JSE Journal of Systematics and Evolution

Journal of Systematics and Evolution 48 (4): 240–248 (2010) doi: 10.1111/j.1759-6831.2010.00080.x

Molecular evolution and phylogeny of the angiosperm *ycf2* gene

[1]Jin-Ling HUANG* [1]Gui-Ling SUN [2]Da-Ming ZHANG

[1](*Department of Biology, East Carolina University,* Greenville, NC 27858, USA)

[2](*State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences,* Beijing 100093, China)

Abstract Much of the recent progress in understanding angiosperm phylogeny has been achieved using multigene or plastid genome datasets. However, it is largely unclear what size of dataset is required to achieve sufficient resolution. The *ycf2* gene is the largest plastid gene in angiosperms and it was used as part of multigene datasets in several earlier investigations into angiosperm relationships. In this study, we show that the *ycf2* gene alone can provide a generally well-supported phylogeny that is consistent with those inferred from the most comprehensive multigene or plastid genome datasets. The phylogenetic signal of the *ycf2* gene is likely derived from the combination of its long sequence length and low rate of nucleotide substitution. The *ycf2* gene may provide a low-cost alternative to comprehensive multigene or genome datasets for investigating angiosperm relationships.

Key words angiosperm phylogeny, indel, multigene dataset, molecular evolution.

Significant progress has been made in elucidating the evolutionary history of angiosperms in the past 20 years. Molecular analyses generated from single-gene or multigene datasets have consistently identified several major clades, the relationships among which often are robustly supported (Chase et al., 1993; Qiu et al., 1999; Savolainen et al., 2000a, 2000b; Soltis et al., 2000; Jansen et al., 2007; Moore et al., 2007). Almost all available studies have been generated through DNA sequence data, often using multiple genes from different genomes (nuclear, plastid, and mitochondrial). Use of larger and larger multigene datasets has increasingly become commonplace as DNA sequencing prices have become more affordable. For example, a recent phylogenetic study of Saxifragales used a combined dataset of 16 genes from all three genomes (Jian et al., 2008), and an endeavor to tackle the relationships among basal angiosperms used a dataset of 61 plastid genes (Moore et al., 2007). The most ambitious study thus far used 81 plastid genes to investigate the relationships among major angiosperm groups (Jansen et al., 2007).

The *ycf2* gene is the largest plastid gene reported in angiosperms (Drescher et al., 2000). The function of *ycf2* is largely unknown, but appears not to be particularly related to photosynthesis (Drescher et al., 2000). In most land plants, two identical *ycf2* copies are located in the inverted repeat regions of plastid genomes, but independent losses of the *ycf2* gene occurred in multiple angiosperm groups including Poaceae. In combination with other genes, *ycf2* DNA sequences have been used in several earlier investigations on angiosperm phylogeny (Jansen et al., 2006, 2007; Moore et al., 2007; Jian et al., 2008). In this study, we investigate the sequence variation and phylogenetic utility of the *ycf2* gene in angiosperms. We show that the *ycf2* gene alone can provide a consistent and generally well-supported phylogeny with those inferred from the most comprehensive multigene data. We suggest that the *ycf2* gene may provide an alternative option to comprehensive multigene or genome approaches for studying the relationships among major angiosperm groups.

1 Material and methods

1.1 Comparison of sequence variation between *ycf2* and other plastid genes

To investigate the tempo and mode of *ycf2* sequence variation, we compared the sequence length, variability, and nucleotide composition between *ycf2* and several other plastid genes (*ndhB*, *ycf1*, *atpB*, *rbcL*, *ndhF*, *rpoC2*, and *matK*) that are often used in plant molecular systematics. For each of these genes, 36 sequences were sampled from major angiosperm groups. All sequences were obtained from the National Center for Biotechnology Information's non-redundant DNA sequence database (nr/nt) and their gene identifiers are shown in Table S1. Multiple and pairwise sequence alignments were carried out using MUSCLE (Edgar, 2004). Ambiguously aligned regions and gap sites were removed.

Received: 9 February 2010 Accepted: 1 April 2010

* Author for correspondence. E-mail: huangj@ecu.edu; Tel.: 1-252-328-5623; Fax: 1-252-328-4178.

Sequence divergence was estimated using uncorrected distance (p-distance). All statistical measurements were calculated using MEGA4.0 (Tamura et al., 2007).

1.2 Phylogenetic analyses

Multiple sequence alignments were visually inspected and manually refined. Ambiguously aligned regions and autapomorphic indels were removed from phylogenetic analyses. For both DNA and protein sequences, phylogenetic analyses were carried out with a maximum likelihood (ML) method using PHYML (Guindon & Gascuel, 2003) and a neighbor-joining distance method using the program *neighbor* of PHYLIP version 3.65 (Felsenstein, 2005). The nucleotide substitution model used was GTR + G + I as determined with ModelTest 3.7 (Posada & Crandall, 1998). The protein substitution matrix was determined using ProtTest (Abascal et al., 2005) and the JTT model was selected. Bootstrap support was estimated using 100 replicates for both ML and distance analyses. Distances for neighbor-joining bootstrap analyses were calculated using TREE-PUZZLE (Schmidt et al., 2002) and PUZZLEBOOT v1.03. Statistical tests for sequence compositional heterogeneity were carried out using TREE-PUZZLE with the default ML model. Branch lengths and topologies of the trees depicted in all figures were calculated with PHYML.

