Third-codon transversion rate-based *Nymphaea* basal angiosperm phylogeny -- concordance with developmental evidence

Xiaohan Yang*†, Gerald A. Tuskan*†, Timothy J. Tschaplinski†, (Max) Zong-Ming Cheng*
*Department of Plant Sciences, University of Tennessee, Knoxville, TN 37996
†Environmental Science Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37831

Corresponding author: (Max) Zong-Ming Cheng**;** E-mail: zcheng@utk.edu

Abstract

Flowering plants (angiosperms) appeared on Earth rather suddenly approximately 130 million years ago and underwent a massive expansion in the subsequent 10-12 million years. Current molecular phylogenies have predominantly identified *Amborella*, followed by *Nymphaea* (water lilies) or *Amborella* plus *Nymphaea*, in the ANITA clade (*Amborella*, Nymphaeales, Illiciaceae, Trimeniaceae and Austrobaileyaceae) as the earliest angiosperm. However, developmental studies suggest that the earliest angiosperm had a 4-cell/4-nucleus female gametophyte and a diploid endosperm represented by *Nymphaea*, suggesting that *Amborella,* having an 8-cell/9-nucleus female gametophyte and a triploid endosperm, cannot be representative of the basal angiosperm. This evolution-development discordance is possibly caused by erroneous inference based on phylogenetic signals with low neutrality and/or high saturation. Here we show that the 3rd codon transversion (P_3Tv), with high neutrality and low saturation, is a robust high-resolution phylogenetic signal for such divergences and that the P_3Tv-based land plant phylogeny cautiously identifies *Nymphaea*, followed by *Amborella*, as the most basal among the angiosperm species examined in this study. This P_3Tv-based phylogeny contributes insights to the origin of angiosperms with concordance to fossil and stomata development evidence.

Keywords: *Amborella*, *Nymphaea*, angiosperm, phylogenetic signal, chloroplast genome

Introduction

Flowering plants (angiosperms) appeared on Earth approximately 130 million years (mya) ago and subsequently experienced a massive expansion over the next 10-12 mya, resulting in much of the today's flora (De Bodt, Maere, and Van de Peer 2005). This phenomenon was described by Darwin in 1879 as an "abominable mystery" (Darwin 1903). Plant evolutionary biologists have long attempted to reconstruct angiosperm evolutionary history in an effort to determine the root of angiosperms. Since 1999, a series of molecular phylogenetic analyses have identified the monotypic *Amborella*, followed by Nymphaeales, or *Amborella* plus Nymphaeales, in the ANITA clade (*Amborella*, Nymphaeales, Illiciaceae, Trimeniaceae and Austrobaileyaceae) as representative of the most basal taxa (Mathews and Donoghue 1999; Parkinson, Adams, and Palmer 1999; Qiu et al. 1999; Soltis, Soltis, and Chase 1999; Graham and Olmstead 2000; Qiu et al. 2000; Zanis et al. 2002; Borsch et al. 2003; Zanis et al. 2003; Stefanovic, Rice, and Palmer 2004; Leebens-Mack et al. 2005; Chang et al. 2006; Qiu et al. 2006). In some cases, monocot-basal trees have also been reported (Goremykin et al. 2003; Goremykin et al. 2004; Chang et al.

2006; Goremykin and Hellwig 2006). Concurrently, developmental biologists have studied two key developmental aspects in the ANITA clade, *i.e.*, the structure of female gametophyte and formation of embryo nourishing tissue (endosperm), to help clarify the origin of the extant angiosperms (Williams and Friedman 2002; Friedman and Williams 2003; Friedman and Williams 2004; Friedman 2006). All ANITA members except *Amborella* have a 4-cell/4-nucleus female gametophyte with one of the double fertilization events yielding a biparental diploid endosperm (Williams and Friedman 2002; Friedman and Williams 2003; Friedman and Williams 2004), whereas the rest of angiosperm species have an 8-cell/9-nucleus (*Amborella*) (Friedman 2006) or a 7-cell/8-nucleus (all non-ANITA angiosperms) female gametophyte, with one of the double fertilization events yielding a biparental triploid endosperm (Friedman and Williams 2003; 2004).

