenzymes. Following Huang et al. (2002), I refer to these as *PGK*-1 (the locus that codes for the nuclear-encoded plastid enzyme) and *PGK*-2 (the cytosolic enzyme). The two forms were derived from an ancient duplication, following the movement of the enzyme from the plastid to the nuclear genome (Longstaff et al. 1989; Brinkmann and Martin, 1996; Martin and Schnarrenberger, 1997). In this study, I also use the term "copy" to refer to these two functionally divergent copies of *PGK*, and "variant" to describe other species-level variation present within copies.

In this study, I sampled the nuclear *PGK* loci for all 12 Hydatellaceae species and outgroups, with multiple samples included per species, where possible. I also added new samples from four species (*T. occidentalis*, *T. australis*, *T. submersa* and *T. bibracteata*) to previously published plastid data (Iles et al., 2012, 2014). I used these new data to address the following questions: (1) to place the duplication history of the *PGK* locus in Hydatellaceae in the overall context of other angiosperms; (2) to investigate the diversity of sequences in the genomes of each species in a phylogenetic context; (3) to apply these data to species-tree inference, and to look for evidence of trans-specific polymorphism that may not be assignable to lineage-sorting events (e.g., Percy et al., in press); (4) to obtain nuclear data relating to the root of Hydatellaceae phylogeny. I was also interested

in adding new data to the existing plastid-based phylogeny of the family, to find molecular evidence to more firmly place *T. occidentalis* in a local phylogenetic context, and to link staminate and pistillate individuals in this species.

## 2. MATERIALS AND METHODS

### 2.1 Sampling

I sampled a portion of the *PGK* (phosphoglycerate kinase) gene for all 12 currently recognized species of Hydatellaceae (30 individuals in total). I included multiple individuals per species (except *T. cookeana, T. cowieana, T. filamentosa, T. polybracteata* and *T. konkanensis*). In six species (*T. austinensis, T. australis, T. bibracteata, T. inconspicua, T. lanterna, T. submersa*), I sampled two-four populations and one-two individuals per population (Appendix 1). I sampled *PGK* for four outgroup taxa: *Amborella, Brasenia, Cabomba* and *Nuphar* (one individual per species). I also added plastid data (*atp*B, *rbc*L, *mat*K and *ndh*F) for eight individuals to the data set of Iles et al. (2012), which came from new field collections in 2013 for four species (*T. australis, T. occidentalis, T. bibracteata* and *T. submersa*), by Isabel Marques (UBC). These included an additional population for *PGK* from *T. australis* that appears to be morphologically distinct from other populations (D. Sokoloff, pers. comm.), and two samples (a staminate and a pistillate individual) from the dioecious and endangered species *T. occidentalis*. I was unable to recover the nuclear ITS region for these samples due to fungal contamination.

## 2.2 Extraction, amplification and cloning

I extracted total genomic DNA from silica-gel dried leaf material using standard protocols (Doyle & Doyle, 1987). I designed primers for the phosphoglycerate kinase (*PGK*) gene between two exons spanning the fourth intron (see Fig. 1) considering a transcriptome of *T. submersa* from M. Barker, L. Rieseberg and S. W. Graham (unpublished data). The corresponding sequenced region is ~350 bp in length, including 241 bp (unaligned length of *T. submersa*, for reference) of exonic sequence. The sequence of the forward primer is: 5'-TCAAAGGTSTCRTCCAAGATTG. The sequence of the reverse primer is: 5'-CAAGTCCCATCCAWCCATCAG.

Primers for plastid genes were based on the following publications: *atp*B (Hoot et al., 1995), *mat*K (Hilu et al., 2003; Löhne et al., 2007; www.kew.org/barcoding/update.html), *ndh*F (Olmstead and Sweere, 1994; Kim and Jansen, 1995), *rbc*L (Terachi et al., 1987; Yamashita and Tamura, 2000; G. Zurawski, Baylor University, personal communication). I amplified these regions using the polymerase chain reaction (PCR) under conditions described in Graham and Olmstead (2000), but using Phusion polymerase (New England Biolabs). I carried out cycle sequencing using BigDye 3.1 chemistry (Applied Biosystems, Foster City, CA, USA). To distinguish different copies of *PGK*, I obtained at least eight successful clones total per individual using the TOPO TA Cloning Kit (Invitrogen Life Science Technologies).

## 2.3 Alignment and phylogenetic analysis

I used Sequencher 4.2.2 (Gene Codes Corp., Ann Arbor, Michigan, USA) to base-call and assemble contigs. I aligned the DNA sequences using MUSCLE (Edgar, 2004), with

manual adjustment in Se-Al v2.0a (Rambaut, 2002), following criteria in Graham et al. (2000). I analyzed the aligned *PGK* matrix using heuristic maximum-parsimony (MP) searches with PAUP* vers. 4.0b10 (Swofford 2003), using tree-bisection-reconnection (TBR) branch swapping, and 200 random-addition replicates, and otherwise using default settings. Garli version 2.0 (Zwickl, 2006) was used to perform a partitioned maximum likelihood (ML) analysis, using PartitionFinder v1.1.1 (Lanfear et al., 2012; Lanfear et al., 2014), and considering the AICc (the Akaike Information Criterion, correcting for sample size) to determine the optimal DNA substitution model and partitioning scheme for: (i) different codon positions in plastid genes for the plastid matrix; (ii) the intron and exonic sequences considered as two separate data partition for the species-level matrix for *PGK*; (iii) different codon positions in the angiosperm-wide matrix for the full *PGK* coding sequence (see below). The models and partitioning scheme can be found in Appendices 2a-c. I ran the ML search with 20 search replicates to find the best ML tree. I used bootstrapping (Felsenstein, 1985) with 200 bootstrap replicates to estimate branch support for ML and MP analyses. I employed MrBayes v2.01 (Huelsenbeck and Ronquist, 2001) to run Bayesian analysis (using the same partitions as in the ML analyses). MrBayes uses a Markov chain Monte Carlo (MCMC) sampling approach: I ran 1,000,000 generations and sampled every 100 generations, and evaluated the result in Tracer version 1.5 (Rambaut and Drummond, 2009). I discarded 10% burnin from each chain. Estimated sample sizes (ESSs) exceeding 200 were found, taken to indicate strong chain convergence. Following Zgurski et al. (2008), I considered well-supported branches to have bootstrap support of ~90% or more, and poorly supported branches to have ~70% or less bootstrap support. For Bayesian analysis, I considered well-supported branches to

have 0.98 Posterior probability (PP) or more, moderate-supported branches to have 0.95-0.98 PP, weak-supported branches to have less than 0.95 PP (see Alfaro and Holder, 2006).

**2.4 Species-tree inference**

To estimate the species tree in an incomplete lineage sorting framework, I used a Bayesian multispecies coalescent approach, *BEAST (Heled and Drummond, 2010). This approach assumes that any incongruence among gene trees is a consequence of incomplete lineage sorting. I included plastid and nuclear ITS data matrix from Iles et al. (2012) in this analysis. Because geographic separation (>1500 km; estimated from Sokoloff et al.) may block gene flow in continentally disjunct populations of *T. submersa*, I provisionally treated eastern and western populations of this species as distinct species, following Iles et al. (2012, 2014). All loci were assigned GTR+Γ substitution models and uncorrelated lognormal clocks. A mean clock rate of 1.0 was assigned to all loci. To date the species tree, I assigned prior distributions to two nodes in Hydatellaceae using posterior ages for these nodes that were determined in a seed-plant wide analysis by Iles et al. (2014). The two calibrated nodes are the crown clade of the family Hydatellaceae, and a node representing the crown clade comprising sections *Hydatella* and *Trithuria*. Setting these priors effectively forces a root for the family, which derives in turn from the plastid-based trees in Iles et al. (2012, 2014). Four independent trials were run for 1,000,000 generations, sampling every 1000. The first 10% of each trial was burned in. ESSs exceeding 200 were found, taken to indicate strong chain convergence.

### 2.5 Reconstructing the history of the PGK duplication

To characterize the *PGK* gene duplication in Hydatellaceae in the context of angiosperm phylogeny I obtained *PGK* exonic sequences from other angiosperms (and a gymnosperm outgroup) with a tblastx search (Altschul et al., 1990), by using a nucleotide sequence (an entire coding sequence of *PGK*-1, which codes for plastidic PGK, from *Arabidopsis thaliana* (AT1G79550, on chromosome I; this is one of three loci in this genome) as a translated query against all six reading frames of subject nucleotide databases, here representing each of the following genomes: *Oryza sativa*, *Zea mays*, *Cucumis sativus*, *Glycine max*, *Theobrama cacoa*, *Populus trichocarpa*, *Arabidopsis thaliana*, *Carica papaya*, *Citrus sinensis* and *Vitis vinifera*. These genomes are all available on Phytozome v9.1 (Goodstein et al., 2012). I recovered *PGK* sequences from *Amborella trichopoda* on the Amborella Genome Database (http://www.amborella.org/), from the *Pinus taeda* genome using Pine Reference Sequences (http://pinegenome.org/pinerefseq/), and from *Trithuria submersa* by using an unpublished 454-based transcriptome of this species produced by M. Barker, L. Rieseberg and S. W. Graham (unpublished data). At least two sequences were recovered in all taxa. I aligned these and performed phylogenetic analysis as above, treating each codon position as a separate data partition for likelihood and Bayesian analysis, the optimal partitioning scheme indicated by PartitionFinder (Appendix 2c).

### 2.6 Estimating the root of Hydatellaceae from *PGK* data

As including all variants (and copies) as independent terminals in an ML (gene-tree) analysis did not recover a well supported root for either copy of the *PGK* gene (see

Results and Suppl. Fig. 1), I ran a concatenated analysis in which I grouped together single representative variants from both copies from a small subset of individuals species chosen to represent each taxonomic section. I included only a subset of individuals/species because of the observed intermingling of variants from different individuals and species at the sectional level in individual gene trees (see below). I included only exons in this analysis, and used a single outgroup, *Cabomba caroliniana* (Cabombaceae) for which I had cloned variants from both copies (i.e., *PGK*-1 and *PGK*-2). A PartitionFinder analysis supported including all both genes as a single data partition, so no partitioned ML or Bayesian analyses were attempted. I ran parsimony, likelihood and Bayesian analysis on this data set using the settings described above.

# 3. RESULTS

## 3.1 Plastid phylogeny

The sequences from eight individuals that I added to the plastid-based analysis place with conspecific sequences in the phylogenetic analysis of the plastid data (new sequences are noted in bold in Fig. 2). Specifically, the two new sequences of *T. australis* recovered from a previously unsampled population (Appendix 1) place together, defining a deep divergence in the *T. australis* clade, with moderate to strong support for this arrangement. The new sequences of *Trithuria bibracteata* and *T. submersa* place in a clade comprising intermingled sequences from other populations of these two species and *T. occidentalis*, as in Iles et al. (2012, 2014), with weak to strong support for their local placement (Fig. 2). Specifically, the two samples from a previously unsampled population of *T. submersa* from western Australia place with the other western Australian *T. submersa* populations; the two samples from a previously unsampled population of *T. bibracteata* place in a mixed clade comprising other *T. bibracteata* and *T. submersa* sequences. The additional *T. occidentalis* sequences group with the previously published sequence from this population (whose placement in Iles et al., 2012, was based only on a short *mat*K sequence retrieved for that individual); the resulting clade of *T. occidentalis* represents a relatively deeply diverging lineage within this mixed clade of three species. The new *T. occidentalis* sequences are from staminate and pistillate individuals that come from the only known population of this species, near Perth, Australia. All other species in the plastid analysis (for which I have multiple individuals sampled per population) are monophyletic at the current level of taxon sampling.