Ginkgo and *Cycas* YCF2 protein sequences were used as outgroups to root the angiosperm phylogeny. Because of the occasional uncertainty in alignment between gymnosperm and angiosperm sequences, we also carried out separate analyses without gymnosperm sequences. In these cases, the sequence of *Amborella*, which has been identified in most recent studies as the earliest branch in angiosperm evolution (Hilu et al., 2003; Soltis et al., 2003; Qiu et al., 2005, 2006; Jansen et al., 2007; Moore et al., 2007), was chosen as the outgroup.

2 Results and discussion

2.1 Comparison of sequence variability between *ycf2* and other plastid genes

The length of the complete *ycf2* gene ranges from 6816 to 6966 bp in most angiosperms. Table 1 shows the comparison of sequence divergence, variability, and other characteristics between *ycf2* and seven other plastid genes that are often used in plant systematics. Overall, the *ycf2* gene possesses one of the lowest sequence substitution rates among the eight sampled genes (Table 1). The mean p-distance of *ycf2* sequences is 4.8047%, which is approximately fourfold lower than that of the widely used *matK* gene (19.2913%). As a result of their low sequence substitution rate, *ycf2* sequences from major angiosperm groups can be easily aligned and they are less likely to suffer from parallel mutations. Nevertheless, because of its greater length, the *ycf2* gene has generated more parsimoniously informative characters in total than the fast-evolving *matK* (826 versus 521) for the same taxonomic sample, although the percentage of parsimoniously informative characters for *ycf2* is much lower (16.2% versus 54.8% for *matK*).

2.2 Indels and their systematic implications

The y*cf2* sequences contain many indels uniquely shared by some sampled sequences. Traditionally, shared indels are often used as rare genomic characters for phylogenetic reconstruction (Rokas & Holland, 2000), particularly for major eukaryotic groups (Baldauf & Palmer, 1993; Huang & Gogarten, 2006; Rice & Palmer, 2006). For example, the initial recovery of the sister relationship between fungi and animals was largely based on the combined evidence of phylogenetic analyses and the identification of a 12-aa insertion shared by the two groups in the otherwise very conserved region of elongation factor-1α (Baldauf

Table 1 Comparison of sequence length, variability, and nucleotide composition for *ycf2* and seven other plastid genes often used for plant molecular systematics

Gene	Length (bp)	Divergence		Transition/transversion		Variability		Nucleotide composition				
		%Divergence	SE	ti/tv	SE	%Variable	%Informative	%GC	%T	%C	%A	%G
ndhB	1475	2.2075	0.002	2.5088	0.5306	18.1	7.3	37.7	34.8	20	27.6	17.7
ycf2	5099	4.8047	0.001	1.4451	0.0081	39.8	16.2	37.5	31.2	19	31.3	18.5
atpB	1462	7.9934	0.004	2.7674	0.2598	35.5	25.5	42.8	27.7	19.3	29.5	23.5
rbcL	1340	8.0126	0.004	2.1141	0.2113	35.2	25.6	44.5	28.7	19.8	26.8	24.7
ndhF	1877	15.4569	0.004	1.2871	0.0571	60.3	42.3	33.0	39.0	15.7	27.9	17.3
rpoC2	2235	14.7955	0.004	1.5595	0.0594	65.6	46.3	36.6	30.0	17.1	33.5	19.5
ycf1	3191	19.0569	0.003	1.1533	0.0326	73.4	50.8	31.6	32.4	15.5	35.9	16.1
matK	951	19.2913	0.006	1.3535	0.0724	77.1	54.8	33.0	36.9	17.7	30.1	15.3

Except for *ycf1*, analyses were carried out with 36 sequences sampled from major angiosperm groups. Length refers to the aligned length of each dataset minus ambiguous and gap sites. Statistics for *ycf1* were calculated with four incomplete Saxifragales sequences removed from the data (i.e. only 32 sequences were used in the analyses). SE, standard error; ti/tv, transition/transversion.