Based on these findings, it has been hypothesized that the angiosperm female gametophytes consist of modular developmental subunits (Friedman and Williams 2003; Friedman 2006). From the ancestral gymnosperms, containing multiple egg cell megagametophytes yielding haploid embryo nourishing tissues upon a single fertilization event, the earliest angiosperm (all ANITA except *Amborella*) evolved into a 4-cell/4-nucleus female gametophyte (a single developmental module) which yielded a diploid endosperm upon double fertilization. In the early angiosperm history, this basic module was duplicated and resulted in 8-cell/9-nucleus female gametophyte in *Amborella* or 7-cell/8-nucleus one in the non-ANITA angiosperms, both of which yielded a triploid endosperm upon double fertilization (Friedman and Williams 2003; Friedman and Williams 2004; Friedman 2006). This hypothesis thus places non-*Amborella* ANITA members (represented by *Nymphaea* in most studies) as representative of the most basal angiosperms, thus conflicting with the current *Amborella*-basal (or *Amborella* plus Nymphaeales co-basal) molecular phylogeny.

If we accept the embryological evidence as consistent and incontrovertible, it is likely that the evolution-development conflict between the current phylogeny and developmental evidence lies on the phylogenetic side. In reconstructing the phylogeny with molecular characters (protein and nucleotide sequences), two types of errors can cause phylogenetic discordance: 1) stochastic error resulting from sampling of limited and/or underrepresented characters and taxa and 2) systematic error generated by the presence of a nonphylogenetic signal in the data (Phillips, Delsuc, and Penny 2004; Jeffroy et al. 2006). In the case of angiosperm phylogeny, because multi-gene concatenated samples and whole-chloroplast genomes of well-represented taxa have been analyzed (Stefanovic, Rice, and Palmer 2004; Leebens-Mack et al. 2005; Chang et al. 2006; Qiu et al. 2006), the stochastic errors should be minimal. Therefore, the discordance among the trees appears to be mainly caused by systematic errors as different reconstruction methods consistently generate discordant, yet highly supported trees (Stefanovic, Rice, and Palmer 2004; Chang et al. 2006; Jansen et al. 2006; Jeffroy et al. 2006; Qiu et al. 2006). Moreover, the inability to separate *Amborella* and *Nymphaea* in *Amborella* plus *Nymphaea* co-basal phylogeny suggests the current molecular phylogenetic signals used in the inference are not robust enough to reconcile the evolutionary history of angiosperms (Williams and Friedman 2002; Friedman and Williams 2003; Friedman and Williams 2004; Friedman 2006).

We report here the use of the 3rd codon transversion (P_3Tv) for inferring phylogenetic divergence. Our objectives were to 1) evaluate P_3Tv as a phylogenetic signal and 2) attempt to resolve the conflict between the molecular phylogenies and developmental evidences. Based on P_3Tv of the whole chloroplast genome coding sequences of 36 land plant taxa, we generated congruent maximum parsimony (MP) and maximum likelihood (ML) phylogenies that suggest that *Nymphaea* corresponds to the sister phyla to all other angiosperms including *Amborella*.

Materials and Methods

Analysis of neutrality and saturation of codon partitions

Synonymous and non-synonymous substitution possibilities were calculated based on the standard genetic code (Griffiths 2007). Uncorrected pairwise sequence divergence (p-distance) and Kimura-corrected distance (Kimura 1980) were calculated using MBEToolbox (Cai et al. 2005), for transitions (Ts), transversions (Tv), or Ts plus Tv at the 3rd-codon positions of the whole chloroplast genome coding sequences of 36 land plant taxa compiled by Qiu *et al*. (Qiu et al. 2006). The use of the 36 land plant taxa compiled by Qiu *et al*. (Qiu et al. 2006) represents a broad set of land plants comprising the full range of ancestral taxa. It is limited by the availability of sequenced chloroplast genomes and does not include less derived members of certain taxa, *e.g.*, Schisandraceae, *Austrobaileya*, and Trimeniaceae in the ANITA grade.

Phylogenetic analysis

The whole chloroplast genome coding sequence alignment matrix of 36 land plant taxa were from Qiu *et al*. (2006); the gaps in the alignment were removed. ML analyses were performed in TreeFinder (Jobb, von Haeseler, and Strimmer 2004), with an optimal model of nucleotide evolution selected by using the Akaike Information Criterion as implemented in ModelGenerator (Keane et al. 2006). MP analyses were carried out by using POWER (Lin et al. 2005). The RY coding (i.e., A or G was changed to R; C or T to Y) was used for phylogeny construction based on transversion only (Phillips and Penny 2003). For both ML and MP, Bootstrap analyses of 100 replicates were performed and the consensus trees were then displayed with bootstrap values (>50%).