### 3.2 The history of the *PGK* duplication in angiosperms

Two copies of the *PGK* gene (plastidic *PGK*-1 and cytosolic *PGK*-2) are present in most angiosperm genomes examined here; I also recovered only two copies by Blast-analysis of the *Trithuria submersa* transcriptome. A subset of species have more than two copies: *Arabidopsis* has three, and *Glycine*, *Oryza*, *Populus* and *Zea* have four copies each (Fig. 3). I observed two major clades of *PGK* in angiosperms, supporting the view that a major duplication happened in this locus around the origin of the angiosperms (Longstaff et al., 1988; Martin and Schnarrenberger, 1997). *Pinus* also has two copies, but these cluster together closely in a well supported clade, and so they may represent a different duplication from the one present in angiosperms (the seed-plant rooting shown in Fig. 3 may not be correct; however, the *Pinus* variants are quite closely related copies, and it is not possible to re-root the tree so that one copy each is the sister group of the two major *PGK* loci in angiosperms). Similarly, although two copies are present in *Amborella*, these form a well-supported clade that appears to be nested in one of two angiosperm copies, supporting a recent duplication in this lineage. It is likely that were also several subsequent duplications in different lineages (distinguished in Fig. 3 with additional labels; 1i, 1ii, etc).

### 3.3 Simultaneous vs. separate analysis of the duplicated *PGK* locus

When all of the amplified sequences (based on the coding and noncoding regions spanning portions of exon 4 to exon 5) of *PGK* are analyzed simultaneously in the species-level analysis, with sequences included from several water lilies and *Amborella*, I

recovered two main clades of *PGK* for Hydatellaceae. All species of *Trithuria* had one or more *PGK* variants that placed in each of these two well-supported clades (Suppl. Fig. 1). These clades are consistent with the two loci inferred in the angiosperm-wide analysis (Fig. 3), and are named accordingly. Both sequences retrieved from the published *Amborella* genome grouped in a single small clade, that in turn grouped with one clade of *PGK* (at least according to the rooting shown in Suppl. Fig. 1); the clade it grouped with is inconsistent between the angiosperm-wide analyses and the species-level analysis, but the support values for the relative placements of *Amborella* are poor in both cases (Fig. 3; Suppl. Fig. 1). In the water lilies, I recovered both major copies for *Cabomba* by PCR (one copy each grouped as the sister group of corresponding clades of Hydatellaceae), but only representatives of one major *PGK* clade (*PGK*-2) for *Brasenia*, and the other major *PGK* clade (*PGK*-1) for *Nuphar* (Fig. 3).

In addition to grouping into two major *PGK* clades, the introns from the two copies were too distant from each other to be aligned with any confidence to each other (only exons are alignable among outgroup taxa and Hydatellaceae, although portions of the *Cabomba* introns were alignable to corresponding copies of *Brasenia* or *Nuphar*). The unalignability of the introns from different copies, and between ingroup and outgroup taxa, also supports the hypothesis of distinct *PGK* loci. The implied root of Hydatellaceae inferred from each copy disagrees with the angiosperm-wide analysis and with the plastid-based root found in Iles et al. (2012) and here (Fig. 2); see arrows in Suppl. Fig. 1 here (and also Fig. 4, discussed below). These roots are likely artifactual, as they are poorly supported and disagree with each other (between copies; Suppl. Fig. 1). In addition, the branch lengths between copies are very long, and are based on the quite

short exon sequences that are alignable (~241 bp, unaligned). In the case of *PGK*-1, the implied rooting would place only two variants of *T. lanterna* as the sister group of all remaining taxa, which seems highly unlikely.

I therefore repeated the analysis by restricting the analysis to sub-matrices that correspond to *PGK-1* and *PGK-2*, individually. I also excluded the outgroup taxa in both case because of the substantial distance between water lilies and Hydatellaceae. The two subtrees recovered for Hydatellaceae, corresponding to *PGK-1* and *PGK-2* (Fig. 4; re-rooted to match the plastid-based root; Fig. 2), are both consistent with the overall phylogenetic structure depicted in Iles et al. (2012, 2014), at least at the sectional level. Arrows in Suppl. Fig. 1 indicate the plastid-based root for Hydatellaceae; arrows in Fig. 4 indicate the roots implied by the simultaneous nuclear gene analysis of the duplicated loci (sections are also noted in Fig. 4). All subsequent discussion of *PGK* gene trees refers to analyses based on *PGK-1* and *PGK-2* separately (Fig. 4a,b), and the rooting implied by the plastid data. Variants within species are arbitrarily distinguished using a single-letter suffix (a, b, c, etc; Fig. 4, Appendix 3). It is not possible here to determine here which variants are due to allelic variation vs. subsequent gene duplications/polyploidy, and they are not distinguished here. However, additional gene duplications (or polyploidy, a species-level form of gene duplication) may be indicated when there are more than two variants present in an individual (see below). Polymerase error during amplification in a lab setting (reflecting mutation or recombination) may also contribute to some sequence variation, although I did not see obvious evidence of recombination or gene conversion between the copies.

Considering the minimum of eight successful clones that I recovered for each individual, I recovered at least a single unique variant from *PGK-1* and *PGK-2* in all cases (Appendix 3; except for one sample of *T. austinensis* where all eight of the clones were *PGK-1* variants). I generally observed a relatively even split in number of variants recovered for *PGK-1* and *PGK-2*. However, it is possible that additional variants would be recovered with further sampling. I explored this for two individuals (one individual each of *T. lanterna* and *T. submersa*; Appendix 3) by doing a replicate amplification and additional cloning. One to three additional variants were picked up in both cases (for the *T. lanterna* example, six of nine variants across the two PGK copies were observed in both amplification/cloning attempts; for *T. submersa*, four of five of the variants recovered the two PGK copies were observed in both attempts).

### 3.4 Relationships within sections *Altofinia* and *Hammania*

Sequences from the two tropical sections, *Altofinia* and *Hammania*, are intermingled for *PGK*-2 here (but not for *PGK*-1), and so I discuss these two sections together here. I sampled one individual for species in sections *Altofinia* and *Hammania*, except for *T. lanterna*, where I sampled two individuals. I recovered multiple variants for all species except *T. cookeana*. When multiple variants were recovered for a *PGK* copy (*PGK*-1 and *PGK*-2, shown in Figs. 4a and 4b, respectively), these copies did not place together in a clade, except for *T. konkanensis*, where five recovered variants comprise a clade for *PGK*-1 (Fig. 4; Appendix 3). In many cases the variants present within a species were quite distantly related to each other, even when copies from a species formed a clade (e.g., *T. konkanensis*). In several cases, variants found within species in sect. *Hamannia* (*T.*

*lanterna*; *T. polybracteata*) or individuals (*T. lanterna*) were more closely related to variants from species in sect. *Altofinia*, than to other variants in sect. *Hamannia*. In a few instances, one or more species had variants that were identical to each other, in contrast to other more distantly related variants from the same individual or species (e.g., the variant found for *T. cowieana*, *T. polybracteata* and *T. lanterna* for *PGK*-2; the variant found in *T. cookeana* and *T. cowieana* for *PGK*-1; Fig. 4; Appendix 3).

### 3.5 Relationships within section *Hydatella*

Two of the three species for which I sampled at least two individuals per species (Appendix 3) were monophyletic for *PGK*-1 and *PGK*-2, respectively (*T. austinensis* and *T. australis*; Fig. 4a,b; Appendix 3). The absence of *PGK*-1 for one individual of *T. austinensis* (the Keighery & Gibson sample; Appendix 3) may reflect failed amplification of this locus. Most individuals of *T. australis* had a single variant per individual, per locus (but two variants in one case); individuals in *T. austinensis* typically had one to two variants per individual, for each copy (three observed in one case). Only one variant was recovered from the single individual sampled for *T. filamentosa*, for both gene copies. Both sampled individuals of *T. inconspicua* contained two-three distinct variants that fell in two disjunct parts of each of the two gene trees; for *PGK*-1, one of these variants was closely related to the *T. australis* sequences (Fig. 4b); for both *PGK*-1 and *PGK*-2, the *T. filamentosa* sequence grouped closely with one set of *T. inconspicua* variants (Fig. 4a,b). A maximum of three variants was recovered from each individual of *T. inconspicua* (Appendix 3).

### 3.6 Relationships within section *Trithuria*

Sequences from the three species in section *Trithuria* are generally extensively intermingled in the species-level *PGK* trees (Fig. 4). There is fairly deep sequence divergence among many of these variants, and moderate to strong support for a subset of the branches. One exception is that the sequences retrieved from *T. occidentalis* comprise a moderately well supported clade for both gene copies (Fig. 4a,b); however, the placement of this clade differs between *PGK*-1 and *PGK*-2. Little geographic structuring was evident within the section *Trithuria* clade as a whole, for either copy: eastern and western populations of *T. submersa* are intermingled, except for a poorly supported clade comprising six western variants of *T. submersa* for *PGK*-2 (plus a single *T. bibracteata* variant), and a small clade of eastern *T. submersa*, also for *PGK*-2 (Fig. 4b). *Trithuria bibracteata*, a western Australia species, has some sequences that are close to western population of *T. submersa*, and others close to eastern populations of it (Fig. 4).

### 3.7 Gene-tree reconciliation and inference of the Hydatellaceae species tree

The species tree that I inferred in a multispecies coalescent-based reconciliation of plastid and ITS data (based on the matrix of Iles et al., 2012, 2014, but updated here with several new plastid sequences from four species: *T. australis*, *T. bibracteata*, *T. occidentalis* and *T. submersa*), is nearly identical to the previously published dated multispecies coalescent analysis in terms of tree topology, branch support, node dates and HPD (highest posterior density); see Table 1. However, there are a few minor differences. First, several nodes have slightly older mean ages here (nodes 1, 10; Table 1). Second, the branch between nodes 7 and 8 (Fig. 5), concerning the relative arrangement of *T. australis* and *T.*

*austinensis* is less well supported here (0.80 here vs. 0.93 posterior probability in Iles et al. 2014), and the branch between nodes 10 and 11, concerning the position of *T. occidentalis* in section *Trithuria*, is better supported here (0.94 vs. 0.59 posterior probability in Iles et al. 2014). For both branches, these involve nodes that are ancestral to extant species where I have added several individuals; in the case of *T. occidentalis*, I also added substantially more data per taxon (previously this species was represented by a short fragment of a single plastid gene, *mat*K, from one individual; Iles et al. 2012).

A species tree based on a reconciliation analysis of the two *PGK* copies to each other recovered the four sections of *Trithuria* that are seen in previous studies (and in Fig. 5 here), with the same relative arrangement of sections (Fig. 6). However, the ages of some of the corresponding nodes differed substantially (e.g., nodes 3, 5 and 7 are substantially older with little overlap in HPD; node 4 is substantially younger, with little overlap in HPD between the PGK and plastid/ITS analysis; node 8 is also younger, but the HPDs overlap for it; Table 1). Within section *Hammania*, the relative arrangement of *T. konkanensis*, *T. lanterna* and *T. polybracteata* (Fig. 6) differed from the reconciliation based on the plastid and ITS data (Fig. 5, and see Iles et al. 2012, 2014), but there was poor support here for this conflicting arrangement. Within section *Trithuria*, the relative arrangement of *T. bibracteata* and the eastern and western *T. submersa* 'species' are also different (Figs. 5, 6); the support for their relative arrangement is poor here (and also in the analysis of the plastid and ITS data; Table 1; Fig. 5).

### 3.8 A nuclear-based root of Hydatellaceae inferred from concatenated *PGK* exons

The best ML tree from a concatenated analysis of combined exonic sequences (both *PGK*-1 and *PGK*-2; alignment length = 536 bp), using representative individuals and taxa (Fig. 7), supported a root of the family that is consistent with the root inferred using plastid data (Iles et al. 2012, 2014; Fig. 2 here). Although this root was not well supported, a clade comprising sections *Hydatella* and *Trithuria* was moderately well supported, as were these both of these sections (75-83% support from ML bootstrap analysis; Fig. 7).

