Reconciling Gene and Genome Duplication Events: Using Multiple Nuclear Gene Families to Infer the Phylogeny of the Aquatic Plant Family Pontederiaceae

Rob W. Ness,*[,1] Sean W. Graham,[2,3] and Spencer C. H. Barrett[1]

[1]Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada

[2]UBC Botanical Garden and Centre for Plant Research, University of British Columbia, Vancouver, British Columbia, Canada

[3]Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada

***Corresponding author:** E-mail: rob.ness@utoronto.ca.

Associate editor: Charles Delwiche

Abstract

Most plant phylogenetic inference has used DNA sequence data from the plastid genome. This genome represents a single genealogical sample with no recombination among genes, potentially limiting the resolution of evolutionary relationships in some contexts. In contrast, nuclear DNA is inherently more difficult to employ for phylogeny reconstruction because major mutational events in the genome, including polyploidization, gene duplication, and gene extinction can result in homologous gene copies that are difficult to identify as orthologs or paralogs. Gene tree parsimony (GTP) can be used to infer the rooted species tree by fitting gene genealogies to species trees while simultaneously minimizing the estimated number of duplications needed to reconcile conflicts among them. Here, we use GTP for five nuclear gene families and a previously published plastid data set to reconstruct the phylogenetic backbone of the aquatic plant family Pontederiaceae. Plastid-based phylogenetic studies strongly supported extensive paraphyly of *Eichhornia* (one of the four major genera) but also depicted considerable ambiguity concerning the true root placement for the family. Our results indicate that species trees inferred from the nuclear genes (alone and in combination with the plastid data) are highly congruent with gene trees inferred from plastid data alone. Consideration of optimal and suboptimal gene tree reconciliations place the root of the family at (or near) a branch leading to the rare and locally restricted *E. meyeri*. We also explore methods to incorporate uncertainty in individual gene trees during reconciliation by considering their individual bootstrap profiles and relate inferred excesses of gene duplication events on individual branches to whole-genome duplication events inferred for the same branches. Our study improves understanding of the phylogenetic history of Pontederiaceae and also demonstrates the utility of GTP for phylogenetic analysis.

Key words: gene tree parsimony, gene tree reconciliation, gene duplication, nuclear phylogenetics, Pontederiaceae.

Introduction

Phylogenetic inference provides an historical framework for understanding fundamental evolutionary processes, such as speciation and adaptation. Phylogenetic reconstruction in plants has largely used data from the plastid (chloroplast) genome, although low-copy nuclear genes are increasingly employed (e.g., Steele et al. 2008; Duarte et al. 2010; Regier et al. 2010), and the growing availability of large genomic data sets from multiple taxa has the potential to revolutionize the inference of plant phylogeny. Nonetheless, the plastid genome continues to be the molecule of choice for phylogenetic studies for a variety of reasons (e.g., Olmstead and Palmer 1994; Graham and Olmstead 2000). For example, it has a conservative DNA substitution rate across a range of genes and a relatively conserved gene order in most land plants. It also has strong conservation in gene number, with generally only one gene copy per genome; plastid genes are occasionally lost but never transferred from other genomes. These properties reduce complications associated with primer design, sequence recovery, and orthology assignment across divergent taxa. However, the entire plastid genome is a single nonrecombining linkage group, which means that different plastid genes do not provide completely independent records of phylogenetic history (e.g., Doyle 1992; Maddison 1997).

The nuclear genome, in contrast, represents a potentially much larger source of information on phylogenetic relationships from multiple independent linkage groups. However, the relatively rapid evolution of nuclear genes, coupled with the overall fluidity of the genome in terms of gene copy number and order, make phylogenetic inferences more challenging. A particular difficulty is uncertainty in the orthology of the characters used. Orthology of genes is almost always unambiguous for plastid data, but nuclear genes are often part of gene families that undergo repeated rounds of expansion and contraction in copy number due to gene duplication and extinction processes (e.g., Kellogg and Bennetzen 2004). Over time these processes can obscure relations among homologous loci. Incorrect assumptions about gene orthology, particularly mistaking paralogs for orthologs, can lead to conflicting

gene trees and uncertain or distorted inferences of the overall species tree from gene tree data (Maddison 1997).

The increasingly inexpensive recovery of large-scale nuclear data sets through advances in DNA sequencing technology has encouraged the development of theoretical and analytical techniques for assessing gene orthology. Orthology assessment requires reconciling apparent conflicts within or among phylogenies inferred from multigene families, which is the basis of the gene tree parsimony (GTP) method (Goodman et al. 1979; Page and Charleston 1997; Slowinski et al. 1997). GTP considers both congruence and conflict in one or a collection of gene tree genealogies, using both to infer an overall (rooted) species tree that minimizes the number of gene duplications, gene losses, or deep coalescences. It does so by reconciling any conflicts among the genealogies considered. When considering multigene families, the reconciliation cost of a tree can be calculated simply as the minimum number of gene duplications needed to reconcile the gene trees to the observed species tree. Gene losses can also be considered as part of the reconciliation cost, but this is not recommended where there is the possibility of incomplete sampling of all members of the gene family in each species (Page and Charleston 1997). The tree with the lowest reconciliation cost among and within gene trees is preferred as the best estimate of species phylogeny. GTP has been implemented in three programs, GeneTree (Page 1998), Gtp (Sanderson and McMahon 2007), and DupTree (Wehe et al. 2008), and has been applied to data from a variety of taxonomic groups (e.g., Maier et al. 2001; Frajman et al. 2009; Holton and Pisani 2010). The advantage of GTP for phylogenetic inference using nuclear sequences is that no a priori knowledge about the orthology of gene copies is necessary to infer species trees; indeed, the reconciliation of conflict among gene genealogies provides core evidence for the overall pattern of species relationships.

Here, we use GTP to reconstruct the species phylogeny of the aquatic flowering plant family Pontederiaceae (Monocotyledoneae: Commelinales, Cantino et al. 2007; Angiosperm Phylogeny Group 2009) using five nuclear gene trees and previously published plastid data (Graham et al. 1998, 2002). This small family is composed of ~35–40 species from five genera: *Eichhornia* (8–9 spp.), *Heteranthera* (13–16 spp., including several species sometimes included in segregate genera), *Monochoria* (7–8 spp.), *Pontederia* (6–9 spp.), and *Hydrothrix* (1 sp.) (Barrett and Graham 1997; Barrett 2004). Members of the family are largely concentrated in the New World tropics, particularly lowland South America and Brazil (Barrett 1978). They display a remarkable diversity of life history and reproductive strategies, ranging from highly clonal, long-lived taxa that inhabit permanent marshes and river systems, to exclusively sexual species that are annual and occur in ephemeral pools, ditches, and rice fields. Linking these extremes are species with various combinations of sexual and asexual reproduction and a variety of floral strategies (tristyly and enantiostyly) and mating systems, including self-incompatible and self-compatible taxa. Evolutionary studies of the family over the past three decades have focused primarily on selected taxa that possess tristylous and monomorphic (homostylous) reproductive systems (reviewed in Barrett 1988, 1993; Barrett et al. 1992). Phylogenetic reconstructions using both morphological (Eckenwalder and Barrett 1986) and plastid sequence data (Graham and Barrett 1995; Kohn et al. 1996; Graham et al. 1998, 2002) have been employed to investigate character evolution and the systematic relationships of taxa within the family and its close relatives.