Results and Discussion

P_3Tv as a phylogenetic signal

The ideal signal for phylogenetic reconstruction should fulfill two important criteria: 1) total or high neutrality and 2) no or low saturation. Compared with nucleotide sequences, protein sequences are less desirable for phylogenetic inference due to their lack of neutrality arising from wobble in the genetic code (20 amino acids coded by 61 corresponding codons). However, in using nucleotide sequence for phylogenetic reconstruction it has long been recognized that phylogenetic signals among codon positions vary greatly (Chaw et al. 2000; Magallon and Sanderson 2002; Qiu et al. 2006; Simmons et al. 2006). Therefore, we compared saturation and neutrality of different codon partitions. Because mutation rates in the 3rd codon position lead to high saturation which can cause long branch attraction (Chaw et al. 2000; Magallon and Sanderson 2002; Van de Peer et al. 2002; Jeffroy et al. 2006), many researchers (Chaw et al. 2000; Nickrent et al. 2000; Stefanovic, Rice, and Palmer 2004) have recommended excluding the 3rd codon position from phylogenetic reconstructions. Conversely, the 1st and 2nd codon positions have less than 5% synonymous mutations (Fig. 1A) and therefore suffer from low neutrality. Mutations in the 3rd codon position are 70% synonymous and are therefore neutral. In this regard, the 3rd codon position may have a superior phylogenetic signal due to its high neutrality. The constrained use of the 3rd codon (*i.e.*, transversions only) may provide both low saturation and high neutrality.

To insure that the 3rd codon transversion (*i.e.*, P_3Tv) functions equally well at shallow and deep phylogenies, we plotted uncorrected pairwise sequence divergence (p-distance) against Kimura-corrected distances for transitions (Ts), transversions (Tv), or Ts plus Tv at the 3rd-codon

positions of the whole chloroplast genome coding sequence alignment matrix of 36 land plant taxa compiled by Qiu *et al*. (2006) and found that the high saturation problem is substantially alleviated after the transitions are removed and only transversions are considered (Fig. 1B). The near linear relationship between uncorrected pairwise sequence divergence (p-distance) and Kimura-corrected distances indicates that high saturation in the 3rd codon position is mainly generated by transitions (A to G or C to T) and suggests that the 3rd codon transversion may be a more powerful phylogenetic signal than the currently-used transitions plus transversions of 1st+2nd, 3rd, or 1st+2nd+3rd positions for constructing shallow and deep phylogenies (Chaw et al. 2000; Nickrent et al. 2000; Stefanovic, Rice, and Palmer 2004; Qiu et al. 2006).

P_3Tv-based phylogeny places *Nymphaea* at the root of angiosperm

To further examine the effects of low saturation and high neutrality on phylogen y, we performed MP and ML phylogenetic analyses based on Tv versus Tv+Ts of different codon partitions (*e.g.*, 3rd, 1st+2nd or 1st+2nd+3rd positions) with the same whole-chloroplast genome coding sequence alignment matrix. Since whole-chloroplast genomes and well-represented taxa were used, the stochastic sampling errors in both taxa and character states should be minimal (Phillips, Delsuc, and Penny 2004; Jeffroy et al. 2006). Due to limited availability of sequenced chloroplast genomes, we recognize that the 36 land plant taxa compiled by Qiu *et al*. (2006) do not include some less derived members of certain taxa, particularly other members in the ANITA grade, *e.g.*, Schisandraceae, *Austrobaileya*, Trimeniaceae. However, since all members in the ANITA clade except *Amborella* show similar female reproductive structures (Williams and Friedman 2002; Friedman and Williams 2003; Friedman and Williams 2004; Friedman 2006), representation by *Nymphaea* for all non-*Amborella* members of ANITA is warranted.