# 4. DISCUSSION

## 4.1 Utility of the *PGK* locus in phylogenetic inference in Hydatellaceae

The clone-based study of variation in the *PGK* loci of individual species of *Trithuria* often revealed substantial variation within species and within individuals, suggestive of substantial incomplete lineage sorting, and possibly other processes that may lead to trans-specific polymorphism (e.g., Omland and Funk, 2003; Percy et al., in press). This contrasts strongly with evidence from plastid genes and the nuclear ITS region (Iles et al. 2012, 2014), which both support species monophyly for most species in the family (except *T. submersa* and *T. bibracteata*). It is likely that the pattern recovered for the *PGK* genes will prove to be more typical for other nuclear genes in this family: plastid genes coalesce faster than nuclear genes (as they are haploid, e.g., Moore, 1995), and the nuclear ITS region is also known to coalesce more rapidly than other nuclear loci (because of concerted evolution among copies, e.g., Baldwin et al. 1995; Buckler and Holtsford, 1996; the nuclear ITS region has various fates after allopolyploidization: different variants may be maintained without recombination, e.g., Soltis et al., 1991, 1995; Baumel et al., 2001, and they may go through various degrees of recombination, e.g., McDade, 1992; Barkman and Simpson, 2002. Only one variant may be left after concerted evolution, e.g., Wendel et al., 1995; Brochmann et al., 1996). Plastid genes and ITS data may therefore each more accurately track species phylogeny and species boundaries than individual nuclear genes. Consistent with this view, in general the phylogenies that I recovered from the *PGK* loci were only consistent at the sectional level and above, and showed substantial intermingling among species within sections. A

similar pattern of apparently slow nuclear coalescence (long retention of ancestral polymorphisms) was also found in a recent analysis of two low-copy nuclear genes for two genera in the Hawaiian flora (Pillon et al., 2013). The *PGK* data recovered here confirm deep relationships in Hydatellaceae inferred using plastid and ITS data, but ultimately provide coarser phylogenetic resolution than plastid or ITS data. They also provide nuclear evidence (nuclear ITS data were not retrieved for my samples due to fungal contamination), here confirming the link between pistillate and staminate individuals of *T. occidentalis*, for example. They may also give insights into the level of trans-specific polymorphism that we might expect in nuclear loci in general.

**4.2 Multiple variants as evidence of polyploidy**

The *PGK* loci may also provide insights into the ploidy level of individual species. The occurence of three or more variants in an individual (Appendix 3) points to a ploidy level that is higher than diploid (because a maximum of two variants are possible at a single locus in diploids if there is heterozygosity), and may also point to allopolyploid origins (recent autopolyploids should have an identical complement to their diploid progenitors). It should be noted, however, that the existence of more than two variants does not rule a diploid status, as individual genes may duplicate by mechanisms other than whole-genome duplication (some variants may also result from mutations introduced during amplification, although the error rate of the Phusion polymerase that I used is very low: 0.462% of PCR products are expected to contain an error after 30 cycles of amplification, calculated from http://www.thermoscientificbio.com/webtools/fidelity/). In general, however, most angiosperms appear to retain a low number of copies of the *PGK* locus

(Fig. 3). Multiple (more than two) variants of *PGK*-1 or *PGK*-2 in a single individual may therefore be consistent with the existence of polyploidy in *T. konkanensis, T. lanterna* and *T. polybracteata* (Appendix 3; in all three cases I observed a maximum of five or six variants per individual, potentially consistent with at least a pentaploid or hexaploid ploidy level), a maximum of four variants per individual in *T. submersa* (potentially consistent with tetraploidy or higher ploidy levels), a maximum of three copies observed in *T. austinensis* (potentially consistent with a tetraploidy in this sexually reproducing species), and a maximum of three copies observed in *T. inconspicua* (possibly consistent with triploidy in this asexually reproducing species). Unpublished results indicate that *T. australis* is a diploid (*2n* = 14; R.G. Kynast, P.J. Rudall et al., unpublished data), which would be consistent with the maximum of two copies observed per individual here. *Trithuria bibracteata, T occidentalis, T. cookeana*, *T. cowieana* and *T. filamentosa* also all have a maximum of two variants observed per individual (although only one individual was sampled for the latter three species), perhaps supporting diploidy in these species. One *PGK-1* variant from *T. inconspicua* groups with *T. australis*, which may suggest that *T. inconspicua* came from an allopolyploidy event involving this lineage. There is 13 bp deletion in the sequence of one variant of *PGK*-2 (*T. inconspicua*_2_Forester, variant_b, and *T. inconspicua*_2_Chapman, variant_b in Fig. 4). This reading-frame interruption may indicate that this variant is a pseudogene that lost its function after polyploidization. If the base chromosome number in the family is that seen in *T. australis* (i.e., *n* = 7), then reported numbers for other taxa are consistent with several of these species being polyploids: *T. konkanensis*, 2n = 40 (Gaikwad and Tadav,

2003; hexaploid?), *T. inconspicua*, 2n = ~24 (de Lange et al., 2004; triploid?), and *Trithuria submersa* with 2n = 56 (Kynast et al., submitted; octoploid?).

Polyploidy may therefore explain the large number of variants observed in some taxa (Fig. 4, Appendix 3). Allopolyploid events may be implied wherever alleles are more closely related between species than within species (Fig. 4): while some of these phylogenetic distributions of allelic variation (patterns of trans-specific polymorphism) may result from incomplete lineage sorting, this may be an insufficient explanation when shared or related variants between species are identical or very closely related (see Percy et al., in press, for a comparable recent example in *Salix*). The identical *PGK-2* variants shared by *T. cookeana*, *T. lanterna* and *T. polybracteata*, or the *PGK-1* variants shared by *T. cookeana* and *T. cowieana* provide particularly striking examples of this (Fig. 2), and the complete lack of diversification of these alleles among species is unlikely to be consistent with the timing of lineage sorting events that would be implied by the fairly ancient speciation events indicated by the plastid and ITS data for affected species (Fig. 6). They may also be inconsistent with allopolyploidy events around the times of these speciation events, but may reflect subsequent gene flow between these lineages.

It should be borne in mind that additional sampling (cloning) could reveal more variants per individual, as I discovered when I re-investigated single individuals of *T. lanterna* and *T. submersa*. In addition, it should be noted that polyploidy is not the only mechanism for increasing gene copy number. However, other research has shown that kinases exhibit high retention rates following polyploidy and low retention rates following non-polyploid duplications (Blanc et al., 2004; Maere et al., 2005). Because of the dosage sensitivity of photosynthesis protein complexes, these enzymes tend to show

retention of polyploid duplicates, and active elimination of non-polyploid duplicates (Coate et al., 2011). In a study of comparative evolution of photosynthetic genes in soybean, barrel and *Arabidopsis* (Coate et al., 2011), *PGK* was found to have no non-polyploid duplicates, but kept all polyploid duplicates (apart from one in *Arabidopsis*). Polyploidy is therefore the most likely explanation for the multiple variants observed in Appendix 3 (i.e., some of these likely represent variants found at additional copies of *PGK*-1 or *PGK*-2).

### 4.3 Conclusions

The duplicated gene history of the PGK loci in Hydatellaceae is likely to be complicated, but at least two copies are found in all species, with patterns of variation and numbers of variants that are consistent with allopolyploidy in several cases. The inference of a species tree using algorithms based on lineage sorting (e.g., *BEAST analyses) is therefore inappropriate, although it is possible that lineage sorting also contributes to some of the shared variation among species. However, when the *PGK* gene trees are reconciled using this approach, the resulting species tree inference conflicts with one inferred using reconciled plastid and nuclear ITS data, both in terms of topology and diversification dates (Figs. 6, 7; Table 1); this conflict may sign-post the contribution of hybridization and allopolyploidy to genome diversity in species of Hydatellaceae.

Inference of the root of Hydatellaceae using a duplicated PGK gene-tree approach (e.g., Mathews and Donoghue, 1999, 2000) was not successful here, likely because the individual *PGK* coding sequences loci are too short, and the major copies too diverged from each other, to reliably infer well the root of the family. However, a concatenated analysis of the two *PGK* loci produces a weakly supported tree root that is at least consistent with the one inferred using plastid data (Fig. 2, 7), and with the sectional classification of the family (Iles et al., 2012). In general, concatenated phylogenetic analysis of duplicated genes like *PGK* should be avoided when allopolyploidy is inferred. If the complex patterns of allele sharing observed here is indicative of other nuclear genes, this indicates that future efforts to infer relationships using concatenated analysis of genes culled from transcriptome or similar genomic data should also be avoided for Hydatellaceae (and for similar taxonomic groups with closely related polyploid species).

It would be useful to use full-length *PGK* genes to further explore the history of duplication of this locus in angiosperms or more broadly in seed plants, but it is currently ambiguous whether the duplication entirely pre-dated angiosperms (note that *Amborella* and *Pinus* each has two copies, but each pair may have a relatively recent origin, Fig. 3).

New plastid data collected here and added to an earlier study recovered deep population diversity in *T. australis*, and confirm the biological and taxonomic link between pistillate and staminate individuals of *T. occidentalis* (Fig. 2), which were previously recognized as different genera; this finding that was also that is also supported by the *PGK* data (Fig. 4). The *PGK* gene trees also support the current sectional classification of the family (Iles et al., 2012).

**Tables and Figures**

Table 1. Estimated ages of splits in Hydatellaceae phylogeny based on a Bayesian multi-species coalescent analysis of plastid vs. nuclear ITS data, and the two *PGK* copies. Mean age and 95% HPD are indicated for the individual nodes labelled in Figs. 5 and 6. 'NA' indicates value not available (split occurred in less than 50% of bipartitions); '--' indicates that clade was not recovered in the Bayesian consensus tree. The plastid and nuclear ITS data from Iles et al. (2014) are shown for comparison here (fewer samples were included in that study).

| | | **Iles et al. (2014):** | | **This study:** | | **This study:** | |
|---|---|---|---|---|---|---|---|
| | **Reconciled data:** | **Plastid & ITS** | | **Plastid & ITS** | | ***PGKI* & *PGKII*** | |
| Node | Crown clade or taxon (if applicable) | Mean age (Ma) | 95% HPD of age (Ma) | Mean age (Ma) | 95% HPD of age (Ma) | Mean age (Ma) | 95% HPD of age (Ma) |
| **1** | Hydatellaceae | 17.55 | 14.69-20.62 | 18.61 | 16.85-20.42 | 18.69 | 16.91-20.76 |
| **2** | | 16.07 | 13.48-18.71 | 16.54 | 14.80-18.19 | 16.34 | 14.48-18.23 |
| **3** | | 6.15 | 4.34-8.06 | 6.44 | 4.63-8.56 | 13.11 | 7.14-18.45 |
| **4** | Sect. *Altofinia* | 4.27 | 2.64-5.97 | 4.38 | 2.45-6.45 | 0.46 | 0.00-1.8 |
| **5** | Sect. *Hammania* | 1.54 | 0.73-2.41 | 1.55 | 0.60-2.56 | 8.91 | 4.31-13.93 |
| **6** | | 0.76 | 0.24-1.33 | 0.74 | 0.12-1.39 | -- | -- |
| **13** | | -- | -- | -- | -- | 7.34 | 2.61-11.54 |
| **7** | Sect. *Hydatella* | 6.27 | 4.45-8.11 | 6.53 | 4.34-8.58 | 10.16 | 4.01-15.62 |
| **8** | | 5.12 | 3.44-6.77 | 5.36 | 3.24-7.38 | 2.94 | 0.65-6.15 |
| **9** | | 0.51 | 0.00-1.12 | 0.34 | 0.00-1.08 | 0.49 | 0.00-1.5 |
| **10** | Sect. *Trithuria* | 1.78 | 0.59-4.64 | 2.35 | 1.16-3.49 | 3.33 | 1.08-6.8 |
| **11** | | 1.04 | 0.57-1.57 | 1.33 | 0.77-1.87 | 1.87 | 0.49-3.67 |
| **12** | | 0.78 | NA | 1.09 | 0.47-1.65 | -- | -- |
| **14** | | -- | -- | -- | -- | 0.77 | 0.00-2.05 |