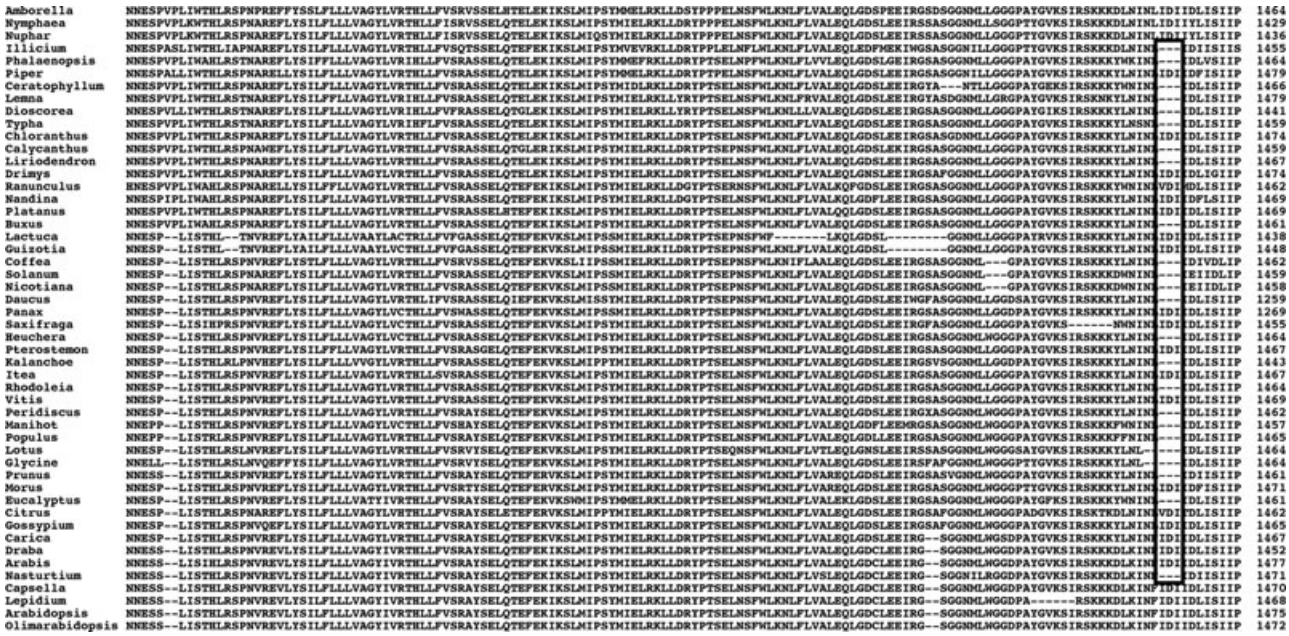

Fig. 1. Distribution of indels in YCF2 protein sequences. Note that a 2-aa deletion is present in all sequences of core eudicots, and that other deletions are often specific to monophyletic groups. The boxed indel is not specific to any monophyletic group.

& Palmer, 1993). A unique indel in 18S rDNA has also been cited as evidence to support the placement of Peridiscaceae in Saxifragales (Davis & Chase, 2004; Soltis et al., 2007) (but also see the following section for discussion).

The angiosperm *ycf2* sequences are generally well aligned and indel boundaries can be easily assessed. We carefully inspected the distribution of indels in the major groups identified in the *ycf2* phylogeny and in earlier multigene analyses (Soltis et al., 2000, 2003; Hilu et al., 2003; Jansen et al., 2007; Moore et al., 2007; Zhu et al., 2007). As expected, many of these indels indeed are predictive of phylogenetic relationships. This is most prominent in Brassicales, where multiple indels are uniquely shared by all sampled sequences from this group (Fig. 1). A highly conserved 5-aa (SDHLS) insertion is present in sequences of all Saxifragales except *Peridiscus* and *Paeonia*, coincident with the often uncertain positions of these two taxa in earlier analyses (Soltis et al., 2007; Jian et al., 2008). Additionally, a 2-aa deletion is identified in all sequences of core eudicots, and multiple other indels are found to be specific to individual minor clades (Fig. 1).

An interesting observation, however, is the occurrence of multiple highly conserved indels shared by more distantly related taxa (Fig. 1). For instance, a 3-aa indel is sporadically distributed in multiple angiosperm groups in a highly conserved sequence region (Fig. 1). A 2-aa insertion is also randomly distributed in Brassicaceae. Theoretically, random distribution of corresponding indels could have resulted from multiple evolutionary scenarios, such as random sequence sampling, differential gene losses, horizontal gene transfer, independent occurrence, or recombination (Keeling & Palmer, 2001). The scenarios resulting from random sequence sampling or differential losses are only possible if the two *ycf2* gene copies differ in possession of these indels. Such scenarios can be reliably excluded because the *ycf2* gene copies are identical in their DNA sequences in our analyses. Gene transfer is also unlikely to be responsible for the observed indel distribution pattern, because this scenario will lead to a gene phylogeny incongruent with the accepted organismal phylogeny, which is untrue in the case of *ycf2*. Therefore, although we cannot confidently pinpoint the cause for the random distribution of these unique indels, this observation points to the potential liability of the longstanding practice of using indels as shared derived characters in molecular systematics.

2.3 Comparison of *ycf2* phylogeny and multigene phylogenies

2.3.1 Overall topology Earlier analyses suggest that slowly evolving genes are most useful in resolving deep relationships of angiosperms (Zhu et al., 2007; Jian et al., 2008). The *ycf2* gene is the longest in plastid genomes of angiosperms and it also has one of the lowest substitution rates among plastid genes (Table 1)