Four published data sets have been used to investigate systematic relationships and character evolution in Pontederiaceae. The first was based on morphological characters, and the resulting trees were poorly supported, likely due to sampling error caused by the relatively small data set (Eckenwalder and Barrett 1986). Later, plastid DNA restriction site variation (Kohn et al. 1996) and DNA sequences from the plastid genes *rbc*L and *nhd*F (Graham and Barrett 1995; Graham et al. 1998, 2002) resulted in well supported trees that were mutually congruent but incongruent with the poorly supported trees inferred from morphological data (Graham et al. 1998). The plastid genealogy strongly supported the monophyly of Pontederiaceae and three of the major genera (*Pontederia*, *Monochoria*, and *Heteranthera*). In contrast, *Eichhornia* was inferred to be nonmonophyletic and to comprise four distinct lineages, two of which are polyploid. Our ability to investigate the polyploid origins of these species has been limited by the nonrecombining nature of the plastid genome. In addition, there remains substantial ambiguity in the location of the root of the phylogeny (see Graham et al. 2002). Nuclear DNA sequence data may provide further insights into these phylogenetic issues.

In this study, we present new data from five nuclear gene families recovered using primers designed from genes from an earlier expressed sequence tag (EST) study (Ness et al. 2010). We used GTP on these nuclear genealogies to infer relationships among 14 exemplar (representative) species that we chose to encompass the broad phylogenetic backbone of Pontederiaceae. Our study addressed the following specific issues: 1) How congruent is the species tree inferred by using GTP to reconcile the five nuclear gene families with that inferred from plastid data alone? 2) Can GTP clarify our understanding of the placement of the root of Pontederiaceae, which is unclear from plastid data alone? 3) GTP assumes that the gene genealogies being reconciled have been inferred correctly, so how can we incorporate estimates of the uncertainty in gene tree inferences when inferring the species tree? and 4) Do phylogenetic inferences of where the gene duplication events occurred correspond to inferences of past episodes of polyploidization in the family?

Materials and Methods

Taxon Sampling

We selected 14 species of Pontederiaceae including species from across the major lineages, based on our current understanding of the phylogeny of the family from the plastid-based analysis of Graham et al. (2002). The species

sampled for nuclear data were: *E. azurea*, *E. crassipes*, *E. meyeri*, *E. paniculata*, *E. paradoxa*, *Eichhornia* sp., *H. multiflora*, *H. seubertiana*, *H. zosterifolia*, *Hy. gardneri*, *M. hastata*, *M. korsakovii*, *P. sagittata*, and *P. subovata*. Source information for these species is presented in supplementary table S1, Supplementary Material online. The taxon we refer to as *Eichhornia* sp. is an undescribed species that was originally identified by Eckenwalder and Barrett (1986) as *E. paradoxa*. However, subsequent evidence indicates that it is a distinct species from *E. paradoxa*, based on morphological differences, hybrid sterility, isozyme differentiation, and plastid DNA sequence variation (Kohn et al. 1996; Barrett SCH, unpublished data).

Primer Development and Sequencing

Primers for nuclear loci were developed based on EST sequences derived from *E. paniculata* and *E. paradoxa* (for details of EST sequencing, see Ness et al. 2010). We identified ESTs with conserved DNA sequence in both *Oryza sativa* and *Zea mays*, under the assumption that these sequences would be more similar across the species used in our study. We annotated the selected ESTs using BLAST2GO (Conesa et al. 2005; Götz et al. 2008), a program that uses BlastX to identify all putative homologs in the nonredundant protein database at NCBI. We designed degenerate primers and used them to amplify gene regions from genomic DNA extractions. We chose a subset of six primer pairs that most reliably amplified across all samples. These loci (described in table 1) were annotated by BLAST2GO as putative homologs of a protein phosphatase (serine/threonine phosphatase), a coatamer subunit, a nucleoside diphosphate kinase, an importin, a protochlorophyllide reductase, and a DNA-J like protein. We cloned the products of each polymerase chain reaction (PCR) and sequenced both forward and reverse strands in four to eight clones per amplification per species using an ABI 3730XL florescent-based capillary sequencer. We assembled forward and reverse sequence strands with Sequencher 4.7 and confirmed all genotypes manually. Sequences from each of the six amplifications were aligned using Muscle (Edgar 2004). We manually adjusted all alignments to ensure the most accurate alignment (e.g., Graham et al. 2000), and sequences with no similarity to any other sample were assumed to be cloning or PCR artifacts and were therefore discarded. Three of the six primer pairs amplified loci with conserved exons flanking a highly variable intron. In these cases, the introns were too divergent amongst the species to align reliably and so we excluded them from subsequent analyses. We also excluded near identical sequence reads ($<$1% of sites were variable) derived from the same species to avoid including allelic variants or sequences that only varied due to errors introduced during amplification, cloning, or sequencing. We excluded one candidate gene family that had limited recovery across the taxa that were sampled because it recovered clearly nonhomologous fragments in different species, presumably due to low specificity of amplification. However, including homologous sequences from this gene family in the analysis has little or no effect on the results or conclusions (data not shown).

Reconstructing Genealogical and Phylogenetic History

We generated maximum likelihood (ML) genealogies for each of the five alignments using the software Garli v1.0 (Zwickl 2006) with a general time reversible (GTR) $+\mathrm{I}+\Gamma$ model in which the substitution matrix, proportion of invariant sites (I), and shape of the gamma distribution (Γ, with four rate categories for the shape parameter α) were estimated from the data. In addition, we generated 2,000 bootstrap replicate genealogies for each alignment to assess support for the genealogies and for use in later analyses.

To fit the best overall species tree for the five genealogies, we used the program DupTree (Wehe et al. 2008). DupTree generates rooted species trees from multiple genealogies using GTP (Goodman et al. 1979; Page and Charleston 1997; Slowinski et al. 1997). It implements a standard rooted subtree pruning and regrafting heuristic search to identify the optimal rooted species tree topology, which is the one that minimizes the number of inferred gene duplications necessary to reconcile each genealogy with this candidate species tree. DupTree minimizes only the number of duplications during reconciliation rather than considering both duplications and losses as part of the reconciliation cost because incomplete taxon and sequence sampling may be erroneously interpreted as gene copy losses (Page and Charleston 1997). We used the unrooted ML genealogies generated in Garli v1.0 as input for DupTree. Local search heuristics are not guaranteed to find the global optimum and so we reran DupTree 1,000 times using random starting trees to increase the probability that all the shortest trees were found.