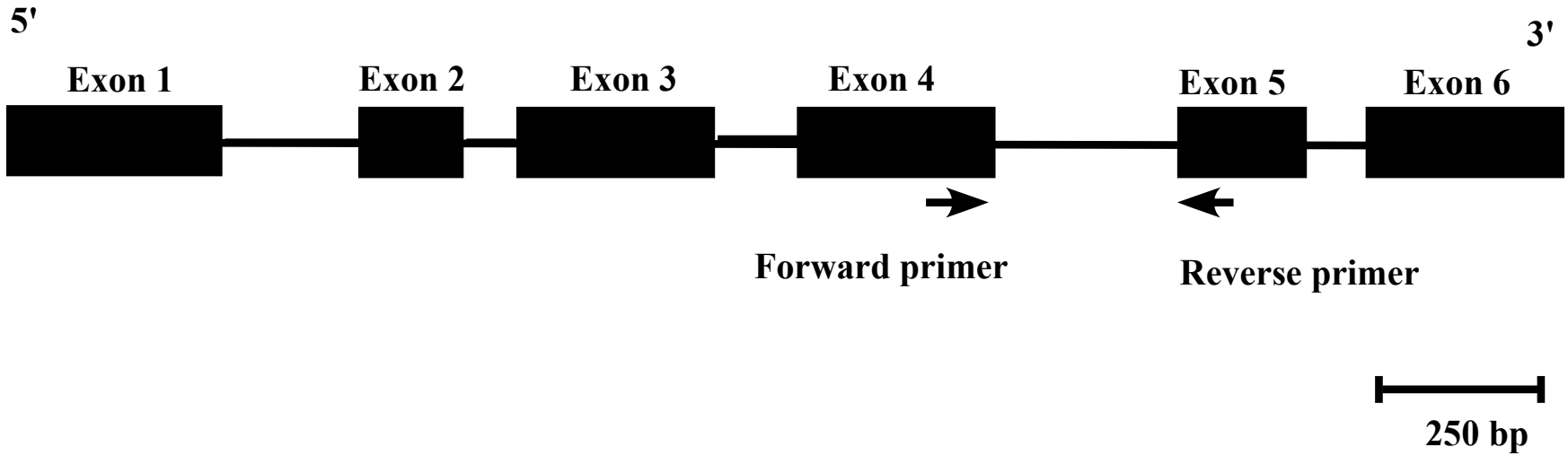



Fig. 1. Primer map for the phosphoglycerate kinase *PGK* gene. Exon and intron information is based on the structure of the *PGK*-1 gene in *Arabidopsis thaliana* (NCBI), primers are not to scale. Scale bar: 250 bp (relative to *Arabidopsis thaliana*).

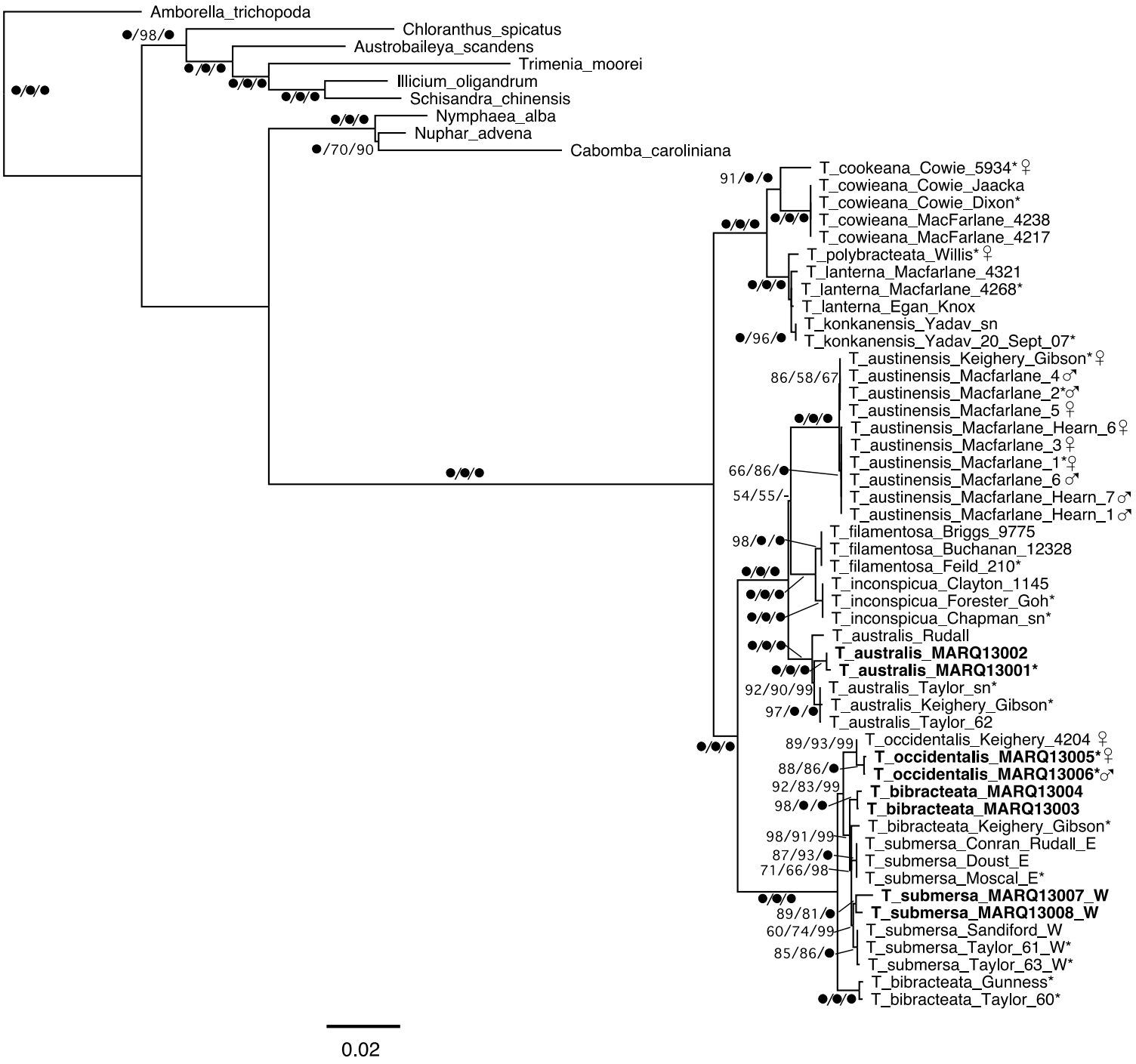



Fig. 2. Maximum likelihood tree of Hydatellaceae and relatives inferred from four plastid loci (*rbc*L, *mat*K, *ndh*F, *atp*B). Each taxon name is followed by source information (collector name and number) and the gender of the sample (when the species is dioecious). Samples with "*" were also sampled in the species-level *PGK* analysis. For *T. submersa* samples: W = south-west Australia. E = south-east Australia. Numbers adjacent to branches are bootstrap support values (respectively: parsimony and likelihood bootstrap values; posterior probabilities expressed as percentages. Filled circle: 100%; dash: <50%). Scale bar: expected substitutions per site. Samples new to this study (compared to Iles et al., 2012, 2014) are highlighted in bold.

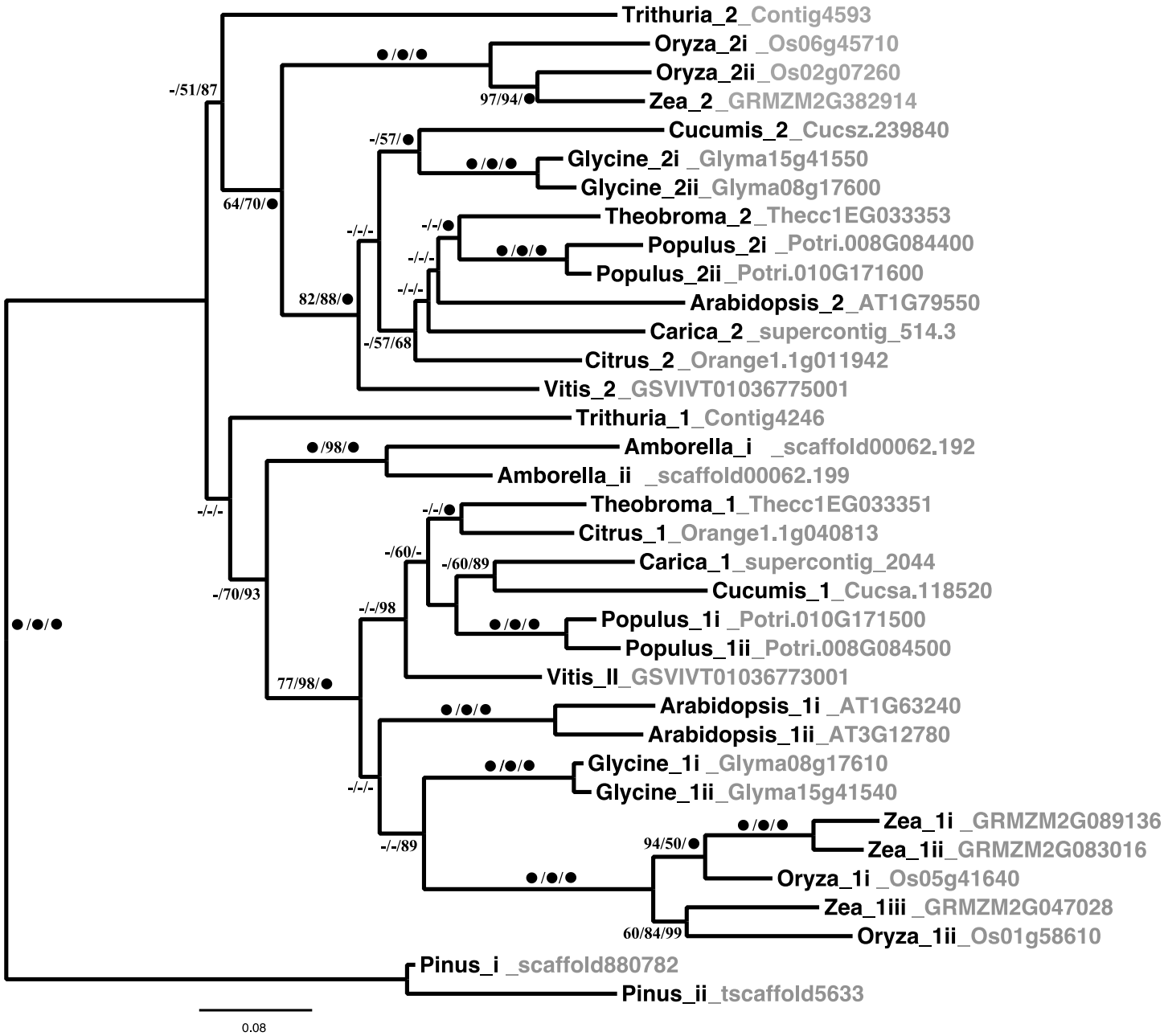



Fig. 3. Maximum likelihood tree of exonic sequences of the angiosperm nuclear-encoded phosphoglycerate kinase (*PGK*) loci (with *Pinus* as an outgroup). The genus name of each taxon is followed by the copy number (1 = *PGK-1* = plastidic; 2 = *PGK-2* = cytosolic; subscripts i, ii, etc, indicate additional copies) and locus tag/contig number (grey font). Numbers adjacent to branches are bootstrap support values (respectively: parsimony and likelihood bootstrap values; posterior probabilities expressed as percentages. Filled circle: 100%; dash: <50%). Scale bar: expected substitutions per site.