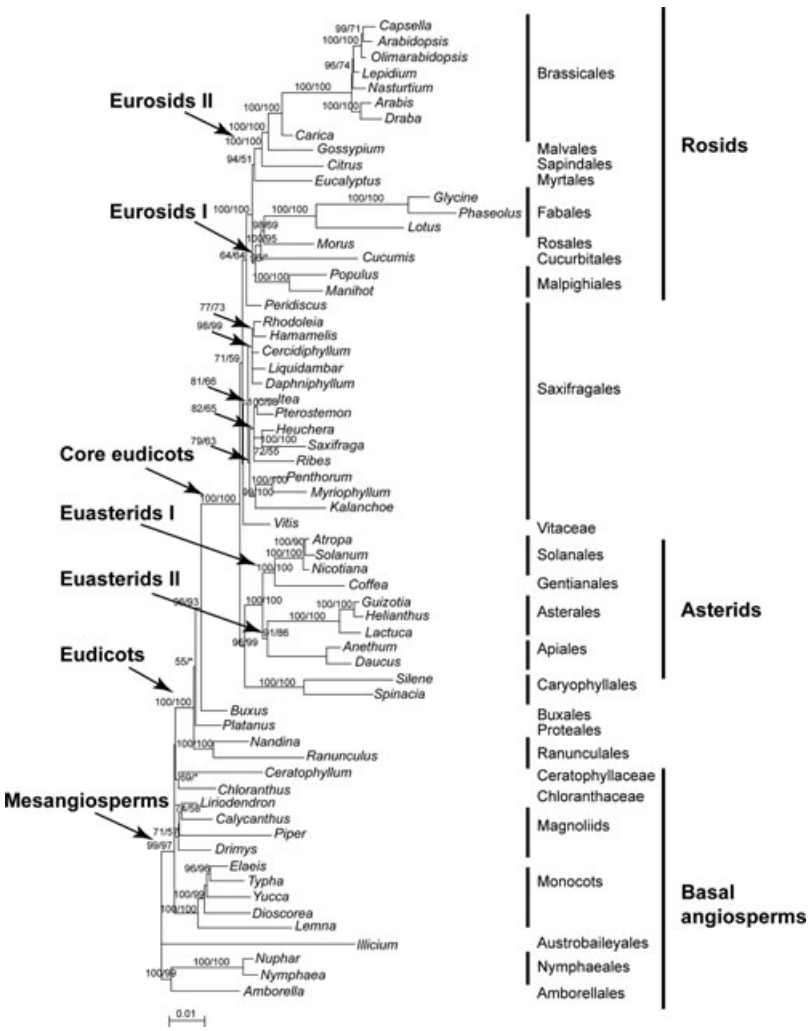

Fig. 2. Phylogenetic analyses of angiosperm *ycf2* DNA sequences. Numbers above the branch show bootstrap values for maximum likelihood and distance analyses, respectively. Branch length and tree topology were calculated from maximum likelihood analyses.

(Jian et al., 2008). Therefore, we decided to compare the phylogeny of *ycf2* with those generated from earlier multigene datasets. For the convenience of comparison, whenever possible, we adopted the same clade names as those used in the whole plastid genome phylogenies (Jansen et al., 2007; Moore et al., 2007). Formal names have been given for these clades under the recently proposed PhyloCode (Cantino et al., 2007).

Phylogeny based on analyses of 6637 bp of *ycf2* DNA sequences is shown in Fig. 2. The YCF2 protein sequence phylogeny has a similar topology and support values (Fig. S1) and, therefore, will not be discussed in detail. With *Ginkgo* and *Cycas* YCF2 protein sequences as outgroups, *Amborella* is identified as the sister group to the remaining angiosperms, although the support for such an early split is not significant (Fig. S2). The major groups identified in the *ycf2* phylogeny and their relationships are largely consistent with those inferred from 61 and 81 plastid genes (hereafter, CP61 and CP81) (Jansen et al., 2007; Moore et al., 2007) (Fig. 2). Nymphaeales and Illiciaceae are successive outgroups to the mesangiosperm clade (Cantino et al., 2007). Ceratophyllaceae and Chloranthaceae form a clade with a bootstrap value of 69% from ML analyses. The monophyly of Chloranthaceae and Ceratophyllaceae was recovered in some recent investigations based on molecular and morphological data (Duvall et al., 2006; Qiu et al., 2006; Endress & Doyle, 2009), but not in the CP61 study, where the two groups were sister to magnoliids and eudicots, respectively (Moore et al., 2007).

As expected, eudicots are highly supported as a monophyletic group in the *ycf2* phylogeny, with Ranunculales, Platanaceae, and Buxaceae being successfully closer outgroups to the core eudicots, which are also

Table 2 Comparisons of statistical support for major angiosperm groups identified in multigene DNA phylogenies and the *ycf2* phylogeny

Group	*matK*[1]	*rbcL*/*atpB*/18S/26S[2]	*matR*/*rbcL*/*atpB*/18S[3]	RMN	CP81[4]	*ycf2*
Eudicots	96	100	NA	100	100	100
Core eudicots	91	100	100	100	100	100
Asterids + Caryophyllales	74	<50	<50	99	100	96
Asterids	98	99	100	100	100	100
Euasterids I	71	100	NA	100	100	100
Euasterids II	91	99	NA	100	100	100
Saxifragales + Rosids + Vitaceae	<50	<50	<50	100	97	71
Rosids	98	79	66	100	100	100
Eurosids I	52	<50	85	100	100	96
Eurosids II	<50	88	99	100	100	100
Eurosids II + Myrtales	<50	<50	<50	96	100	94

Bootstrap values generated from maximum likelihood analyses are used for *ycf2* and *rpoC2*, *matK*, and *ndhF* (RMN) data. The highest bootstrap or jackknife values are chosen for other analyses. Because of the different taxonomic samples used in other analyses, values are chosen from groups with similar taxon composition. Note that these values may not be calculated using the same phylogenetic algorithms. [1]Hilu et al., 2003; [2]Soltis et al., 2003; [3]Zhu et al., 2007; [4]Jansen et al., 2007. NA, not applicable.

recovered as a clade. Two major clades are identified within core eudicots, one including Caryophyllales and asterids and the other consisting of Saxifragales, Vitaceae, and rosids (Fig. 2). Within rosids, two major clades (eurosids I and eurosids II + Myrtales) are recovered, and their relationships are in agreement with those inferred from CP61 and CP81 (Jansen et al., 2007; Moore et al., 2007). The relationships in asterids in the *ycf2* phylogeny are also consistent with those inferred from CP61 and CP81 (Jansen et al., 2007; Moore et al., 2007). Both asterids I and asterids II are recovered as monophyletic groups. Caryophyllales are found to be the sister group of asterids (Fig. 2).