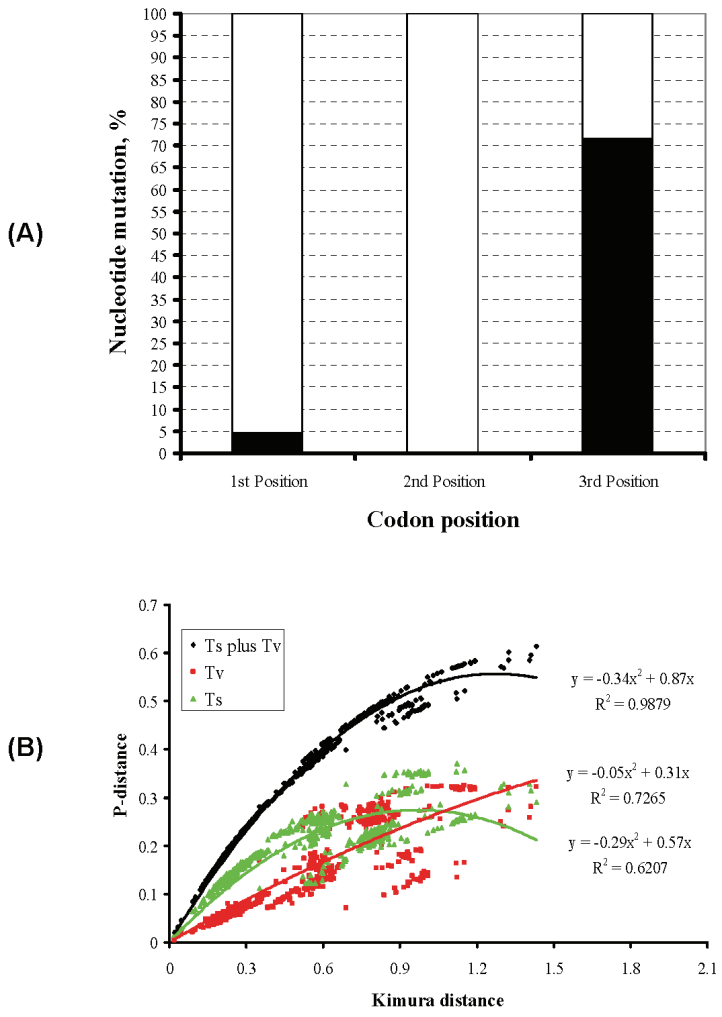

Figure 1. Analysis of neutrality and saturation of codon partitions. (**A**) Synonymous (the solid bar) and non-synonymous (the open bar) substitution possibilities calculated according to the standard genetic code (Griffiths 2007). (**B**) Relationship between uncorrected pairwise sequence divergence (p-distance) and Kimura-corrected distances, for transitions (Ts), transversions (Tv), or Ts plus Tv at the 3rd-codon positions of the whole chloroplast genome coding sequence alignment matrix of 36 land plant taxa compiled by Qiu *et al*. (2006).

Among the 12 phylogenetic trees we constructed (Table 1; Figs. 2-3, S1-10), the P_3Tv-based MP and ML trees were nearly identical (except *Calycanthus* placement) (Figs. 2-3). Moreover, the P_3Tv signal showed internally consistent

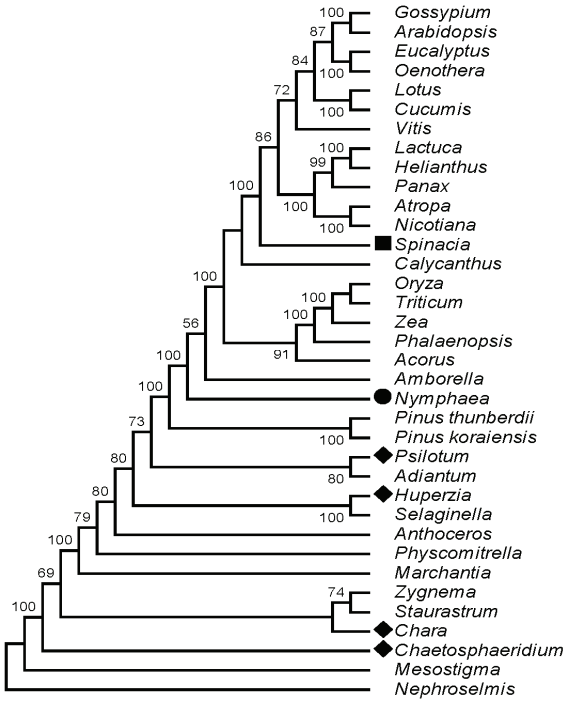

Figure 2. Phylogenetic tree created by maximum parsimony (MP) analysis based on P_3Tv of the whole chloroplast genome coding sequence alignment matrix of 36 land plant taxa compiled by Qiu *et al*. (Qiu et al. 2006). See the text for details about the improvement compared with Fig. 3A in Qiu *et al*. (Qiu et al. 2006). Inconsistencies are marked by the filled diamonds regarding deep divergence, the filled square in shallow divergence and the filled circle for most basal angiosperm.