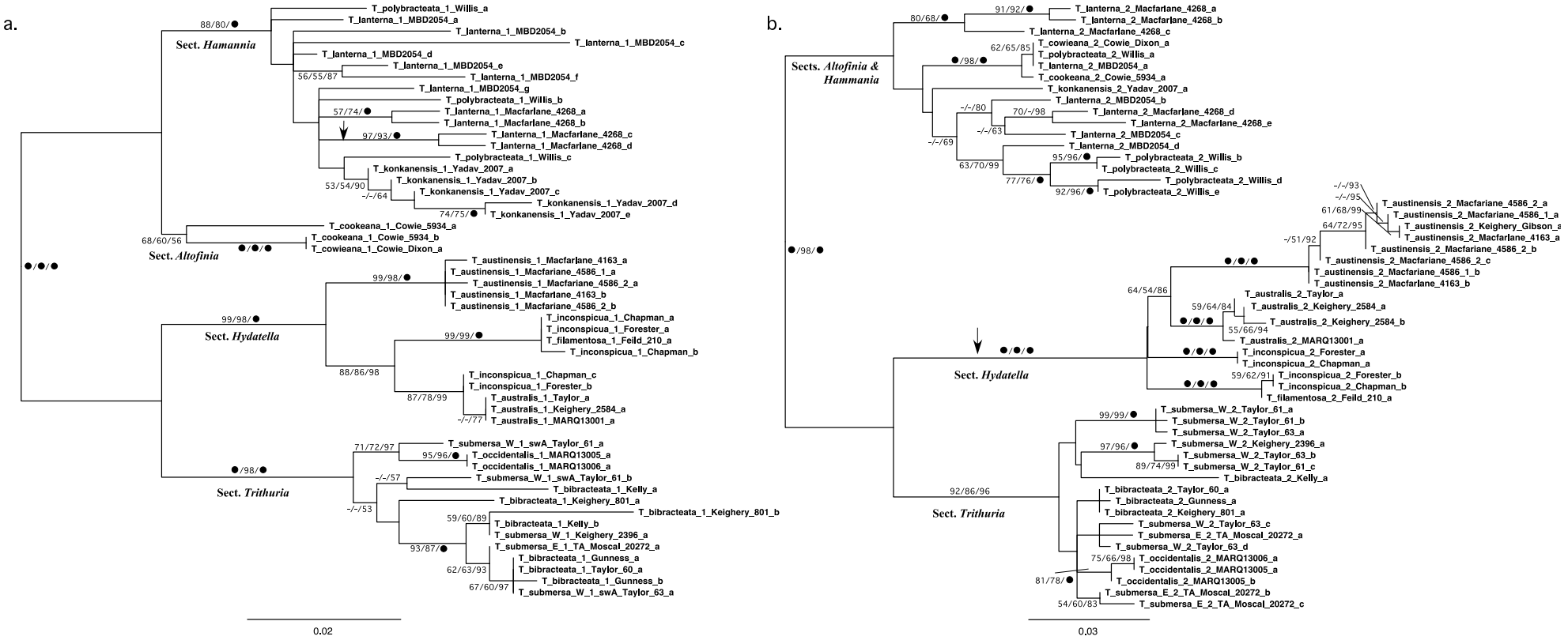



Fig. 4. Maximum likelihood trees of two copies of the nuclear phosphoglycerate kinase (*PGK*) loci for Hydatellaceae (a) *PGK-1*, the plastidic copy; (b) *PGK-2*, the cytosolic copy. Each taxon name is followed by the inferred *PGK* gene copy number (1 vs. 2), source information (collector name and number); the last character indicates clone type (letter a, b, c, etc for different clone variants), in that order. For *T. austinensis*_Macfarlane_4586, the second last number indicates the individual number (two individuals sampled in this population); for *T. submersa* samples: W = south-west Australia. E = south-east Australia. Numbers adjacent to branches are bootstrap support values (respectively: parsimony and likelihood bootstrap values; posterior probabilities expressed as percentages. Filled circle: 100%; dash or no support value: <50%). Scale bars: expected substitutions per site. The arrows show inferred roots of the family from a separate analysis that included both loci and outgroup samples. (Suppl. Fig. 1)

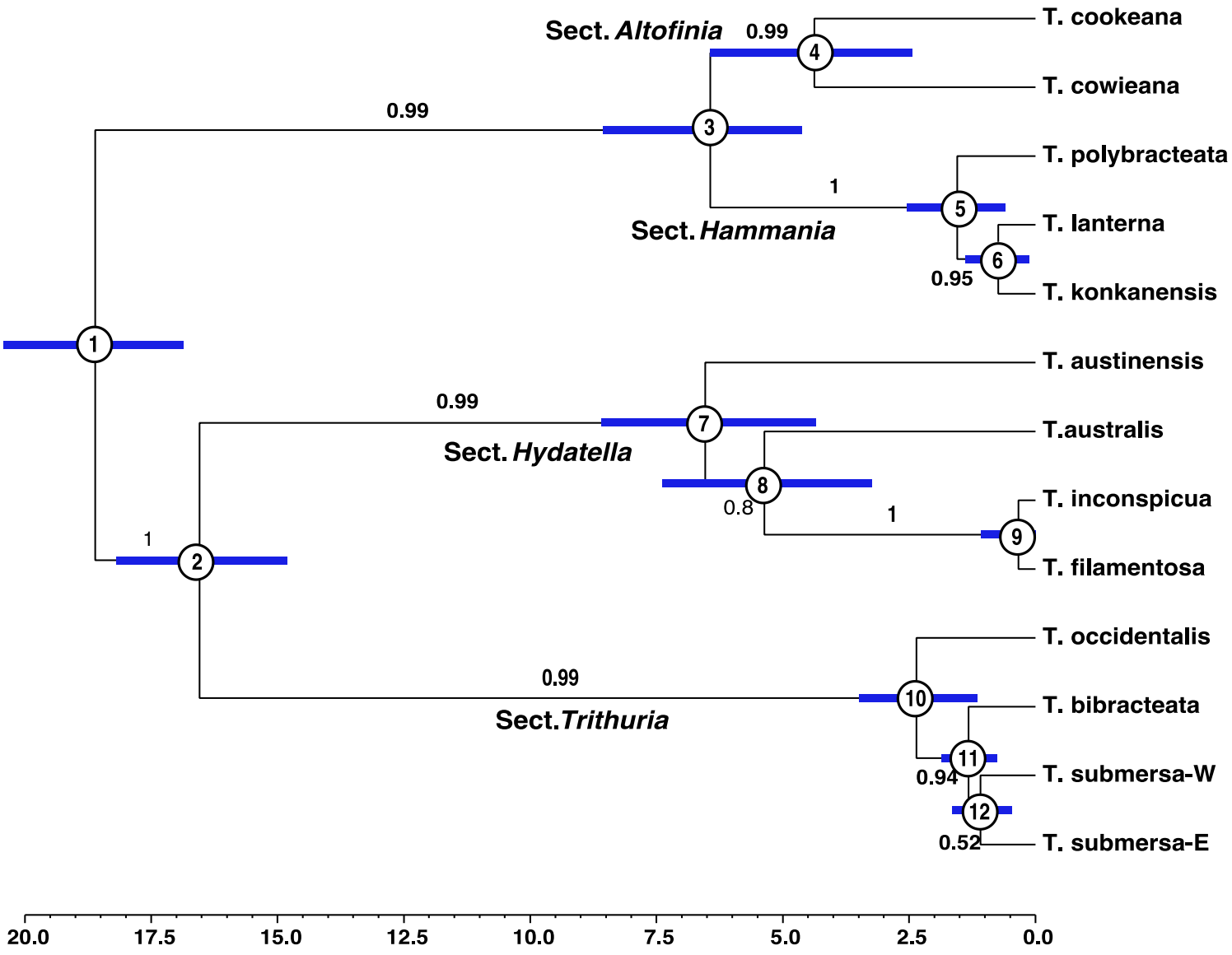



Fig. 5. Species tree and timing of divergences in Hydatellaceae inferred from a Bayesian multispecies coalescent analysis of plastid and nuclear ITS data, with prior dating estimates for the first two nodes based on Iles et al. (2014). Labelled nodes are referred to in Table 1. *Trithuria submersa* is provisionally divided into two "species" (see text). Numbers adjacent to branches are posterior probabilities. The time scale is in Ma. Divergence time uncertainty is shown by blue bars, representing 95% HPD.

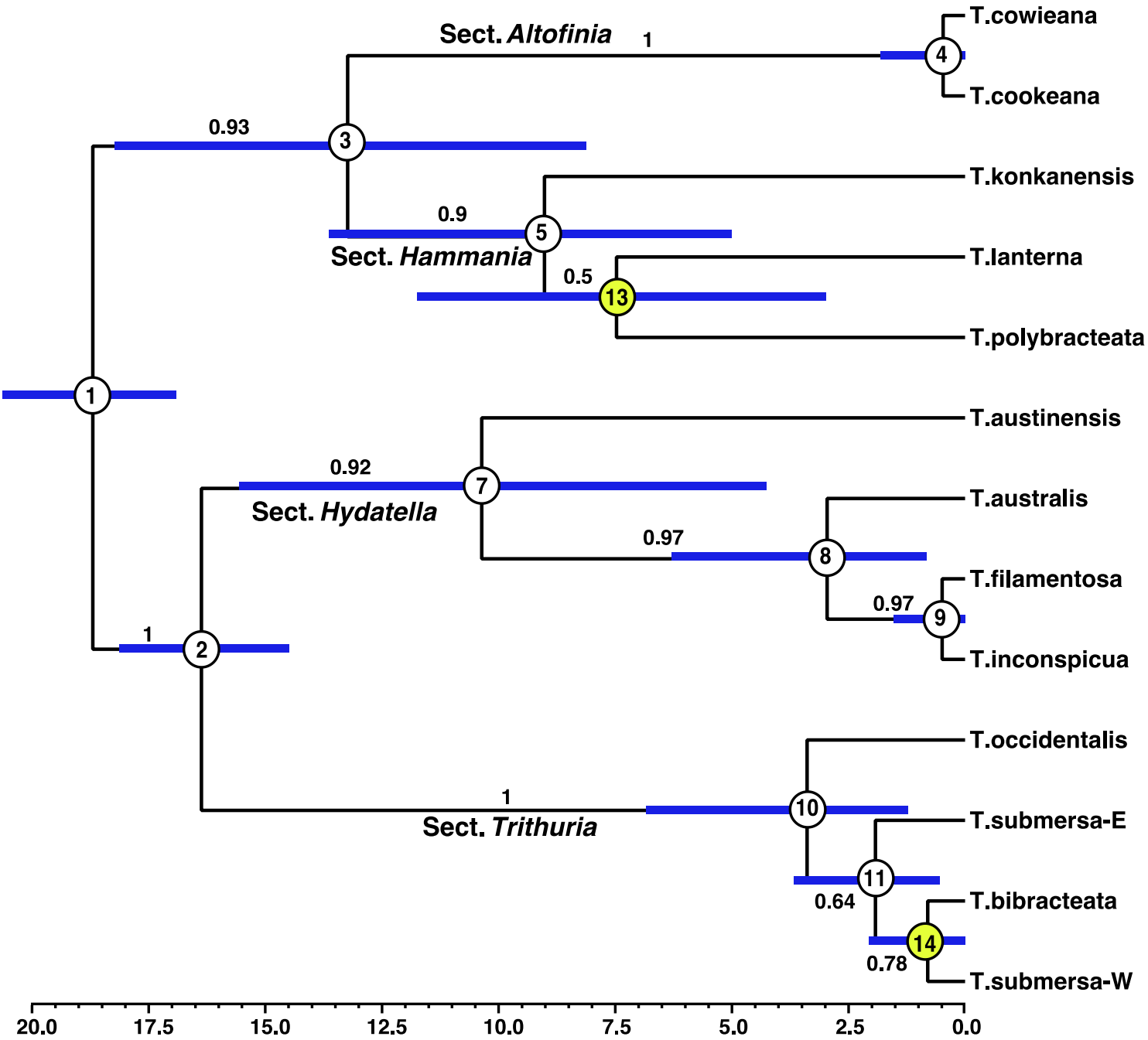



Fig. 6. Species tree and timing of divergences in Hydatellaceae inferred from a Bayesian multispecies coalescent analysis of the two *PGK* loci, with prior dating estimates for the first two nodes based on Iles et al. (2014). Labelled nodes are referred to in Table 1 (nodes not seen in Fig. 5 are highlighted in yellow). *Trithuria submersa* is provisionally divided into two "species" (see text). Numbers adjacent to branches are posterior probabilities. The time scale is in Ma. Divergence time uncertainty is shown by blue bars, representing 95% HPD.