DupTree assumes that the gene genealogies used are fully resolved (bifurcating) and does not take into account uncertainty in each gene tree. We therefore attempted to assess the impact of uncertainty in genealogical inferences using two methods. In the first case ("Bootstrap-1"), we evaluated the sensitivity of the inferred species tree to uncertainty in each of the five genealogies, in turn. We ran 1,000 iterations of DupTree where one of the five gene trees was represented by an individual tree inferred from a bootstrap pseudoreplicate, generated in Garli v1.0, and the best ML gene tree was used for the other four. We repeated this for each of the five gene trees and summarized the effect of each on the species tree by calculating the proportion of iterations that supported each clade pooled across results from the separate analyses of each locus. We computed this using the majority-rule consensus function in PAUP* (Swofford 2003). In the second method ("Bootstrap-2"), we assessed the effect of uncertainty in genealogical inferences of all alignments simultaneously. This approach is similar to a previously published method from Cotton and Page (2002). We ran 1,000 DupTree iterations in which we simultaneously sampled single ML bootstrap pseudoreplicate genealogies for each of the five gene families per iteration. We again summarized the results as the fraction of analyses that recovered each clade of interest across iterations.

We repeated all of these analyses considering the unrooted plastid-based genealogy of Graham et al. (2002)

Table 1. Summary Information on the Six ESTs Used to Design Primers for Amplifying Nuclear Regions in Pontederiaceae.

EST label	Blast Protein Description[a]	Number of Sequences	Unique Sequences[b]	Aligned Length (bp)	Parsimony Informative Sites	Singletons
EX0014	Protein phosphatase	59	18	334	65	24
EX0042	Coatomer subunit epsilon	21	15	441	68	56
P0133	Nucleoside diphosphate kinase	39	22	255	55	19
P0249	Importin alpha	61	32	1,040	271	69
P0263	Protochlorophyllide reductase precursor	46	21	454	101	59
P0508[c]	DNA-J like protein	25	10	567	101	63

[a] Protein descriptions are based on BLAST2GO annotations of EST translations.
[b] Number of sequences included after removing nearly identical (>99% identical) sequences from the same species.
[c] This locus was excluded from all analyses presented in this paper.

as an additional gene tree. We pruned taxa from the plastid dataset that were not included in our study to maintain similar taxon sampling and interpolated the position of two taxa sampled here but not in the plastid data. Specifically, *P. subovata* was interpolated as the sister group of *P. sagittata* in the plastid tree (representing an assumption of monophyly for this genus and based on unpublished plastid sequence data), and *H. multiflora* was interpolated in the position of *H. reniformis* in the study of Graham et al. (2002), as these three *Heteranthera* species were predicted to be closely related to each other by Horn (1985).

Cost of Suboptimal Roots of the Species Tree

The current plastid genealogy of Pontederiaceae is well resolved, but its exact rooting remains unclear (see Graham et al. 2002). We used the program GeneTree v1.3.0 (Page 1998) to calculate the reconciliation cost for rerooting the overall species tree, estimated from each of the five nuclear gene trees and the plastid data, on all of its suboptimal branches. We reconciled each of the five nuclear gene tree genealogies in turn to the optimal species tree (GeneTree can only perform pairwise reconciliations), repeating this for each of its 26 possible roots, and then summing the minimum number of duplications required to fit each of the genealogies for each rerooting. GeneTree requires rooted genealogies as input for this analysis. To root the individual nuclear gene trees generated in Garli v1.0, we used the most common, shortest rooted genealogy for each alignment that was estimated by running 1,000 iterations of DupTree; rooted gene trees are part of the output from the tree reconciliation exercise. We used this method to examine suboptimal species tree roots, including those that are nearly optimal. Using GeneTree in this manner might be expected to recover the same best species tree as DupTree, as the two programs use the same underlying reconciliation principle.

Gene Duplication and Polyploidy

We assessed correspondence between instances of whole-genome duplications (including paleopolyploidy events inferred on internal branches) and the number and position of inferred gene duplications on one of the four species tree inferred from the nuclear and plastid gene trees (see below). We first mapped polyploidization or demi-polyploidization events onto the tree using the program ChromEvol v1.1 (Mayrose et al. 2010). ChromEvol defines polyploidization as a doubling of the haploid chromosome number, whereas demi-polyploidization is an increase of 1.5× in the chromosome number due to the union of a reduced and an unreduced gamete, eventually leading, for example, to hexaploid formation (Mayrose et al. 2010). The program uses the current chromosome number of each species and the rooted species tree to reconstruct ancestral changes in chromosome number in a phylogeny. We estimated the parameters for each of eight models included with the program and chose the model with the best fit based on the Akaike information criterion. The model chosen estimates a constant rate of chromosome loss and gain and constrained the rate of demi-polyploidization and polyploidization to be equal and constant across the tree. This allowed us to map changes in ploidy along each branch in the tree. We then used GeneTree to identify the location on the tree where gene duplications occurred. The number of gene duplications inferred for each branch was summed across all five nuclear gene tree reconciliations, and these were visually compared with the occurrence of partial or whole-genome duplications.

Results

In total, we recovered 226 alignable sequences for the five primer pairs totaling 2,524 bp of aligned nuclear sequence per taxon, after intron exclusion. Following the removal of nearly identical sequences from the same species, 108 "unique" sequences were identified (mean 21.6 sequences/alignment). In this reduced set, there were 560 parsimony informative sites (table 1). The bootstrap majority consensus for each of the five unrooted genealogies that we reconstructed is presented in figure 1. The proportion of branches resolved with greater than 50% support varies from 0.57–0.80, with a mean bootstrap value of 82.7% for resolved branches across the five genealogies. As expected in genealogies with paralogous sequences, the relations among species in each of the genealogies did not unambiguously reflect the known relationships or the taxonomy of the family.