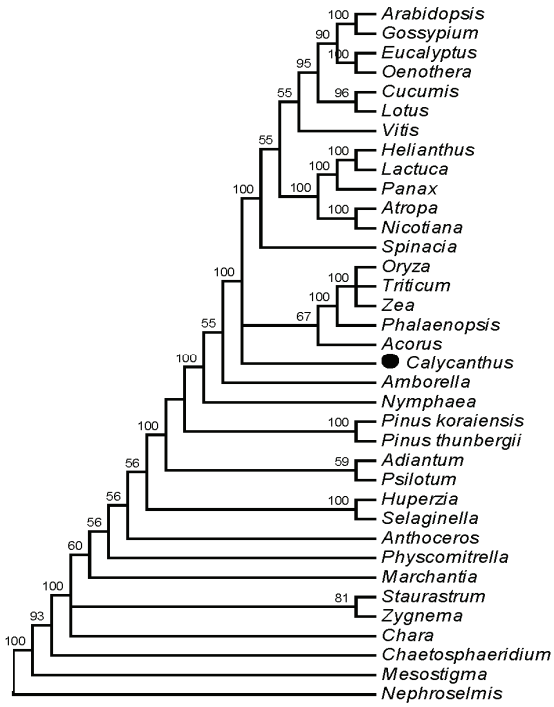

Figure 3. Maximum likelihood (ML) tree created based on transversions of 3rd codon positions of the whole chloroplast genome coding sequence alignment matrix of 36 land plant taxa compiled by Qiu *et al*. (2006). The filled circles represent conflicts with the maximum parsimony tree (Fig. 2) based on P_3Tv.

placements of *Chaetosphaeridium* basal to *Chara* and *Huperzia* basal to *Psilotum* in the derived phylogeny. These placements in the deep divergence are congruent with those in the multigene supermatrix tree (Fig. 1 in Qiu *et al*. (2006)) which was considered to be appropriate for deep divergence because it was reconstructed with six genes with slow to moderate evolutionary rates (*i.e.*, low saturation rate) and a large number of taxa for breaking long branches. Furthermore, the P_3Tv-based trees show proper resolution in shallow divergence, as exemplified by the appropriate placement of *Spinacia*, which is sister to both asterids and rosids (APG II 2003), as opposed to the chloroplast-genome tree (Fig. 3A in Qiu *et al*. (2006)) which grouped *Spinacia* with asterids using all codon positions as the phylogenetic signal.

If we accept the deep and shallow phylogenetic consistency of the P_3Tv signal as processing minimal stochastic and systematic error, then we should consider the P_3Tv-based mid-depth phylogeny as reliable. At the node of *Nymphaea–Amborella*, the core controversy in the angiosperm origin, both P_3Tv-based MP and ML trees clearly separated *Amborella* and *Nymphaea* (Figs. 2-3) and consistently indicate that *Nymphaea* is basal to *Amborella* with bootstrap values of 100. The best tree in Qiu *et al*. (2006) placed *Amborella* and *Nymphaea* as a co-basal sister clade. Among the other five pairs of MP and ML trees, only the 3rd codon Tv+Ts based MP and the ML trees (Figs. S1-S2) consistently placed *Amborella* basal to *Nymphaea*, while those based on the 1st+2nd codon Tv+Ts (Figs. S5-S6) placed with *Nymphaea* basal to

Table 1. Comparison of phylogenies created by maximum parsimony (MP) or maximum likelihood (ML) based on transversion (Tv) or transition (Ts) plus Tv of 1st, 2nd and 3rd codon partitions. The MP phylogeny (Fig. 2) based on P_3Tv was set as a reference for the comparisons, with "yes" indicating consistency and "no" indicating conflict.

		3rd positions		1st+2nd positions				1st+2nd+3rd positions			
		Ts + Tv		Tv		Ts + Tv		Tv		Ts + Tv	
		MP	ML	MP	ML	MP	ML	MP	ML	MP	ML
VASCULAR PLANTS	Corresponding to	Fig	Fig	Fig	Fig	Fig	Fig	Fig	Fig	Fig	Fig
Angiosperms		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Core eudicots											
Rosids	*Gossypium*	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Arabidopsis	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Eucalyptus	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Oenothera	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Lotus	yes	yes	**no**	**no**	yes	yes	yes	yes	yes	yes
	Cucumis	yes	yes	**no**	yes	yes	yes	yes	yes	yes	yes
	Vitis	yes	yes	**no**	**no**	**no**	yes	yes	yes	**no**	yes
Asterids	*Lactuca*	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Helianthus	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Panax	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Atropa	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Nicotiana	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Caryophyllales	*Spinacia*	**no**	**no**	**no**	**no**	**no**	**no**	yes	**no**	**no**	**no**
Magnolids	*Calycanthus*	yes	**no**	yes	yes	yes	yes	yes	yes	yes	yes
Monocots											
Poales	*Oryza*	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Triticum	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Zea	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Asparagales	*Phalaenopsis*	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Acorales	*Acorus*	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Basal angiosperms											
	Amborella	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Nymphaea	**no**	**no**	**no**	**no**	yes	yes	yes	**no**	**no**	**no**
Gymnosperms	*Pinus thunberdii*	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Pinus koraiensis	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Monilophytes	*Psilotum*	**no**	yes	yes	yes	yes	yes	yes	yes	**no**	**no**
	Adiantum	yes	yes	yes	yes	yes	yes	yes	yes	**no**	**no**
Lycophytes	*Huperzia*	**no**	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Selaginella	**no**	yes	yes	yes	yes	yes	yes	yes	yes	yes
HORNWORTS	*Anthoceros*	yes	yes	**no**	yes	yes	yes	yes	yes	yes	yes
MOSSES	*Physcomitrella*	yes	yes	yes	**no**	yes	yes	yes	yes	yes	yes
LIVERWORTS	*Marchantia*	yes	yes	**no**	**no**	**no**	**no**	yes	yes	yes	yes
CHAROPHYTES	*Zygnema*	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Staurastrum	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Chara	yes	yes	**no**	**no**	**no**	**no**	**no**	yes	**no**	**no**
	Chaetosphaeridium	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Mesostigma	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Nephroselmis	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Number of conflicts		**5**	**3**	**8**	**7**	**4**	**3**	**1**	**2**	**6**	**5**