Not only is the *ycf2* phylogeny largely consistent with those inferred from CP61 and CP81, it also provides robust support for most identified groups (Fig. 2). Except for relationships among basal angiosperm groups (magnoliids, monocots, Ceratophyllaceae and Chloranthaceae) and Saxifragales, the vast majority of other identified clades in the *ycf2* phylogeny are supported by bootstrap values of over 85%.

2.3.2 Phylogenetic signal, sequence length, and rate of nucleotide substitution Compared with the widely used plastid genes such as *rbcL* (Savolainen et al., 2000b) and *matK* (Hilu et al., 2003), the *ycf2* sequence data provide better support in resolving deep relationships of angiosperms (Table 2). Because *ycf2* is the largest gene in plastid genomes, we decided to investigate whether the phylogenetic signal of *ycf2* is derived from its greater length. Independent phylogenetic analyses were carried out on a separate dataset that includes an identical taxonomic sample and an approximately equal sequence length (6640 bp versus 6637 bp for *ycf2*) from three fast-evolving plastid genes (*rpoC2*, *matK*, and *ndhF*) (RMN). Both *ycf2* and RMN phylogenies provide generally strong and consistent relationships among major groups of eudicots; such strong support is somewhat expected because these major groups can often be recovered by other single-gene data (Table 2). The biggest differences between *ycf2* and RMN phylogenies are in their relationships within Saxifragales (see below) and among basal angiosperms. RMN places Peridiscaceae within Saxifragales with strong support, but provides poor support for the overall relationships among major clades of Saxifragales and within wood Saxifragales (Fig. S3); additionally, it also fails to recover magnoliids as a monophyletic group. The *ycf2* gene, on the other hand, provides decent support and resolution for Saxifragales, but places Peridiscaceae as the sister group to rosids (see next section for further discussion). This suggests that, although larger datasets can almost always generate a stronger phylogenetic signal, the slower rate of nucleotide substitution of the *ycf2* gene may be responsible for the additional resolution for some ancient angiosperm groups that were subject to earlier rapid radiation.

The *ycf2* data also generate a better resolution than several datasets combining sequences from different genome compartments, including four genes (*rbcL*, *atpB*, 18S rDNA, and 26S rDNA) from plastid and nuclear genomes (Soltis et al., 2003) and four genes from all three genomes (mitochondrial *matR*, nuclear 18S rDNA, and plastid *rbcL* and *atpB*) (Zhu et al., 2007) (Table 2). For instance, the *ycf2* phylogeny provides strong support (96% and 99% from ML and distance analyses, respectively) for the sister-group relationship between Caryophyllales and asterids, which was either unrecovered or only weakly supported by multigene data from different genomes (Soltis et al., 2000, 2003; Zhu et al., 2007) (Table 2). Currently, we are unsure whether this difference in resolution may have resulted from factors such as sampling density, gene selection, or the method of phylogenetic analysis.

Table 3 Comparisons of bootstrap support for relationships within Saxifragales inferred from multigene DNA sequence data and *ycf2* sequences

Group	Fast (3)	Intermediate (6)	Slow (8)	Total (16+IR)	*ycf2*
Saxifragales minus Peridiscaceae	NP	NP	98	99	81
Woody clade	91	NP	82	98	98
Core Saxifragales	54	83	95	100	79
Saxifragaceae alliance	100	92	100	100	82
Saxifragaceae + Grossulariaceae	100	93	53	100	72
Iteaceae + Pterostemonaceae	100	100	<50	199	100
Halograceae + Crassulaceae	100	98	100	100	99

Bootstrap support values in multigene analyses are from Jian et al. (2008). Numbers in parentheses indicate the number of genes used in the multigene analyses. All bootstrap values except those for Iteaceae/Pterostemonaceae, where data from maximum parsimony are shown, are from maximum likelihood analyses, but the phylogenetic algorithms used in Jian et al. (2008) and the current study may differ. Fast, intermediate, and slow genes are plastid genes, whereas total evidence includes four mitochondrial genes, two nuclear genes, 10 plastid genes and the entire plastid inverted repeat region. Note that the *ycf2* phylogeny is consistent with those inferred from slow genes and total evidence. Note also that *ycf2* is included in the slow genes and total evidence datasets used in the original multigene DNA phylogenetic analyses. IR, inverted repeat; NP, not present in topology.

2.3.3 Systematic position of Saxifragales and relationships within the group Saxifragales are a major angiosperm group including species of diverse morphology and were likely subject to a rapid radiation during their early evolution (Soltis & Soltis, 1997; Jian et al., 2008). Although the monophyly of Saxifragales has been supported by several analyses, the systematic position of Saxifragales varies in these analyses (Soltis et al., 2000, 2003; Hilu et al., 2003;). Except for a combined three-gene dataset (18S rDNA, *rbcL*, and *atpB*) (Soltis et al., 2000), no other earlier studies were able to identify the sister group of Saxifragales with bootstrap or jackknife support of over 50%. The phylogeny generated from the three-gene dataset placed Saxifragales within core eudicots as the sister group of rosids and Vitaceae with weak bootstrap support of 60% (Soltis et al., 2000). However, an enlarged dataset with the addition of 26S rDNA sequences failed to provide a consistent topology, and placed Saxifragales with Caryophyllales (Soltis et al., 2003). No Saxifragales were sampled in recent CP61 and CP81 analyses (Jansen et al., 2007; Moore et al., 2007). Therefore, despite the significant progress in angiosperm phylogeny overall, the systematic position of Saxifragales remains largely uncertain (Soltis & Soltis, 2004).