Nuclear-Based Species Tree

The rooted species tree inferred from the five genealogies is contrasted with the plastid topology of Graham et al. (2002; modified here with two interpolated taxa) in figure 2. Using

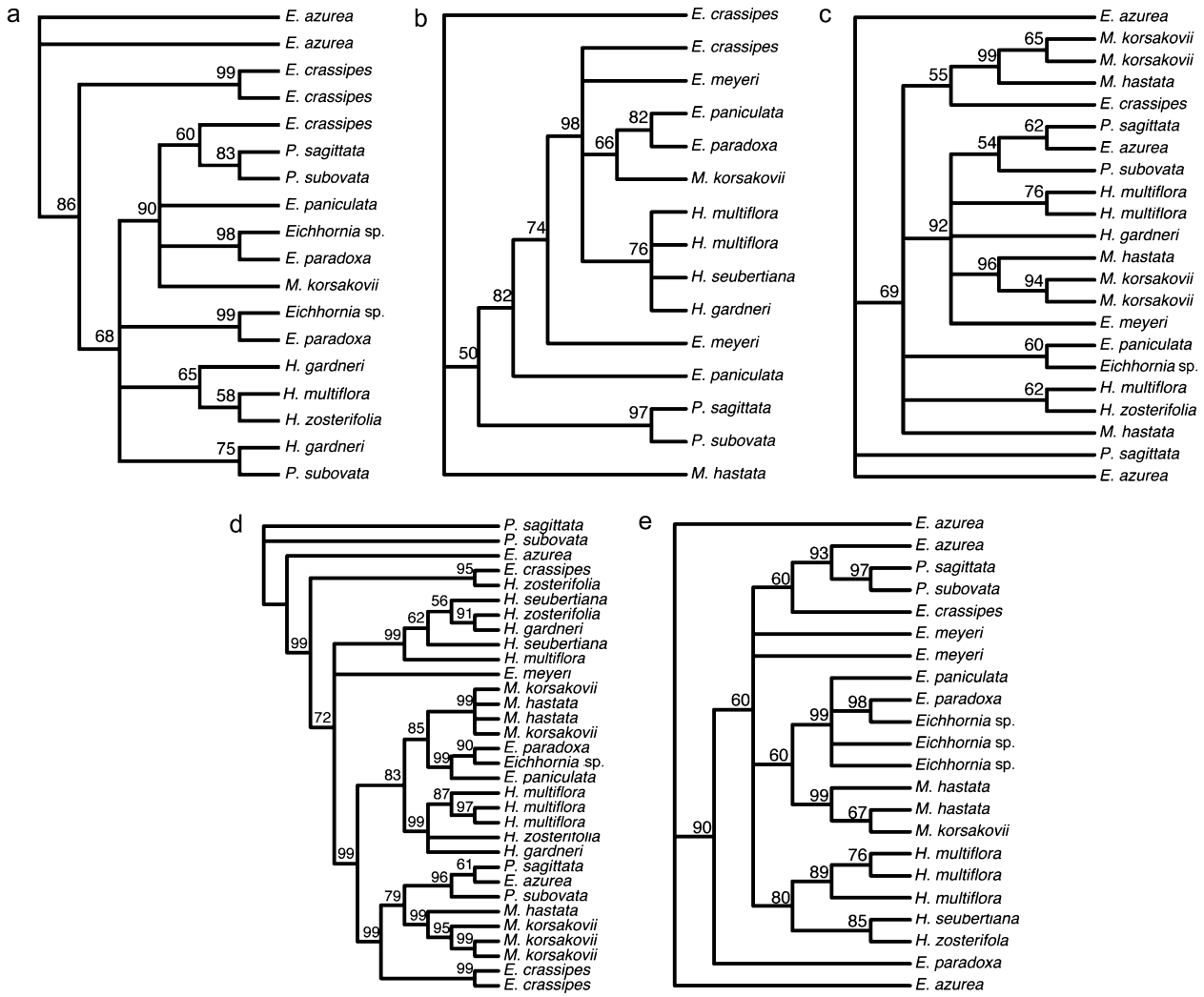

Fig. 1. ML reconstructions of each of five nuclear gene trees generated using Garli v1.0. Bootstrap support values from 500 replicates are shown on each branch. The trees correspond to those generated using the following primer sets: (*a*) EX0014, (*b*) EX0042, (*c*) P0133, (*d*) P0249, and (*e*) P0263. These unrooted gene trees are depicted with arbitrary roots for display purposes only.

our Bootstrap-1 method to assess support for this tree, we found 10 of 12 branches were supported by at least 70% of the iterations. On average, 73% of bootstrap replicates from each of the five genealogies supported a rooting of Pontederiaceae on the branch leading to *E. meyeri* as the best choice. The nuclear-based tree is similar to the current plastid topology with two exceptions. First, *Hy. gardneri* is nested within *Heteranthera s.s.*, whereas in the most recent plastid trees, *Hy. gardneri* is depicted as the sister group of *Heteranthera* Second, unlike the plastid-based tree, the nuclear tree did not support the monophyly of *Pontederia* but instead indicated that *P. subovata* was sister to *E. azurea* and *P. sagittata*, among taxa sampled here.

Combined Nuclear and Plastid Species Tree

We recovered four shortest species trees in this case; one of the four trees matched the unrooted plastid genealogy in its underlying topology, and we used this tree for display purposes (fig. 3). All subsequent analyses that are presented here are based on this species tree; however, the results are consistent across all four trees. We assessed the effect of uncertainty in individual genealogies on the species tree using the two methods that incorporate bootstrap pseudoreplicates from each gene family. In both cases, a high proportion of species tree inferences using individual or multiple bootstrapped gene trees recovered the same clades found using the species tree inference illustrated in figure 3. This pattern remained when we bootstrapped one genealogy at a time, pooling results for each iteration (Bootstrap-1), and when we used bootstrap replicates for all genealogies simultaneously for an iteration (Bootstrap-2). In both cases, a high fraction of bootstrapped reconciliations supported a root for the species tree with *E. meyeri* as the sister group to the rest of the family (88% with Bootstrap-1 and 68% with Bootstrap-2). Most branches were recovered with at least 50% support from both bootstrap methods (excluding one branch in *Heteranthera* and the monophyly of *Pontederia* with Bootstrap-1).

Rooting the Pontederiaceae Phylogeny

We used the program GeneTree to calculate the reconciliation cost of every possible rooting of the

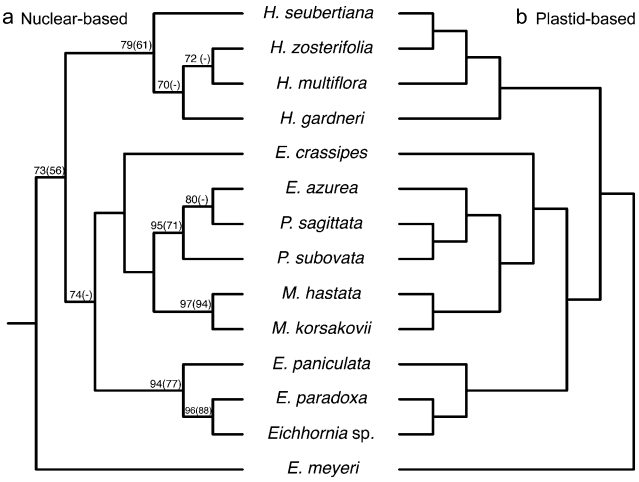

Fig. 2. *a*) Rooted species tree for the Pontederiaceae inferred in DupTree using five nuclear genealogies. The first support value was generated by sampling a bootstrap replicate for one of five gene trees in each iteration, with the remaining four input as the best ML gene trees (Bootstrap-1). Values in brackets were generated by randomly sampling bootstrapped gene trees for all five alignments for each iteration (Bootstrap-2). (*b*) Unrooted plastid genealogy from Graham et al. (2002) with interpolated positions of *Pontederia subovata* and *Heteranthera multiflora* (see text).