Amborella, and remaining three pairs of MP and ML trees based on 1st+2nd codon Tv (Figs. S3-S4), 1st+2nd+3rd codon Tv (Figs. S7-S8), and 1st+2nd+3rd codon Tv+Ts (Figs. S9-S10) were inconsistent, with ML trees showing *Amborella* and *Nymphaea* co-basal (Figs. S4, S8, S10), one

Nymphaea-basal-to-*Amborella* MP tree (Fig. S7), and two *Amborella*-basal-to-*Nymphaea* MP trees (Figs. S3, S9). These incongruent trees suggest the effect of either saturation or non-synonymous mutations (Chaw et al. 2000; Nickrent et al. 2000; Stefanovic, Rice, and Palmer 2004; Qiu et al. 2006). Furthermore, all non-P_3Tv trees (Figs. S1-S10) contain weak support for the shallow and/or deep portion of their phylogenies. Considering the drastic developmental differences between *Nymphaea* (and the other non-*Amborella* ANITA members) and *Amborella* (see below), the inability to separate *Nymphaea* from *Amborella* suggests the presence of some systematic errors in using these non-P_3Tv-based signals for phylogenetic inference. Despite the support for both shallow and deep elements of the land plant phylogeny, the use of P_3Tv did not definitively resolve the relationship between *Amborella* and the rest of the angiosperm tree. The P_3Tv-based trees placed *Amborella* with 100% bootstrap support as separated from *Nymphaea* and sister to all other angiosperms 55-56% of the time (Figs. 2-3).

Nymphaea–basal angiosperm tree concordant to developmental evidence

Since all phylogenetic inferences are reconstructions based on the molecular sequences of extant plant species, the incongruence among various trees constructed by different phylogenetic signals or by different tree-construction methods are difficult to cross verify. The recent emergence of evolutionary developmental biology, or "evo-devo", opened a new approach for cross verifying molecular phylogenies with developmental data (Goodman and Coughlin 2000). The evolution of the female gametophyte is one such developmental characteristic. Our P_3TV-based angiosperm phylogeny in the overall land plant phylogeny appears concordant with the evolution of female gametophytes from 4-cell/4-nucleus (ANITA except *Amborella*) to 8-cell/9-nucleus (*Amborella*) or 7-cell/8-nucleus (the non-ANITA angiosperms) (Williams and Friedman 2002; Friedman and Williams 2003; Friedman and Williams 2004; Friedman 2006) and the evolution of endosperm from 1n (ancestral gymnosperm) to 2n (ANITA except *Amborella*) to 3n (*Amborella* and the non-ANITA angiosperms), as illustrated in Fig. 4. Such an interpretation cannot rule out the possibility that the progenitor of the modern monocots, eudicots and