The *ycf2* phylogeny indicates that Saxifragales plus Vitaceae form a monophyletic group with rosids, with modest bootstrap values of 71% and 59% from ML and distance analyses, respectively (Fig. 2). The monophyly of these three taxa is also strongly supported by ML analyses of RMN, with a bootstrap support value of 97% (Fig. S3). These are the most significant support values obtained thus far regarding the systematic position of Saxifragales. Given the overall topological agreement of *ycf2*, CP61, and CP81 phylogenies and the fact that this affiliation was also recovered as the best resolution from earlier multigene datasets (Soltis et al., 2000), it suggests that the monophyly of Saxifragales, Vitaceae, and rosids may likely be real.

The *ycf2* DNA sequences were previously used as part of a multigene dataset to investigate the relationships within Saxifragales (Jian et al., 2008). We, therefore, decided to compare the *ycf2* phylogeny with those generated from the above multigene data. Except *Paeonia brownii,* whose *ycf2* sequence has an apparently elevated substitution rate based on alignment inspection and a heterogeneity test, 16 species of Saxifragales sampled in the earlier study (Jian et al., 2008) are included in our current analyses (Fig. 2). The topology of the *ycf2* phylogeny is consistent with those inferred from slowly evolving genes and total evidence data (Table 3), which are considered to be more reliable (Jian et al., 2008). In each of the phylogenies generated from these three datasets, all Saxifragales except Peridiscaceae form two major clades, one including woody Saxifragales (Altingiaceae, Hamamelidaceae, Daphniphyllaceae, and Cercidiphyllaceae) and the other the core Saxifragales (Saxifragaceae, Grossulariaceae, Iteaceae, Pterostemonaceae, Crassulaceae, and Haloragaceae). The same topology with similar bootstrap support values was also recovered when the *ycf2* dataset was truncated to the size of fast-evolving genes (5847 bp) (Fig. S4). However, phylogenetic analyses using genes of intermediate and fast substitution rates provided inconsistent topologies, particularly regarding the position of Peridiscaceae and the monophyly of woody Saxifragales (Jian et al., 2008) (Table 3). These results suggest that the resolution discrepancy generated from different datasets are largely due to the rates of nucleotide substitution and that slow genes are more suitable for resolving relationships of ancient lineages that were subject to earlier rapid radiation (Jian et al., 2008).

The most distinct difference between phylogenies generated from *ycf2*, RMN, and earlier multigene datasets is the systematic position of Peridiscaceae (Soltis et al., 2007; Jian et al., 2008). Depending on the genes and methods used, Peridiscaceae were placed in

various positions in earlier analyses. Maximum parsimony analyses using total evidence data placed Peridiscaceae and Paeoniaceae as a clade that was sister to the remaining Saxifragales, whereas those analyses using plastid genes of intermediate and fast substitution rates placed Peridiscaceae either as sister to Saxifragaceae plus allied groups or with Hamamelidaceae (Jian et al., 2008). Similar observations were also made in analyses of a combined five-gene dataset (*rbcL*, *atpB*, *matK*, 18S rDNA, and 26S rDNA) (Soltis et al., 2007). None of these affiliations received bootstrap support of over 84%. The most significant and consistent support came from ML and Bayesian analyses using total evidence data and slowly evolving genes, where Peridiscaceae were placed as sister to the remaining Saxifragales (Jian et al., 2008).

The earlier multigene analyses largely hinged on an assumption that Peridiscaceae and other Saxifragales were a monophyletic group (Soltis et al., 2007; Jian et al., 2008). Although monophyly of this group was indeed recovered from an analysis using combined DNA sequences of *ndhF*, *rbcL*, and PHYC (Davis & Chase, 2004), only a few sequences from other angiosperm groups were selected in that study. A limited sequence sampling from other angiosperm groups may potentially fail to recover the true affiliation of Peridiscaceae even if other Saxifragales are sufficiently sampled. In this study with an enlarged dataset, the RMN phylogeny strongly supports a monophyly including Peridiscaceae and other Saxifragales (Fig. S3), but the *ycf2* phylogeny places Peridiscaceae as sister to rosids; ML and distance analyses provide the same bootstrap value of 64% for *ycf2* DNA sequences (Fig. 2), and 87% and 80%, respectively, for YCF2 protein sequences (Fig. S1). Therefore, the systematic position of Peridiscaceae appears to be uncertain, and further detailed studies using other genes and a higher sampling density are necessary.

2.4 Utility of the *ycf2* gene in angiosperm phylogenetics

With the availability of the whole plastid phylogeny generated by recent studies (Jansen et al., 2007; Moore et al., 2007), the overall framework of angiosperm phylogeny appears to be well established. Such a framework provides a yardstick to evaluate the utility of other molecular markers in plant systematics. It is not our intention in this study to provide additional insights into the relationships among major angiosperm groups. Rather, we show that the *ycf2* gene alone can provide a generally well-supported angiosperm phylogeny that is consistent with those inferred from the whole plastid genome data (i.e. CP61 and CP81). Therefore, sufficient phylogenetic resolution may be achieved from single genes of suitable length without resorting to comprehensive genome data, in particular for relationships among groups of eudicots.