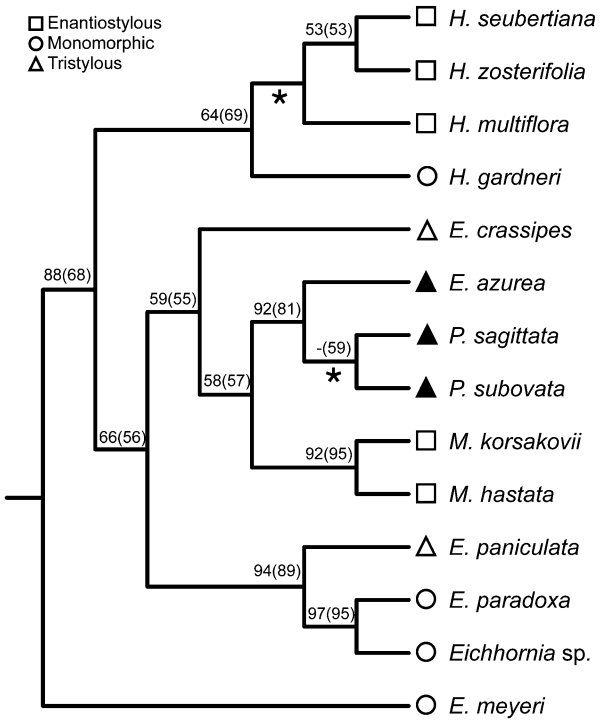

Fig. 3. One of four most parsimonious species trees based on simultaneous reconciliation of five nuclear gene trees and a plastid gene tree. Values on branches represent estimates of support made by considering bootstrapped gene trees during tree reconciliation, using our two bootstrap-based methods. The first value was generated by sampling a bootstrap replicate for one of five gene trees in each iteration, with the remaining four input as the best ML gene tree (Bootstrap-1). Values in brackets were generated by randomly sampling bootstrap genealogies for all five alignments for each iteration (Bootstrap-2). The two branches marked with asterisks (*) collapse in the strict consensus of the four most parsimonious trees. The floral form of each species is indicated at branch tips; squares, triangles, and circles refer to enantiostyly, tristyly, and floral monomorphism, respectively, and open symbols and closed symbols represent self-compatible or self-incompatible species, respectively.

Pontederiaceae phylogeny. The cost expressed in total number of duplications required to reconcile all five genealogies varied from 63 to 47 across the 26 possible roots. The reconciliation costs for each of the possible roots are shown in figure 4. As expected, because Duptree and GeneTree use the same underlying reconciliation method, both agree that the branch leading to *E. meyeri* is the lowest cost root for the species tree that reconciles all of the nuclear and plastid gene trees (figs. 3 and 4). The branch leading to *Heteranthera* and the branch connecting *Heteranthera* and *E. meyeri* to the rest of the family were the next two lowest cost possibilities, with one or two additional duplication(s) required to reconcile the genealogies compared with the optimal root.

Gene Duplication and Polyploidy

To investigate the association of polyploidy with gene duplications, we mapped polyploidization events and gene duplications onto the species tree reconstructed from the nuclear and plastid gene trees (fig. 5). We inferred four full genome duplications and one demi-polyploidization within the crown clade of Pontederiaceae; three of these polyploidization events occurred on the terminal branches leading to *E. azurea*, *E. crassipes*, and *M. korsakovii*, respectively, and the fourth occurred on the stem lineage leading to *Heteranthera*. With GeneTree, we inferred a total of 46 nuclear gene duplications distributed across 14 branches in the phylogeny, including 12 duplications on the stem lineage leading to the family (i.e., before the radiation of the extant species, assuming the current sampling represents the whole crown). All instances of polyploidization were associated with gene duplication events.

Discussion

A major finding of this study is that the species tree we obtained from five nuclear gene trees was broadly congruent with previous phylogenetic estimates of the Pontederiaceae using plastid data alone, and is reasonably robust relative to the uncertainty inferred in the underlying gene trees. We also recovered a single best solution for the root of Pontederiaceae using these new data, although there are also several nearly optimal alternatives to this rooting. Below we contrast our results with previous phylogenetic studies and discuss the potential causes of the relatively minor topological differences that were revealed between nuclear- versus plastid-based species trees. We also consider the implications of polyploidy and gene duplication for phylogenetic inference using nuclear sequence data.

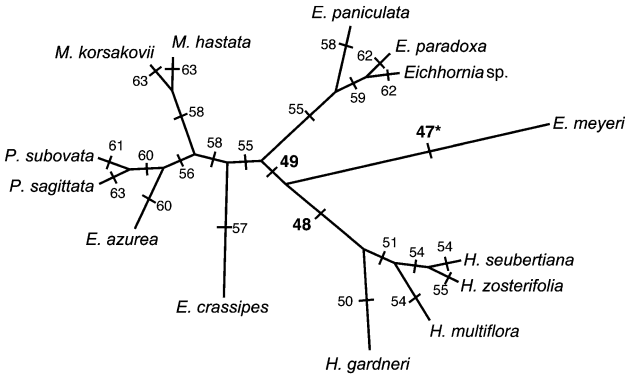

FIG. 4. Species tree of the Pontederiaceae based on combined nuclear and plastid sequences, presented as an unrooted cladogram, showing reconciliation costs for root placements on each branch. Reconciliation costs are expressed in terms of the number of gene duplications necessary to reconcile all five nuclear gene trees to the species tree based on nuclear and plastid data, when rooted on the branch indicated. The lowest cost branch is marked with an asterisk (*), and the next two lowest costs root placements are bolded.