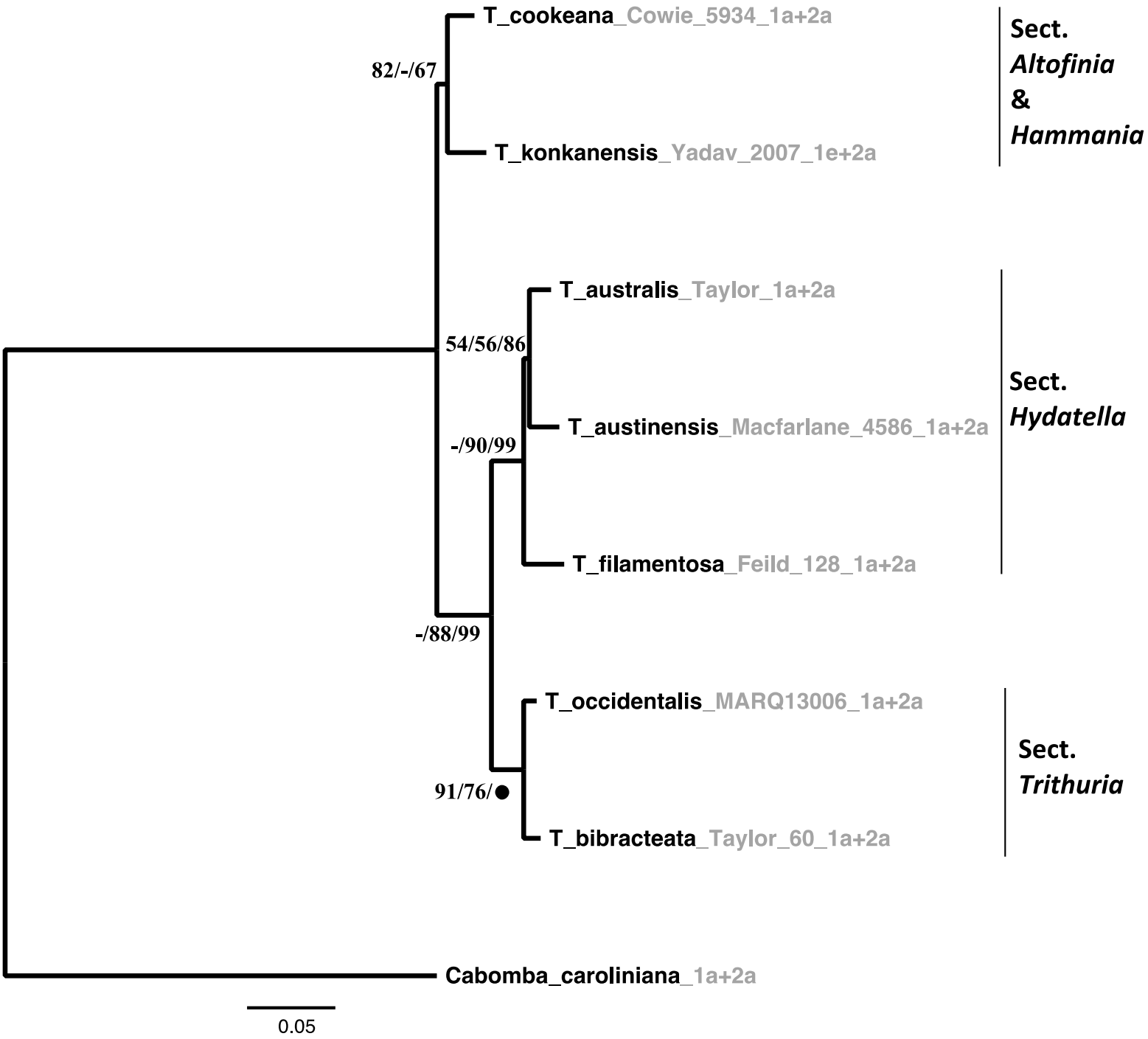
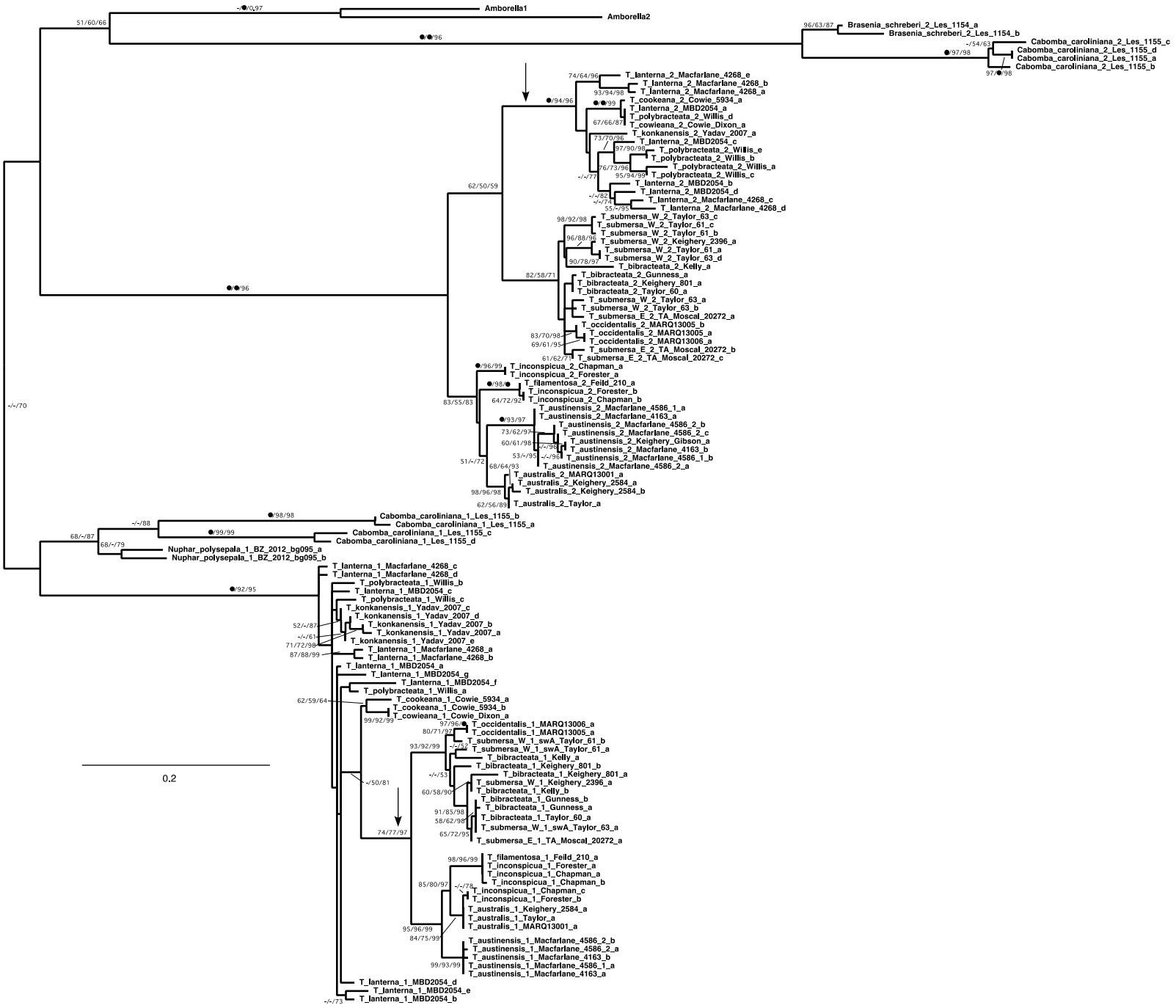



Fig. 7. Maximum likelihood tree of Hydatellaceae inferred from concatenated analysis of the two *PGK* loci using representative variants from representative individuals and species for each taxonomic section. Each taxon name is followed by source information and variant tags (variant number corresponds to those in Fig. 4 and Suppl. Fig. 1). Numbers adjacent to braches are likelihood bootstrap support values ("--" = <50%).




Suppl. Fig. 1. Maximum likelihood tree of the nuclear phosphoglycerate kinase (PGK) loci for Hydatellaceae and relatives. The tree is rooted between two inferred copies, with *Amborella* arbitrarily shown as the sister group of one copy. Each taxon name is followed by the *PGK* gene copy number (1 vs. 2), voucher information (the collector name and number); see Fig. 4 for further details. Numbers adjacent to branches are bootstrap support values (respectively: parsimony and likelihood bootstrap values; posterior probabilities expressed as percentages. Filled circle: 100%; dash or no support value: <50%). Scale bar: expected substitutions per site. Arrows show the root of Hydatellaceae inferred in Iles et al. (2012, 2014), based on plastid data.

# BIBLIOGRAPHY

**Adderley S, Sun G. 2014.** Molecular evolution and nucleotide diversity of nuclear plastid phosphoglycerate kinase (*PGK*) gene in Triticeae (Poaceae). *Gene* **533**: 142–148.

**Alfaro ME, Holder MT. 2006.** The posterior and the prior in Bayesian phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* **37**: 19–42.

**Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990.** Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403–410.

**Amborella Genome Project. 2013.** The *Amborella* genome and the evolution of flowering plants. *Science* **342**: 1241089.

**Anderson LE, Advani VR. 1970.** Chloroplast and cytoplasmic enzymes: three distinct isoenzymes associated with the reductive pentose phosphate cycle. *Plant Physiology* **45**: 583–585.

**Angiosperm Phylogeny Group. 2009.** An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society* **161**: 105–121.

**Baldwin BG, Sanderson MJ, Porter JM, Wojciechowski MF, Campbell CS, Donoghue MJ. 1995.** The ITS region of nuclear ribosomal DNA–a valuable source of evidence on angiosperm phylogeny. *Annual of Missouri Botanical Garden*. **82**: 247–277.

**Barkmane J, Simpson BB. 2002.** Hybrid origin and parentage of *Dendrochilum acuiferum* (Orchidaceae) inferred in a phylogenetic context using nuclear and plastid DNA sequence data. *Systematic Botany* **27**: 209–220.

**Baumel A, Ainouche ML, Levasseur JE. 2001.** Molecular investigations in populations of *Spartina anglica* C.E. Hubbard (Poaceae) invading coastal Brittany (France). *Molecular Ecology* **10**: 1689–1701.

**Blanc G, Wolfe KH. 2004.** Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* **16**: 1667–1678.

**Brinkmann H, Martin W. 1996.** Higher-plant chloroplast and cytosolic 3-phosphoglycerate kinases: a case of endosymbiotic gene replacement. *Plant Molecular Biology* **30**: 65–75.

**Brochmann C, Nilsson T, Gabrielsen TM. 1996.** A classic example of postglacial allopolyploid speciation re-examined using RAPD markers and nucleotide sequences: *Saxifraga osloensis* (Saxifragaceae). *Symbolae Botanicae Upsalienses* **31**: 75–89.

**Buckler ES, Holtsford TP. 1996.** *Zea* ribosomal repeat evolution and substitution patterns. *Molecular Biology and Evolution* **13**: 623–632.

**Chen Q, Kang H-Y, Fan X, Wang Y, Sha L-S, Zhang H-Q, Zhong M-Y, Xu L-L, Zeng J, Yang R-W et al. 2013.** Evolutionary history of *Triticum petropavlovski* Udacz. et Migusch. inferred from the sequences of the 3-Phosphoglycerate Kinase gene. *PloS ONE* **8**: e71139.

**Coate JE, Schlueter JA, Whaley AM, Doyle JJ. 2011.** Comparative evolution of photosynthetic genes in response to polyploidy and nonpolyploid duplication. *Plant Physiology* **155**: 2081–2095.

**Cooke DA. 1987.** Hydatellaceae. In A. S. George [ed.], Flora of Australia, vol. 45, Hydatellaceae to Liliaceae, 1–5. Australian Government Publishing Service, Canberra, Australia.

**Costa M, Pereira AM, Rudall PJ, Coimbra S. 2013.** Immunolocalization of arabinogalactan proteins (AGPs) in reproductive structures of an early-divergent angiosperm, *Trithuria* (Hydatellaceae). *Annals of Botany* **111**: 183–190.

**de Lange PJ, Murray BG, Datson PM. 2004.** Contributions to a chromosome atlas of the New Zealand flora – 38. Counts for 50 families. *New Zealand Journal of Botany* **42**: 873–904.

**Doyle JJ. 1992.** Gene trees and species trees: molecular systematics as one-character taxonomy. *Systematic Botany* **17**: 144–163.