Compared with the fast-evolving plastid RMN dataset, the *ycf2* gene appears to have provided better overall resolution for more ancient clades that were under rapid radiation. Therefore, the phylogenetic signal of *ycf2* is most likely derived from a combination of its greater sequence length and slower rate of nucleotide substitution. We reason that the *ycf2* gene is valuable for future investigations to understand the finer details of the angiosperm tree of life. This should not diminish the value of comprehensive multigene or genome data, but provides an alternative low-cost option for investigating angiosperm relationships. Even with increasingly reduced expenses of sequencing, it is impractical and unnecessary to sequence whole plastid (or mitochondrial) genomes for many studies of plant molecular systematics. It should be noted that no current phylogenomic analysis has used complete genome sequences. Both CP61 and CP81 are essentially the most comprehensive multigene datasets constructed from plastid genomes. The construction of such datasets, including genome assembly, annotation, and further data processing, requires considerable efforts and expertise that may not be available to many plant systematists. Furthermore, although many multigene datasets have often provided desirable results in resolving relationships among organisms, it has also been shown that multigene datasets, in particular those combining sequences from different genomes, may potentially generate a well supported phylogeny that reflects the evolutionary history of neither the organisms nor the individual genes (Zhaxybayeva et al., 2006). In view of this potential complication, reliable single-gene markers such as *ycf2* have their own outstanding merits in inferring organismal evolutionary history.

It should also be noted that, although *ycf2* sequences provide a consistent and generally well-supported topology with CP61 and CP81, all three datasets are from plastid genomes. Strongly resolved angiosperm phylogeny based on independent nuclear and mitochondrial evidence is still lacking. It is our opinion that exploring other nuclear and mitochondrial genes for their utility in reconstructing angiosperm phylogeny remains a task for the near future.

Acknowledgements We are grateful to Jim DOYLE and two anonymous reviewers for their insightful comments and suggestions. We also thank Doug SOLTIS, Peter GOGARTEN, Zhiduan CHEN, and Jason BOND

for critical reading of an early manuscript draft and helpful discussion. This study is supported in part by a National Science Foundation (USA) Assembling the Tree of Life (ATOL) grant (DEB-0830024).

References

Abascal F, Zardoya R, Posada D. 2005. ProtTest: Selection of best-fit models of protein evolution. Bioinformatics 21: 2104–2105.

Baldauf SL, Palmer JD. 1993. Animals and fungi are each other's closest relatives: Congruent evidence from multiple proteins. Proceedings of the National Academy of Sciences USA 90: 11558–11562.

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.

Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR, Price RA, Hills HG, Qiu YL, Kron KA, Rettig JH, Conti E, Palmer JD, Manhart JR, Sytsma KJ, Michaels HJ, Kress WJ, Karol KG, Clark WD, Hedren M, Gaut BS, Jansen RK, Kim K-J, Wimpee CF, Smith JF, Furnier GR, Strauss SH, Xiang Q-Y, Plunkett GM, Soltis PS, Swensen SM, Williams SE, Gadek PA, Quinn CJ, Eguiarte LE, Golenberg E, Learn GH, Graham SW, Barrett SCH, Dayanandan S, Albert VA. 1993. Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene rbcL. Annals of the Missouri Botanical Garden 80: 528–580.

Davis CC, Chase MW. 2004. Elatinaceae are sister to Malpighiaceae; Peridiscaceae belong to Saxifragales. American Journal of Botany 91: 262–273.

Drescher A, Ruf S, Calsa T Jr, Carrer H, Bock R. 2000. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. The Plant Journal 22: 97–104.

Duvall MR, Mathews S, Mohammad N, Russell T. 2006. Placing the monocots: Conflicting signal from trigenomic analyses. Aliso 22: 79–90.

Edgar RC. 2004. Muscle: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792–1797.

Endress PK, Doyle JA. 2009. Reconstructing the ancestral angiosperm flower and its initial specializations. American Journal of Botany 96: 22–66.

Felsenstein J. 2005. Phylip (phylogeny inference package) version 3.65. Seattle: Department of Genome Sciences, University of Washington. Available from: http://evolution.genetics.washington.edu/phylip.html [Accessed 20 June 2010].

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52: 696–704.

Hilu KW, Borsch T, Muller K, Soltis DE, Soltis PS, Savolainen V, Chase MW, Powell MP, Alice LA, Evans R, Sauquet H, Neinhuis C, Slotta TAB, Rohwer JG, Campbell CS, Chatrou LW. 2003. Angiosperm phylogeny based on matk sequence information. American Journal of Botany 90: 1758–1776.

Huang J, Gogarten JP. 2006. Ancient horizontal gene transfer can benefit phylogenetic reconstruction. Trends in Genetics 22: 361–366.

Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack J, Muller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, Chumley TW, Lee SB, Peery R, McNeal JR, Kuehl JV, Boore JL. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proceedings of the National Academy of Sciences USA 104: 19369–19374.

Jansen RK, Kaittanis C, Saski C, Lee SB, Tomkins J, Alverson AJ, Daniell H. 2006. Phylogenetic analyses of Vitis (Vitaceae) based on complete chloroplast genome sequences: Effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evolutionary Biology 6: 32.

Jian S, Soltis PS, Gitzendanner MA, Moore MJ, Li R, Hendry TA, Qiu YL, Dhingra A, Bell CD, Soltis DE. 2008. Resolving an ancient, rapid radiation in Saxifragales. Systematic Biology 57: 38–57.

Keeling PJ, Palmer JD. 2001. Lateral transfer at the gene and subgenic levels in the evolution of eukaryotic enolase. Proceedings of the National Academy Sciences USA 98: 10745–10750.