Nuclear-Based Species Tree

Gene tree reconciliation as currently implemented does not attempt to take account of uncertainty in the underlying gene trees—these are assumed to be correct. However, individual gene trees are undoubtedly subject to stochastic error, and it would be appropriate to consider bootstrap support for the individual branches that are being reconciled. We attempted to do this by developing two different methods for incorporating the tree uncertainty captured by bootstrap analysis. The support for the species tree measured with our Bootstrap-1 method was quite strong (in this method, we examined the effect of including in each iteration, for each of the gene families, one pseudoreplicate bootstrap tree along with the best trees for the four other multigene families). Most species-tree branches in this case were recovered in at least 70% of iterations. Not surprisingly given the potential for a greater diversity of tree topologies in each iteration, support values from our Bootstrap-2 method were substantially lower (in this case, each iteration included only bootstrap pseudoreplicate gene trees from all five multigene families). The real robustness estimate for tree reconciliation may lie between these two estimates, but until a better method is developed for taking account of tree uncertainty we offer this approach as a possible means for estimating the certainty of species tree inferences using GTP. Despite the overall similarity between the five gene nuclear reconciliation presented here and previous estimates of phylogeny using plastid data alone (Graham et al. 1998, 2002), there were two instances of incongruence (fig. 2). One of these concerns *Heteranthera* (including *Hy. gardneri*) and the other the placement of *E. azurea* relative to the genus *Pontederia*.

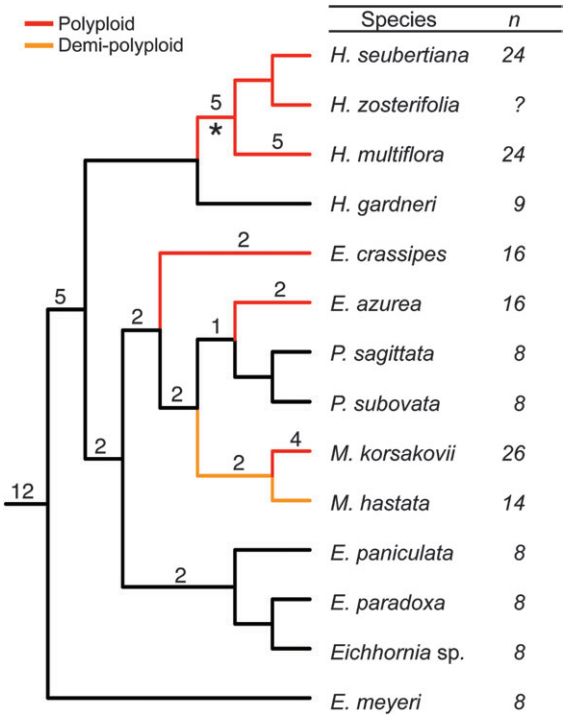

FIG. 5. Species tree of Pontederiaceae inferred using five nuclear gene trees and the plastid gene tree. Polyploid branches were mapped using ChromEvol, with inferred shifts marked in red for full genome duplication and orange for demi-polyploidization. Values indicate the inferred location and number of gene duplications made when reconciling individual nuclear gene trees to the species tree. Branch tips are labeled with the species name and haploid chromosome number (*n*) (Eckenwalder and Barrett 1986; Barrett 1988). The branch marked with an asterisk (*) indicates that both a demi-polyploidization and full genome duplication are inferred to have occurred on this branch.

First, our nuclear-based phylogeny was potentially incongruent with previous estimates of the phylogeny using plastid sequence data with respect to the placement of *H. seubertiana* relative to other species of *Heteranthera*. The nuclear-based species tree depicts *H. seubertiana* as the sister group of a clade comprising the remaining species of *Heteranthera*, defined broadly here to include *Hy. gardneri* (fig. 2), with weak to moderate support from the two different bootstrap methods. In contrast, previously published plastid data sets strongly group *H. seubertiana* in a clade with *H. zosterifolia* to the exclusion of the other two species sampled here. The earlier studies were also unable to robustly resolve the root of *Heteranthera* (in the largest study, Graham et al. 1998 recovered a basal trichotomy in *Heteranthera* between *Hy. gardneri*, *H. limosa*–*H. rotundifolia* and a clade of species that included *H. zosterifolia* and *H. seubertiana*).

One line of evidence supporting *Hy. gardneri* as sister to the rest of *Heteranthera* comes from analysis of chromosome number evolution. Reconstructions using ChromEvol (not shown) using the topology of *Heteranthera* in the nuclear-based species tree (fig. 2; left-hand tree) require two independent genome duplications in this clade rather than the single one inferred if *Hydrothrix* is assumed to have the placement observed in the plastid tree (fig. 3). Thus, a more parsimonious arrangement considering chromosome number is one that places the morphologically distinctive *Hy. gardneri* as sister to the rest of *Heteranthera*.

The second discordant result between our nuclear-based species tree and published plastid-based trees (e.g., Graham et al. 1998) is that *Pontederia* is not supported as monophyletic

(fig. 2: *P. subovata* is depicted as the sister group of *P. sagittata*–*E. azurea*). This result has only weak to moderate bootstrap support and so it may reflect stochastic error in one or more nuclear gene trees. Unpublished plastid sequence data from *ndh*F recovers the monophyly of *Pontederia*, placing *P. subovata* as sister to all other species of *Pontederia*. Our analysis with ChromEvol inferred a polyploidization event on the terminal branch leading to *E. azurea* (fig. 5), and it is possible that this genome duplication is the result of an ancient hybridization event between a species of *Pontederia* and the ancestor of *E. azurea*. However, if this were true, it would not necessarily lead to recovery of the species tree inferred using nuclear data seen in figure 2, as GTP assumes strictly bifurcating gene trees and species trees and is unable to accommodate discordance due to hybridization events (Page and Charleston 1997). Therefore, the recovery here of a relationship that is consistent with this type of polyploidization event would be at best a coincidence and at worst an artifact. Nonetheless, it should be possible to resolve this issue by additional sampling of taxa (more *Pontederia* species and species of *Eichhornia* inferred to be closely related to *E. azurea* in plastid-based analyses, such as *E. diversifolia* and *E. heterosperma*).

Combined Nuclear and Plastid Species Tree

When we combined our five nuclear gene trees with the plastid gene tree of Graham et al. (2002) in a joint GTP analysis, we recovered a species tree that was completely congruent with the previous plastid-based phylogeny (fig. 3). We also recovered three additional trees with the same reconciliation cost. The four shortest trees each have combinations of the two differences found in the nuclear-based tree discussed above (i.e., two locations of *E. azurea* × two topologies of *Heteranthera*). Gene duplications in the plastid tree are extremely unlikely. In fact, none have been observed in the long course of land plant evolution, with the exception of duplications in the plastid inverted repeat region, where duplicated genes do not diverge (Goulding et al. 1996). Nonetheless, the plastid tree is worth including in reconciliations, in the same manner that a strictly single-copy nuclear gene would be worth including in tree reconciliations, as it is an independent estimate of species phylogeny and thus could help uncover additional gene tree conflicts.