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

**Duarte JM, Wall PK, Edger PP, Landherr LL, Ma H, Pires JC, Leebens-Mack J, dePamphilis CW. 2010.** Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology* **10**: 1–18.

**Edgar RC. 2004.** MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792–1797.

**Felsenstein, J. 1985.** Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**: 783–791.

**Friedman WE. 2008.** Hydatellaceae are water lilies with gymnospermous tendencies. *Nature* **453**: 94–97.

**Friedman WE, Bachelier JB. 2013.** Seed development in *Trimenia* (Trimeniaceae) and its bearing on the evolution of embryo-nourishing strategies in early flowering plant lineages. *American Journal of Botany* **100**: 906–915.

**Friedman WE, Bachelier JB, Hormaza JI. 2012.** Embryology in *Trithuria submersa* (Hydatellaceae) and relationships between embryo, endosperm, and perisperm in early-diverging flowering plants. *American Journal of Botany* **99**: 1083–1092.

**Gaikwad SP, Yadav SR. 2003.** Further morphotaxonomical contribution to the understanding of family Hydatellaceae. *Journal of the Swamy Botanical Club* **20**: 1–10.

**Gilg-Benedict, C. 1930.** Centrolepidaceae. *In* A. Engler, and K. A. E. Prantl [eds.], Die Natürlichen Pflanzenfamilien, 15a, 27–33. W. Engelmann, Leipzig, Germany.

**Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS. 2012.** Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* **40**: D1178–D1186.

**Graham SW, Kohn JR, Morton BR, Eckenwalder JE, Barrett SCH. 1998.** Phylogenetic congruence and discordance among one morphological and three molecular data sets from Pontederiaceae. *Systematic Biology* **47**: 545–567.

**Graham SW, Olmstead RG. 2000.** Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *American Journal of Botany* **87**: 1712–1730.

**Hamann U. 1998.** Hydatellaceae. *In* K. Kubitzki [ed.], The families and genera of vascular plants, vol. IV, Flowering plants—Monocotyledons—Alismatanae and Commelinanae, 231–234. Springer, Berlin, Germany.

**Heled J, Drummond AJ. 2010.** Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* **27**: 570 – 580.

**Hieronymus G. 1888.** Centrolepidaceae. *In* A. Engler, and K. A. E. Prantl [eds.], Die Natürlichen Pflanzenfamilien, vol. II, part 4, 11–16. W. Engelmann, Leipzig, Germany.

**Hilu KW, Borsch T, Muller K, Soltis DE, Soltis PS, Savolainen V, Chase MW, Powell MP, Alice LA, Evans R et al. 2003.** Angiosperm phylogeny based on *mat*K sequence information. *American Journal of Botany* **90**: 1758–1776.

**Hoot SB, Culhama A, Crane PR. 1995.** The utility of *atp*B gene sequences in resolving phylogenetic relationships: Comparison with *rbc*L and 18S ribosomal DNA sequences in the Lardizabalaceae. *Annals of the Missouri Botanical Garden* **82**: 194–207.

**Huang S, Sirikhachornkit A, Faris JD, Su X, Gill BS, Haselkorn R, Gornicki P. 2002.** Phylogenetic analysis of the acetyl-CoA carboxylase and 3-phosphoglycerate kinase loci in wheat and other grasses. *Plant Molecular Biology* **48**: 805–820.

**Huelsenbeck JP, Ronquist F. 2001.** MrBayes: Bayesian inference of phylogeny. *Bioinformatics* **17**:754–755.

**Iles WJD, Lee C, Sokoloff DD, Remizowa MV, Yadav SR, Barrett MD, Barrett RL, Macfarlane TD, Rudall PJ, Graham SW. 2014.** Reconstructing the age and historical biogeography of the ancient flowering-plant family Hydatellaceae (Nymphaeales). *BMC Evolutionary Biology* **14**: 102**.**

**Iles WJD, Rudall PJ, Sokoloff DD, Remizowa MV, Macfarlane TD, Logacheva MD, Graham SW. 2012.** Molecular phylogenetics of Hydatellaceae (Nymphaeales): Sexual-system homoplasy and a new sectional classification. *American Journal of Botany* **99**: 663–676.

**Kim KJ, Jansen RK. 1995.** *ndh*F sequence evolution and the major clades in the sunflower family. *Proceedings of the National Academy of Sciences, USA* **92**: 10379–10383.

**Kynast RG, Joseph JA, Pellicer J, Ramsay MM, Rudall PJ.** Chromosome behavior at the base of the angiosperm radiation: karyology of *Trithuria submersa* (Hydatellaceae, Nymphaeales). In press, *American Journal of Botany*.

**Lanfear R, Calcott B, Ho SYW, Guindon S. 2002.** PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution* **29**: 1695–1701.

**Lanfear R, Calcott B, Kainer D, Mayer C, Stamakakis A. 2014.** Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evolutionary Biology* **14**: 82.

**Löhne C, Borsch T, Wiersema JH. 2007**. Phylogenetic analysis of Nymphaeales using fast-evolving and noncoding chloroplast marker. *Botanical Journal of the Linnean Society* **154**: 141–163.

**Longstaff M, Raines CA, McMorrow EM, Bradbeer JW, Dyer T. 1989.** Wheat phosphoglycerate kinase: evidence for recombination between the genes for the chloroplastic and cytosolic enzymes. *Nucleic Acids Res* **17**: 6569–6580

**Maddison WP. 1997.** Gene trees in species trees. *Systematic Biology* **46**: 523–526.

**Maere S, de Bodt S, Raes J, Casneuf T, Van Montagu M, Van de Peer Y. 2005.** Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences* USA **102**: 5454–5459.

**Martin W, Schnarrenberger C. 1997.** The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. *Current Genetics* **32**: 1–18.

**Mathews S, Donoghue MJ. 1999.** The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* **286**: 947–950.

**Mathews S, Donoghue MJ. 2000.** Basal angiosperm phylogeny inferred from duplicate Phytochromes A and C. *International Journal of Plant Sciences* **161**: S41–S55.

**McDade LA. 1992.** Hybrids and phylogenetic systematics II. The impact of hybrids on cladistic analysis. *Evolution* **46**: 1329–1346.

**Moore MJ, Hassan N, Gitzendanner MA, Bruenn RA, Croley M, Vandeventer A, Horn JW, Dhingra A, Brockington SF, Latvis M et al. 2011.** Phylogenetic analysis of the plastid inverted repeat for 244 species: Insights into deeper-level angiosperm relationships from a long, slowly evolving sequence region. *International Journal of Plant Sciences* **172**: 541–558.

**Moore WS. 1995.** Inference of phylogenies from mtDNA variation: mitochondrial-gene tree versus nuclear-gene trees. *Evolution* **49**: 718–726.

**Ness RW, Graham SW, Barrett SCH. 2011.** Reconciling gene and genome duplication events: using multiple nuclear gene families to infer the phylogeny of the aquatic plant family Pontederiaceae. *Molecular Biology and Evolution* **28**: 3009–3018.

**Olmstead RG, Sweere JA. 1994.** Combining data in phylogenetic systematics: An empirical approach using three molecular data sets in the Solanaceae. *Systematic Biology* **43**: 467–481.

**Omland KE, Funk DJ. 2003.** Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics* **34**: 397–423.

**Percy DM, Argus GW, Cronk QC, Fazekas AJ, Kesanakurti PR, Burgess KS, Husband BC, Newmaster SG, Barrett SCH, Graham SW. 2014.** Understanding the spectacular failure of DNA barcoding in willows (*Salix*): Does this result from a trans-specific selective sweep? In press, *Molecular Ecology*.

**Pillon Y, Johansen J, Sakishima T, Chamala S, Barbazuk WB, Roalson EH, Price DK, Stacy EA. 2013.** Potential use of low-copy nuclear genes in DNA barcoding: a comparison with plastid genes in two Hawaiian plant radiations. *BMC Evolutionary Biology* **13**: 35.

**Qiu Y-L, Li L, Wang B, Xue J-Y, Hendry TA, Li R-Q, Brown JW, Liu Y, Hudson GT, Chen ZD. 2010.** Angiosperm phylogeny inferred from sequences of four mitochondrial genes. *Journal of Systematics and Evolution* **48**: 391–425.

**Rambaut A. 2002.** Se-Al version 2.0a11 [computer program].

**Rambaut A, Drummond AJ. 2009.** Tracer v1.5 [computer program].

**Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW. 2010.** Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* **463**: 1079–1083.

**Remizowa MV, Sokoloff DD, Macfarlane TD, Yadav SR, Prychid CJ, Rudall PJ. 2008.** Comparative pollen morphology in the early-divergent angiosperm family Hydatellaceae reveals variation at the infraspecific level. *Grana* **47**: 81–100.

**Rudall PJ, Eldridge T, Tratt J, Ramsay MM, Tuckett RE, Smith SY, Collinson ME, Remizowa MV, Sokoloff DD. 2009.** Seed fertilization, development, and germination in Hydatellaceae (Nymphaeales): implications for endosperm evolution in early angiosperms. *American Journal of Botany* **96**: 1581–1593.

**Rudall PJ, Remizowa MV, Prenner G, Prychid CJ, Tuckett RE, Sokoloff DD. 2009.** Nonflowers near the base of extant angiosperms? Spatiotemporal arrangement of organs in reproductive units of Hydatellaceae and its bearing on the origin of the flower. *American Journal of Botany* **96**: 67–82.

**Rudall PJ, Sokoloff DD, Remizowa MV, Conran JG, Davis JI, Macfarlane TD, Stevenson DW. 2007.** Morphology of Hydatellaceae, an anomalous aquatic family recently recognized as an early-divergent angiosperm lineage. *American Journal of Botany* **94**: 1073–1092.

**Saarela JM, Rai HS, Doyle JA, Endress PK, Mathews S, Marchant AD, Briggs BG, Graham SW. 2007.** Hydatellaceae identified as a new branch near the base of the angiosperm phylogenetic tree. *Nature* **446**: 312–315.

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

**Sokoloff DD, Remizowa MV, Conran JG, Macfarlane TD, Ramsay MM, Rudall PJ. 2014.** Embryo and seedling morphology In *Trithuria lanterna* (Hydatellaceae, Nymphaeales): new data for infrafamilial systematics and a novel type of syncotyly. *Botanical Journal of the Linnean Society* **174**: 551–573.

**Sokoloff DD, Remizowa MV, Macfarlane TD, Conran JG, Yadav SR, Rudall PJ. 2013.** Comparative fruit structure in Hydatellaceae (Nymphaeales) reveals specialized pericarp dehiscence in some early-divergent angiosperms with ascidiate carpels. *Taxon* **62** (1): 40–61.

**Sokoloff DD, Remizowa MV, Macfarlane TD, Rudall PJ. 2008.** Classification of the early-divergent angiosperm family Hydatellaceae: one genus instead of two, four new species and sexual dimorphism in dioecious taxa. *Taxon* **57**: 179–200.

**Sokoloff DD, Remizowa MV, Macfarlane TD, Tuckett RE, Ramsay MM, Beer AS, Yadav SR, Rudall PJ. 2008.** Seedling diversity in Hydatellaceae: Implications for the evolution of angiosperm cotyledons. *Annals of Botany* **101**: 153–164.

**Sokoloff DD, Remizowa MV, Macfarlane TD, Yadav SR, Rudall PJ. 2011.** Hydatellaceae: A historical review of systematics and ecology. *Rheedea* **21** (2): 115–136.

**Soltis PS, Plunkett GM, Novak SJ, Soltis DE. 1995.** Genetic variation in *Tragopogon* species: Additional origins of the allotetraploids *T. mirus* and *T. miscellus* (Compositae). *American Journal of Botany* **82**: 1329–1341.

**Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC, Brockington SF, Refulio-Rodriguez NF, Walker JB, Moore MJ et al. 2011.** Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany* **98**: 704–730.

**Soltis PS, Soltis DE. 1991.** Multiple origins of the allotetraploid *Tragopogon mirus* (Compositae): rDNA evidence. *Systematic Botany* **16**: 407–413.