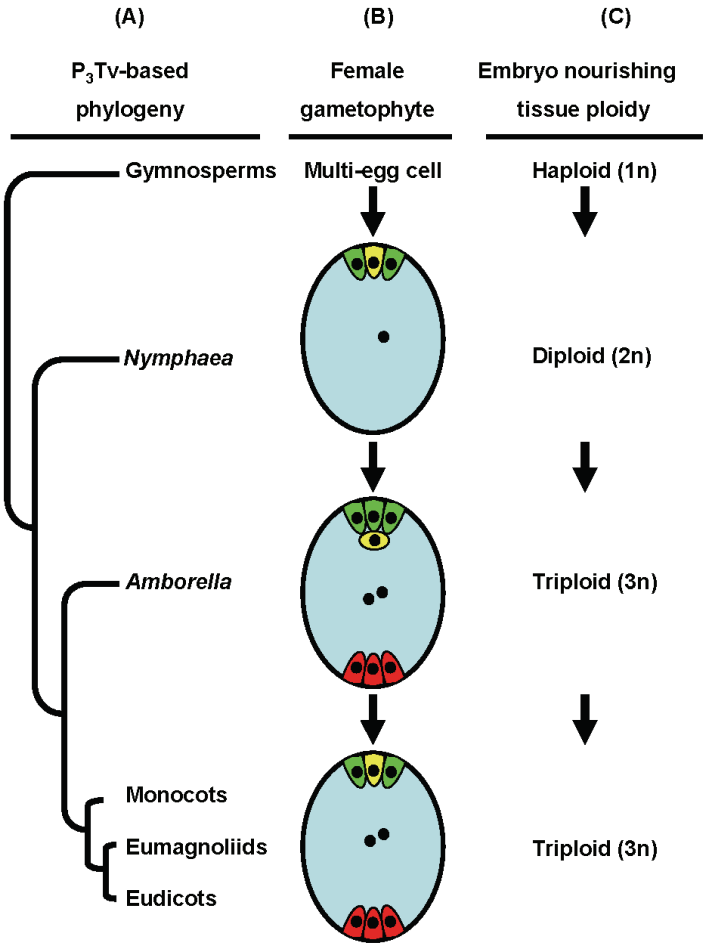

Figure 4. The reconciliation of the P_3Tv-based molecular phylogeny (**A**) and the corresponding evolution of female gametophyte (**B**) and ploidy of endosperm (**C**) (Williams and Friedman 2002; Friedman and Williams 2003; Friedman and Williams 2004; 2006). During double fertilizations, the central cell is fused with a sperm cell (1n), producing diploid (2n) endosperm in *Nymphaea*, and triploid (3n) endosperm in *Amborella* and the non-ANITA angiosperms. Egg cell, yellow; central cell, blue; synergid cells, green; antipodal cells, red; and nucleus, black.

eumagnoliids acquired triploid endosperms independently. Still the most parsimonious conclusion is that the progenitor of modern angiosperms descended from *Nymphaea*. Our *Nymphaea*-basal tree is also supported by the evolution of angiosperm leaf stomata. *Nymphaea* showed little stomata modification from the hypothetical ancestral angiosperm stomata, while *Amborella* exhibited extensive stomata modification (*e.g.*, tangential divisions of contact cells) (Carpenter 2005). Our *Nymphaea*-basal phylogeny also agrees with the fossil evidence that *Nymphaea* extended back to the early Cretaceous (125-115 mya) and into the oldest fossil assemblages that contain unequivocal angiosperm stamens and carpels (Friis, Pedersen, and Crane 2001).

In conclusion, our study demonstrates that P_3Tv is a robust phylogenetic signal for reconstruction of plant evolutionary history and may have broad applications for phylogenetic inference of other biological lineages. Using this new signal, we generated a congruent phylogeny that is consistent across land plant lineages in both deep and shallow divergences and contributes to the resolution of the conflict between the current molecular phylogeny and the developmental evidence on the root of angiosperms.

Literature Cited:

APG II. 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Botanical Journal of the Linnean Society **141**:399-436.

Borsch, T., K. W. Hilu, D. Quandt, V. Wilde, C. Neinhuis, and W. Barthlott. 2003. Noncoding plastid trnT-trnF sequences reveal a well resolved phylogeny of basal angiosperms. Journal of Evolutionary Biology **16**:558-576.

Cai, J. J., D. K. Smith, X. Xia, and K. Y. Yuen. 2005. MBEToolbox: a MATLAB toolbox for sequence data analysis in molecular biology and evolution. BMC Bioinformatics **6**:64.

Carpenter, K. J. 2005. Stomatal architecture and evolution in basal angiosperms. American Journal of Botany **92**:1595-1615.

Chang, C. C., H. C. Lin, I. P. Lin et al. 2006. The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. Mol Biol Evol **23**:279-291.

Chaw, S. M., C. L. Parkinson, Y. Cheng, T. M. Vincent, and J. D. Palmer. 2000. Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. Proc Natl Acad Sci U S A **97**:4086-4091.

Darwin, C. 1903. Letter to J. D. Hooker *in* F. Darwin, and A. C. Seward, eds. More letters of Charles Darwin. John Murray, London.