Corroboration of the general shape of the Pontederiaceae phylogeny supports previous conclusions derived from plastid DNA sequence (e.g., Graham et al. 1998, 2002). In particular, our results support the finding that the genus *Eichhornia* is highly paraphyletic. Modifications to the taxonomy of the family to take account of our findings could either recognize distinct genera to accommodate these lineages (note that as *E. azurea* is the type species of the genus, only it and its close relatives could retain this name), combine one or more genera (*Pontederia* has nomenclatural priority in the family), or employ rank-free taxonomy (i.e., the draft Phylocode, available online at www.phylocode.org).

Rooting the Pontederiaceae Species Tree

Rooting the Pontederiaceae tree has remained a difficult problem because of the erosion in phylogenetic signal of divergent outgroups which may lead to artifactual or ambiguous rootings. Graham et al. (2002) used multiple outgroups to root Pontederiaceae and reported two potential root locations for different optimality criteria, and a range of suboptimal rootings that could not be ruled out statistically (see fig. 5 in Graham et al. 2002). The optimal root based on parsimony split the family into *Heteranthera* (including *Hydrothrix*) versus all other species, whereas ML favored a root on the branch leading to both *Heteranthera* and *E. meyeri*. More extensive sequence data may help resolve the root of the plastid-based tree (Graham SW, unpublished data) but may still be subject to problems associated with the long branch connecting Pontederiaceae and its closest relatives. Our gene tree reconciliations did not include outgroup taxa but instead considered gene duplication evidence alone to root the family tree, which may be a useful alternative when the nearest sister group is distantly related to the ingroup (Mathews and Donoghue 1999).

Our analyses support a rooting of the family that places *E. meyeri* as an isolated lineage that is sister to the remainder of the family (figs. 2–4). Five gene duplications support this rooting (fig. 5), and there is also weak to moderately strong support for this considering our different bootstrap measures (figs. 2 and 3). However, there are several root placements implied by the gene duplication data that are nearly as optimal; only two gene duplications separate the reconciliation costs of the top three root placements. The second and third most likely root placements are the optimal placements found for the best ML and MP trees for plastid-based phylogenies in Graham et al. (2002). Thus, our results cannot be considered definitive, although the nearly congruent rooting arrangement with this earlier study is encouraging. Clearly, however, reconciliation of nuclear and plastid gene trees has provided new insight into the difficult problem of rooting the Pontederiaceae phylogeny and with larger sequence data sets and the addition of outgroup samples a solution to the problem may be achievable.

The finding that *E. meyeri* may be the sister group to the rest of Pontederiaceae is of particular interest for several reasons. First, the species is rather poorly known, having only been reported from a few localities in Northern Argentina, Paraguay, and Brazil, sites that are largely associated with the wetlands of the Grand Chaco and Matto Grosso. *Eichhornia meyeri* most closely resembles *E. paniculata* morphologically, which it has often been confused with in the literature (Eckenwalder and Barrett 1986). Earlier studies of the floral biology of this species (Barrett 1988) demonstrated that it was monomorphic for style length, with two sets of stamens positioned close to the stigmas. This arrangement results in high levels of autonomous self-pollination and resembles the semi-homostylous condition found in several tristylous species of *Eichhornia* in which tristyly has broken down, resulting in the evolution of self-pollinating forms (reviewed in Barrett 1988). However, the results from this study support the view that the monomorphic floral condition of *E. meyeri* is likely plesiomorphic (preceding the evolution of tristyly in the family). *Eichhornia meyeri* is self-compatible and the root position

inferred here for Pontederiaceae, with *E. meyeri* sister to the remainder of the family, is still consistent with our earlier studies in which self-incompatibility was inferred to be plesiomorphic (e.g., Barrett and Graham 1997), in contrast to most other self-incompatible flowering-plant families. The rarity of *E. meyeri* and its restricted distribution may be associated with progressive extinctions and loss of its wetland habitats during its long evolutionary history. The species clearly deserves conservation attention and additional study.

Gene Duplication and Polyploidy

Using contemporary chromosome numbers, we inferred four polyploidization events and one demi-polyploidization in the crown clade of Pontederiaceae (fig. 5). These numbers are almost certainly underestimates because other species in the family not included here have polyploid chromosome numbers (e.g., *Pontederia* includes other diploid and tetraploid taxa; Eckenwalder and Barrett 1986). Some taxa (e.g., *H. zosterifolia*, included in our study) have no chromosome number estimates. The polyploidization events we inferred may be one of the major causes of the 46 gene duplications revealed by our analysis, as 15 of the total gene duplications occur on branches where shifts in ploidy are inferred to have occurred. However, 12 gene duplications are inferred outside the crown clade. One possible explanation for this large number of early gene duplications is that they represent paralogous gene copies that either arose on the stem lineage or before the split of Pontederiaceae from its sister group, Haemodoraceae (Saarela et al. 2008); the former should be more informative about the root of the family, being closer in time to the root split and presumably less prone to saturation effects. Alternatively, the duplications could reflect incomplete sampling of members of a gene family. For example, divergent gene copies for which we have not sampled other orthologous sequences may be interpreted as duplications at the base of a tree because no information exists to support placement elsewhere. The number of gene duplications and their dispersion across the tree highlights the complexity of identifying orthologous sequences and the potential difficulties in using nuclear gene trees to infer species trees. Despite these complexities, the species tree inferred here is highly congruent with previous estimates based on plastid data alone, supporting the use of GTP as a method for mining further phylogenetic information from the nuclear genomes of these and other plants.

Conclusions

The aim of our study was to explore the utility of multiple nuclear gene trees for inferring the phylogenetic history of Pontederiaceae. Our investigations have demonstrated that the shape of the nuclear-based species tree is generally consistent with the plastid tree, and when we combined our nuclear gene trees with the plastid genealogy, we recovered a species tree that was completely congruent with the previous phylogenetic estimates. Our analysis also provided new evidence supporting a root placement in which *E. meyeri* is the sister group of the rest of the family, although several other placements are nearly as optimal. Lastly, by modeling the evolution of chromosome number, we showed that polyploidy could be responsible for a sizeable fraction of gene duplications in our gene trees. The history of gene and genome duplication complicates the relationships among homologous nuclear loci. However, using the phylogenetic signal present in multiple genealogies can provide a valuable method for inferring the relations among species by reconciling conflicts and clarifying the identity of orthologous versus paralogous gene copies. This approach will become increasingly relevant as the number of large-scale nuclear genome sequencing projects burgeons in future years.

Supplementary Material

Supplementary table S1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The research was supported by grants to S.C.H.B. and S.W.G. from the Natural Sciences and Engineering Research Council of Canada Discovery Grants Program, National Science Foundation grant DEB 0830036 to S.W.G., and a Premier's Discovery Award in Life Sciences and Medicine from the Ontario Government to S.C.H.B. R.W.N. was supported by student fellowships from the University of Toronto and the Canada Research Chair's Program to S.C.H.B.