**Steele PR, Guisinger-Bellian M, Linder CR, Jansen RK. 2008.** Phylogenetic utility of 141 low-copy nuclear regions In taxa at different taxonomic levels in two distantly related families of rosids. *Molecular Phylogenetics and Evolution* **8**: 349–362.

**Strand AE, Leebens-Mack J, Milligan BG. 1997.** Nuclear DNA-based markers for plant evolutionary biology. *Molecular Ecology* **6**: 113–118.

**Swofford DL. 2003.** PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4. Sinauer, Sunderland, Massachusetts, USA.

**Taylor ML, Macfarlane TD, Williams JH. 2010.** Reproductive ecology of the basal angiosperm *Trithuria submersa* (Hydatellaceae). *Annals of Botany* **106**: 909–920.

**Taylor ML, Williams JH. 2012.** Pollen tube development in two species of *Trithuria* (Hydatellaceae) with contrasting breeding systems. *Sexual Plant Reproduction* **25**: 83–96.

**Terachi T, Ogihara Y, Tsunewaki K. 1987.** The molecular basis of genetic diversity among cytoplasms of *Triticum* and *Aegilops*. VI. Complete nucleotide sequences of the *rbcL* genes encoding Hand L-type Rubisco large subunits in common wheat and *Ae. Crassa* 4x. *Japanese Journal of Genetics* **62**: 375–387.

**Thiers B. 2011.** [continuously updated]. Index Herbariorum, part 1: The herbaria of the world. New York Botanical Garden, Bronx, New York, USA. Website http://sweergum.nybg.org/ih/ [accessed 31 October 2011].

**Wendel JF, Schnabel A, Seelanan T. 1995.** Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proceedings of the National Academy of Sciences* USA **92**: 280–284.

**Yadav SR, Janarthanam MK. 1994.** Hydatellaceae: a new family for the Indian flora with a new species. *Rheedea* **4**: 17–20.

**Yamashita J. Tamura M. 2000.** Molecular phylogeny of the Convallariaceae (Asparagales). *In* K. Wilson and D. Morrison [eds.], Monocots: Systematics and evolution, 387–400. CSIRO, Melbourne, Australia.

**Zgurski JM, Rai HS, Fai QM, Bogler DJ, Francisco-Ortega J, Graham SW. 2008.** How well do we understand the overall backbone of cycad phylogeny? New insights from a large, multigene plastid data set. *Molecular Phylogenetics and Evolution* **47**: 1232–1237.

**Zwickl DJ. 2006**. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD Thesis, University of Texas at Austin, Austin, Texas, USA.

# APPENDICES

Appendix 1. Specimen voucher and accession details. Collection information, herbarium in parenthesis (following abbreviations in Thiers [continuously updated]), gender of plant (if not cosexual).

***Trithuria austinensis*** **D. D. Sokoloff, Remizowa, T. D. Macfarl. & Rudall: Australia: Western Australia:** *Keighery B.J. & Gibson s.n.* (PERTH), Lake Pindicup, Pindicup Nature Reserve, PGK; *Macfarlane 4586* (PERTH?), ♀ 1—PGK, ♂ 2—PGK. *Macfarlane 4163 & Hearn* (PERTH?), ♀—PGK.

***Trithuria australis*** **(Diels) D.D. Sokoloff, Remizowa, T.D. Macfarl. & Rudall: Australia: Western Australia:** *Keighery G.J. & Gibson 2584* (PERTH), PGK; *Taylor s.n.* (TENN), PGK; *MARQ13001 & Macfarlane*, PGK, *atp*B, *mat*K, *ndh*F, *rbc*L; *MARQ13002 & Macfarlane*, *atp*B, *mat*K, *ndh*F, *rbc*L.

***Trithuria bibracteata*** **Stapf ex D.A. Cooke: Australia: Western Australia:** *Gunness & al. 13/37* (PERTH), PGK; *Keighery B.J. & Gibson 801* (PERTH), PGK; *Taylor 60* (TENN), PGK. *Kelly* (PERTH), PGK. *MARQ13003 & Macfarlane* (PERTH), *atp*B, *mat*K, *ndh*F, *rbc*L; *MARQ13004 & Macfarlane* (PERTH), *atp*B, *mat*K, *ndh*F, *rbc*L.

***Trithuria cookeana*** **D.D. Sokoloff, Remizowa, T.D. Macfarl. & Rudall: Australia: Northern Territory:** *Cowie 5934* (DNA), ♀, PGK.

***Trithuria cowieana*** **D.D. Sokoloff, Remizowa, T.D. Macfarl. & Rudall: Australia: Northern Territory:** *Cowie & Dixon s.n.* (DNA), PGK.

***Trithuria filamentosa*** **Rodway: Australia: Tasmania:** *Feild 210* (TENN), ♀, PGK.

***Trithuria inconspicua*** **Cheesem.: New Zealand: North Island:** *Chapman s.n.* (NSW), ♀, PGK. **South Island:** *Forester & Goh s.n.* (AK), ♀, PGK.

***Trithuria konkanensis*** **Yadav & Jarnarthanam: India: Maharashtra:** *Yadav s.n.* (MW), 2007,PGK.

***Trithuria lanterna*** **D.A. Cooke: Australia: Northern Territory:** *Macfarlane & al. 4268* (MW), PGK. *MBD2054* (PERTH), PGK.

***Trithuria occidentalis*** **Benth.: Australia: Western Australia:** *MARQ13005 & Macfarlane* (PERTH), ♀, PGK, *atp*B, *mat*K, *ndh*F, *rbc*L; *MARQ13006 & Macfarlane* (PERTH), ♂, PGK, *atp*B, *mat*K, *ndh*F, *rbc*L.

***Trithuria polybracteata*** **D.A. Cooke ex D.D. Sokoloff, Remizowa, T.D. Macfarl. & Rudall. Australia: Western Australia:** *Willis s.n.* (MEL), PGK.

***Trithuria submersa*** **Hook. f.: Australia: Tasmania:** *Moscal 20272* (HO), PGK. **Western Australia:** *Taylor 61* (TENN), PGK; *Taylor 63* (TENN), PGK. *Keighery B.J. & Gibson 2396*

(PERTH), PGK. *MARQ13007 & Macfarlane* (PERTH), *atp*B, *mat*K, *ndh*F, *rbc*L; *MARQ13008 & Macfarlane* (PERTH), *atp*B, *mat*K, *ndh*F, *rbc*L.

**Outgroups:**

***Amborella trichopoda***: Sequences are blasted from Amborella Genome Database. http://www.amborella.org/ ***Brasenia schreberi*:** *Les1155* (CONN), PGK. ***Cabomba caroliniana***: *Les1156* (CONN), PGK. ***Nuphar polysepala***: *Cuili Zhuang_2012-bg095* (UBC), PGK.

Appendix 2. Optimal DNA substitution models and partitioning scheme for different data sets, as chosen by PartitionFinder using the AICc for: (a) the plastid gene matrix; (b) the species-level analysis of the amplified portion of the PGK loci; (c) the angiosperm-wide analysis of the full coding sequence for the PGK loci.

(a)

| **Gene region** | **DNA substitution model** |
|---|---|
| *rbc*L codon position1 | GTR+I+G |
| *rbc*L codon position2 | HKY+I+ G |
| *rbc*L codon position3 | TIM+ G |
| *atp*B codon position1 | HKY+I |
| *atp*B codon position2 | HKY+I |
| *atp*B codon position3 | TVM+ G |
| *mat*K codon positions 1+2, *ndh*F codon position 2 | TVM+ G |
| *mat*K codon position3 | TVM+ G |
| *ndh*F codon postion1 | TVM+ G |
| *ndh*F codon postion3 | GTR+ G |

(b)

| **Gene region** | **DNA substitution model** |
|---|---|
| PGK exons 4+5 | GTR+I+ G |
| PGK intron | TVM+I |

(c)

| **Gene region** | **DNA substitution model** |
|---|---|
| PGK codon position1 | GTR+ G |
| PGK codon position2 | GTR+I+ G |
| PGK codon position3 | GTR+I+ G |

Appendix 3. List of clone variants observed for two different *PGK* copies in individuals from different species of *Trithuria* (Hydatellaceae), represented by different letters for each of the two *PGK* genes; numbers in brackets indicate how many times I obtained this variant in different clones. Distant alleles have been separated from each other in different lines. Specimen voucher information and the monophyly status of each species according to that *PGK* copy are also noted. Herbaria abbreviations follow Thiers (2011).

| **Section/species** | **Voucher information** | ***PGK-1*** | **Species monophyletic?** | ***PGK-2*** | **Species monophyletic?** |
|---|---|---|---|---|---|
| **Section *Altofinia*** | | | | | |
| *T. cookeana* | Cowie_5934 (DNA) | a(6)b(1) | No | a(1) | -- |
| *T. coweiana* | Cowie_Dixon (DNA) | a(3) | -- | a(5) | -- |
| **Section *Hamannia*** | | | | | |
| *T. konkanensis* | Yadav s.n. 2007 (MW) | a(2)b(1)c(1)d(1)e(1) | Yes | a(2) | -- |
| *T. lanterna* | Macfarlane & al. 4268 (MW) | a(7)b(1)<br>c(2)d(2) | No | a(4)b(1)c(1)<br>d(1)e(2) | No |
| | MBD2054 (PERTH) | a(2+3)<br>b(0+1)<br>c(1+1)<br>d(1+0)<br>e(2+0)f(3+1) | | a(1+1)<br>b(1+1)<br>c(0+1)<br>d(1+1) | |
| *T.polybracteata* | Willis s. n. (MEL) | a(1)<br>b(1)<br>c(3) | No | a(2)<br>b(1)c(1)<br>d(1)e(1) | No |

| Section/species | Voucher information | *PGK-1* | Species monophyletic? | *PGK-2* | Species monophyletic? |
|---|---|---|---|---|---|
| **Section *Hydatella*** | | | | | |
| *T. austinensis* | Keighery B. J. & Gibson s. n. (PERTH) | -- | Yes | a(8) | Yes |
| | Macfarlane 4163 & Hearn (PERTH?) | a(1)b(3) | | a(1)<br>b(3) | |
| | Macfarlane 4586 (PERTH)_1 | a(4) | | a(1)<br>b(3) | |
| | Macfarlane 4586 (PERTH)_2 | a(1)b(4) | | a(1)b(1)<br>c(1) | |
| *T. australis* | Keighery & Gibson 2584 (PERTH) | a(3) | Yes | a(7)b(3) | Yes |
| | Taylor s.n. (TENN) | a(4) | | a(4) | |
| | MARQ13001 & Macfarlane (PERTH) | a(4) | -- | a(4) | -- |
| *T. filamentosa* | Feild 210 (TENN) | a(5) | -- | a(5) | -- |
| *T. inconspicua* | Chapman s. n. (NSW) | a(3)b(1)<br>c(3) | No | a(3)<br>b(1) | No |
| | Forester & Goh s. n. (AK) | a(2)<br>b(3) | | a(4)<br>b(1) | |

| Section/species | Voucher information | *PGK-1* | Species monophyletic? | *PGK-2* | Species monophyletic? |
|---|---|---|---|---|---|
| **Section *Trithuria*** | | | | | |
| *T. bibracteata* | Gunness & al. 13/37 (PERTH) | a(3)<br>b(1) | No | a(5) | No<br>(but close) |
| | Kelly (PERTH) | a(1)<br>b(4) | | a(3) | |
| | Keighery B. J. & Gibson 801 (PERTH) | a(3)<br>b(2) | | a(3) | |
| | Taylor 60 (TENN) | a(4) | | a(4) | |
| *T. occidentalis* | MARQ13005 & Macfarlane (PERTH) | a(5) | Yes | a(3)(b(1) | Yes |
| | MARQ13006 & Macfarlane (PERTH) | a(5) | | a(3) | |
| *T. submersa* E | Moscal-E 20272 (HO) | a(3) | -- | a(2)<br>b(2)c(1) | No |
| *T. submersa* W | Taylor 61-W (TENN) | a(2)<br>b(1) | No | a(2)b(1)<br>c(1) | No |
| | Taylor 63_W (TENN) | a(2+2) | | a(0+3)<br>b(2+2)<br>c(3+1)d(1+1) | |