Moore MJ, Bell CD, Soltis PS, Soltis DE. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proceedings of the National Academy of Sciences USA 104: 19363–19368.

Posada D, Crandall KA. 1998. ModelTest: Testing the model of DNA substitution. Bioinformatics 14: 817–818.

Qiu Y-L, Dombrovska O, Lee J, Li L, Whitlock BA, Bernasconi-Quadroni F, Rest JS, Davis CC, Borsch T, Hilu KW, Renner SS, Soltis DE, Soltis PS, Zanis MJ, Cannone JJ, Gutell RR, Powell M, Savolainen V, Chatrou LW, Chase MW. 2005. Phylogenetic analyses of basal angiosperms based on nine plastid, mitochondrial, and nuclear genes. International Journal of Plant Sciences 166: 815–842.

Qiu Y-L, Li L, Hendry TA, Li R, Taylor DW, Issa MJ, Ronen AJ, Vekaria ML, White AM. 2006. Reconstructing the basal angiosperm phylogeny: Evaluating information content of mitochondrial genes. Taxon 55: 837–856.

Qiu Y-L, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainen V, Chase MW. 1999. The earliest angiosperms: Evidence from mitochondrial, plastid and nuclear genomes. Nature 402: 404–407.

Rice DW, Palmer JD. 2006. An exceptional horizontal gene transfer in plastids: Gene replacement by a distant bacterial paralog and evidence that haptophyte and cryptophyte plastids are sisters. BMC Biology 4: 31.

Rokas A, Holland PW. 2000. Rare genomic changes as a tool for phylogenetics. Trends in Ecology and Evolution 15: 454–459.

Savolainen V, Chase MW, Hoot SB, Morton CM, Soltis DE, Bayer C, Fay MF, De Bruijn AY, Sullivan S, Qiu Y-L. 2000a. Phylogenetics of flowering plants based on combined

analysis of plastid *atpB* and *rbcL* gene sequences. Systematic Biology: 306–362.

Savolainen V, Fay MF, Albach DC, Backlund A, van der Bank M, Cameron KM, Johnson SA, Lledo MD, Pintaud J-C, Powell M, Sheahan MC, Soltis DE, Soltis PS, Weston P, Whitten WM, Wurdack KJ, Chase MW. 2000b. Phylogeny of the eudicots: A nearly complete familial analysis based on *rbcL* gene sequences. Kew Bulletin 55: 257–309.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502–504.

Soltis DE, Clayton JW, Davis CC, Gitzendanner MA, Cheek M, Savolainen V, Amorim AM, Soltis PS. 2007. Monophyly and relationships of the enigmatic family Peridiscaceae. Taxon 56: 65–73.

Soltis DE, Senters AE, Zanis MJ, Kim S, Thompson JD, Soltis PS, De Craene LPR, Endress PK, Farris JS. 2003. Gunnerales are sister to other core eudicots: Implications for the evolution of pentamery. American Journal of Botany 90: 461–470.

Soltis DE, Soltis PS. 1997. Phylogenetic relationships in saxifragaceae sensu lato: A comparison of topologies based on 18s rdna and rbcl sequences. American Journal of Botany 84: 504–522.

Soltis DE, Soltis PS, Chase MW, Mort ME, Albach DC, Zanis M, Savolainen V, Hahn WH, Hoot SB, Fay MF, Axtell M, Swensen SM, Prince LM, Kress WJ, Nixon KC, Farris JS. 2000. Angiosperm phylogeny inferred from 18s rdna, rbcl, and atpb sequences. Botanical Journal of the Linnean Society 133: 381–461.

Soltis PS, Soltis DE. 2004. The origin and diversification of angiosperms. American Journal of Botany 91: 1614–1626.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Molecular Biology Evolution 24: 1596–1599.

Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT. 2006. Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. Genome Research 16: 1099–1108.

Zhu XY, Chase MW, Qiu YL, Kong HZ, Dilcher DL, Li JH, Chen ZD. 2007. Mitochondrial *matR* sequences help to resolve deep phylogenetic relationships in rosids. BMC Evolutionary Biology 7: 217.

Supporting Information

The following supporting information is available for this article:

Fig. S1. Phylogenetic analyses of angiosperm YCF2 protein sequences. Numbers above the branch show bootstrap values for maximum likelihood and distance analyses, respectively. Asterisks indicate values lower than 50%.

Fig. S2. Phylogenetic analyses of angiosperm YCF2 protein sequences. Numbers above the branch show bootstrap values for maximum likelihood analyses. The tree is rooted with gymnosperm sequences.

Fig. S3. Phylogenetic analyses of a combined dataset of *rpoC2*, *matK*, and *ndhF* of 6640 bp. Numbers above the branch show bootstrap values for maximum likelihood and distance analyses, respectively. Asterisks indicate values lower than 50%.

Fig. S4. Phylogenetic analyses of a truncated dataset of *ycf2* DNA sequences for Saxifragales. The total sequence length is 5858 bp, which is the same size of the fast-gene dataset used in Jian et al (2008). The tree shows the bootstrap consensus tree from maximum likelihood analyses. Numbers above the branch show bootstrap values for maximum likelihood analyses.

Table S1. Taxonomic sampling and sequence identifiers used for comparative analyses of sequence divergence, variability, and other characteristics. The identifier for the whole genome is used for gene sequences that are retrieved from the same genome.

Table S2. GenBank identifiers for *ycf2* DNA sequences used in phylogenetic analyses.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.