References

Angiosperm Phylogeny Group. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot J Linn Soc*. 161:105–121.

Barrett SCH. 1978. Pontederiaceae. In: Heywood VH, editor. In: Flowering plants of the world. Oxford: Oxford University Press. p. 309–311.

Barrett SCH. 1988. Evolution of breeding systems in *Eichhornia*, a review. *Ann Mo Bot Gard*. 75:741–760.

Barrett SCH. 1993. The evolutionary biology of tristyly. In: Futuyma DJ, Antonovics J, editors. Oxford surveys in evolutionary biology. Oxford: Oxford University Press. p. 283–326.

Barrett SCH. 2004. Pontederiaceae (water hyacinth family). In: Smith N, Mori SA, Henderson A, Stevenson DW, Heald SV, editors. Flowering plants of the neotropics. Princeton (NJ): Princeton Unversity Press. p. 474–476.

Barrett SCH, Graham SW. 1997. Adaptive radiation in the aquatic plant family Pontederiaceae: insights from phylogenetic analysis. In: Givnish TJ, Sytsma K, editors. Molecular evolution and adaptive radiation. Cambridge (UK): Cambridge University Press. p. 225–258.

Barrett SCH, Kohn JR, Cruzan MB. 1992. Experimental studies of mating-system evolution: the marriage of marker genes and floral biology. In: Wyatt R, editor. Ecology and evolution of plant reproduction: new approaches. New York: Chapman & Hall. p. 192–230.

Cantino PD, Doyle JA, Graham SW, Judd WS, Olmstead RG, Soltis DE, Soltis PS, Donoghue MJ. 2007. Towards a phylogenetic nomenclature of Tracheophyta. *Taxon*. 56:822–846.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and

analysis in functional genomics research. *Bioinformatics* 21:3674–3676.
Cotton JA, Page RDM. 2002. Going nuclear: gene family evolution and vertebrate phylogeny reconciled. *Proc R Soc Lond B Biol Sci*. 269:1555–1561.
Doyle JJ. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. *Syst Bot*. 17:144–163.
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 Evol Biol*. 10:1–18.
Eckenwalder JE, Barrett SCH. 1986. Phylogenetic systematics of Pontederiaceae. *Syst Bot*. 11:373–391.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792–1797.
Frajman B, Eggens F, Oxelman B. 2009. Hybrid origins and homoploid reticulate evolution within *Heliosperma* (Sileneae, Caryophyllaceae)—a multigene phylogenetic approach with relative dating. *Syst Biol*. 58:328–345.
Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst Zool*. 28:132–163.
Götz S, García-Gómez JM, Terol J, et al. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res*. 36:3420–3435.
Goulding SE, Olmstead RG, Morden CW, Wolfe KH. 1996. Ebb and flow of the chloroplast inverted repeat. *Mol Gen Genet*. 252:195–206.
Graham SW, Barrett SCH. 1995. Phylogenetic systematics of Pontederiales: implications for breeding system evolution. In: Rudall PJ, Cribb PJ, Cutler DF, Humphries CJ, editors. Monocotyledons: systematics and evolution. Kew (UK): Royal Botanic Gardens. p. 415–441.
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. *Syst Biol*. 47:545–567.
Graham SW, Olmstead RG. 2000. Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *Am J Bot*. 87:1712–1730.
Graham SW, Olmstead RG, Barrett SCH. 2002. Rooting phylogenetic trees with distant outgroups: a case study from the commelinoid monocots. *Mol Biol Evol*. 19:1769–1781.
Graham SW, Reeves PA, Burns ACE, Olmstead RG. 2000. Microstructural changes in noncoding chloroplast DNA: interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference. *Int J Plant Sci*. 161:S83–S96.
Holton TA, Pisani D. 2010. Deep genomic-scale analyses of the metazoa reject Coelomata: evidence from single- and multigene families analyzed under a supertree and supermatrix paradigm. *Genome Biol Evol*. 2:310–324.
Horn CN. 1985. A systematic revision of the genus *Heteranthera* (*sensu lato*, Pontederiaceae). [Tuscaloosa (AL)]: University of Alabama.[PhD Thesis]
Kellogg EA, Bennetzen JL. 2004. The evolution of nuclear genome structure in seed plants. *Am J Bot*. 91:1709–1725.
Kohn JR, Graham SW, Morton BR, Doyle JJ, Barrett SCH. 1996. Reconstruction of the evolution of reproductive characters in Pontederiaceae using phylogenetic evidence from chloroplast DNA restriction-site variation. *Evolution* 50:1454–1469.
Maddison WP. 1997. Gene trees in species trees. *Syst Biol*. 46:523–536.
Maier SA, Podemski L, Graham SW, McDermid HE, Locke J. 2001. Characterization of the adenosine deaminase-related growth factor (ADGF) gene family in *Drosophila*. *Gene* 280:27–36.
Mathews S, Donoghue MJ. 1999. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science*. 286:947–950.
Mayrose I, Barker MS, Otto SP. 2010. Probabilistic models of chromosome number evolution and the inference of polyploidy. *Syst Biol*. 59:132–144.
Ness RW, Wright SI, Barrett SCH. 2010. Mating-system variation, demographic history and patterns of nucleotide diversity in the tristylous plant *Eichhornia paniculata*. *Genetics* 184:381–392.
Olmstead RG, Palmer JD. 1994. Chloroplast DNA systematics: a review of methods and data analysis. *Am J Bot*. 81:1205–1224.
Page RDM. 1998. GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics* 14:819–820.
Page RDM, Charleston MA. 1997. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol Phylogenet Evol*. 7:231–240.
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.
Saarela JM, Prentis PJ, Rai HS, Graham SW. 2008. Phylogenetic relationships in the monocot order Commelinales, with a focus on Philydraceae. *Botany* 86:719–731.
Sanderson MJ, McMahon MM. 2007. Inferring angiosperm phylogeny from EST data with widespread gene duplication. *BMC Evol Biol*. 7:1–14.
Slowinski JB, Knight A, Rooney AP. 1997. Inferring species trees from gene trees: a phylogenetic analysis of the Elapidae (Serpentes) based on the amino acid sequences of venom proteins. *Mol Phylogenet Evol*. 8:349–362.
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. *Mol Phylogenet Evol*. 48:1013–1026.
Swofford DL. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). v. 4. Sunderland (MA): Sinauer Associates.
Wehe A, Bansal MS, Burleigh JG, Eulenstein O. 2008. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics*. 24:1540–1541.
Zwickl DJ. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion [PhD Thesis]. [Austin (TX)]: The University of Texas at Austin.