De Bodt, S., S. Maere, and Y. Van de Peer. 2005. Genome duplication and the origin of angiosperms. Trends Ecol Evol **20**:591-597.

Friedman, W. E. 2006. Embryological evidence for developmental lability during early angiosperm evolution. Nature **441**:337-340.

Friedman, W. E., and J. H. Williams. 2003. Modularity of the angiosperm female gametophyte and its bearing on the early evolution of endosperm in flowering plants. Evolution Int J Org Evolution **57**:216-230.

Friedman, W. E., and J. H. Williams. 2004. Developmental evolution of the sexual process in ancient flowering plant lineages. Plant Cell **16 Suppl**:S119-132.

Friis, E. M., K. R. Pedersen, and P. R. Crane. 2001. Fossil evidence of water lilies (Nymphaeales) in the Early Cretaceous. Nature **410**:357-360.

Goodman, C. S., and B. C. Coughlin. 2000. Introduction. The evolution of evo-devo biology. Proc Natl Acad Sci U S A **97**:4424-4425.

Goremykin, V. V., and F. H. Hellwig. 2006. A new test of phylogenetic model fitness addresses the issue of the basal angiosperm phylogeny. Gene **381**:81-91.

Goremykin, V. V., K. I. Hirsch-Ernst, S. Wolfl, and F. H. Hellwig. 2004. The chloroplast genome of Nymphaea alba: whole-genome analyses and the problem of identifying the most basal angiosperm. Mol Biol Evol **21**:1445-1454.

Goremykin, V. V., K. I. Hirsch-Ernst, S. Wolfl, and F. H. Hellwig. 2003. Analysis of the Amborella trichopoda chloroplast genome sequence suggests that amborella is not a basal angiosperm. Mol Biol Evol **20**:1499-1505.

Graham, S. W., and R. G. Olmstead. 2000. Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. American Journal of Botany **87**:1712-1730.

Griffiths, A. J. F. 2007. Introduction to genetic analysis. W.H,. Freeman and Co., New York, NY.

Jansen, R. K., C. Kaittanis, C. Saski, S. B. Lee, J. Tomkins, A. J. Alverson, and H. Daniell. 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 Evol Biol **6**:32.

Jeffroy, O., H. Brinkmann, F. Delsuc, and H. Philippe. 2006. Phylogenomics: the beginning of incongruence? Trends Genet **22**:225-231.

Jobb, G., A. von Haeseler, and K. Strimmer. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol **4**:18.

Keane, T. M., C. J. Creevey, M. M. Pentony, T. J. Naughton, and J. O. McLnerney. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol Biol **6**:29.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol **16**:111-120.

Leebens-Mack, J., L. A. Raubeson, L. Cui, J. V. Kuehl, M. H. Fourcade, T. W. Chumley, J. L. Boore, R. K. Jansen, and C. W. depamphilis. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Mol Biol Evol **22**:1948-1963.

Lin, C. Y., F. K. Lin, C. H. Lin, L. W. Lai, H. J. Hsu, S. H. Chen, and C. A. Hsiung. 2005. POWER: PhylOgenetic WEb Repeater--an integrated and user-optimized framework for biomolecular phylogenetic analysis. Nucleic Acids Res **33**:W553-556.

Magallon, S., and M. J. Sanderson. 2002. Relationships among seed plants inferred from highly conserved genes: Sorting conflicting phylogenetic signals among ancient lineages. American Journal of Botany **89**:1991-2006.

Mathews, S., and M. J. Donoghue. 1999. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. Science **286**:947-950.

Nickrent, D. L., C. L. Parkinson, J. D. Palmer, and R. J. Duff. 2000. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. Mol Biol Evol **17**:1885-1895.

Parkinson, C. L., K. L. Adams, and J. D. Palmer. 1999. Multigene analyses identify the three earliest lineages of extant flowering plants. Curr Biol **9**:1485-1488.

Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol **21**:1455-1458.

Phillips, M. J., and D. Penny. 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. Mol Phylogenet Evol **28**:171-185.
Qiu, Y. L., J. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, E. A. Zimmer, Z. Chen, V. Savolainen, and M. W. Chase. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature **402**:404-407.
Qiu, Y. L., J. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, E. A. Zimmer, Z. Chen, V. Savolainen, and M. W. Chase. 2000. Phylogeny of basal angiosperms: Analyses of five genes from three genomes. International Journal of Plant Sciences **161**:S3-S27.
Qiu, Y. L., L. Li, B. Wang et al. 2006. The deepest divergences in land plants inferred from phylogenomic evidence. Proc Natl Acad Sci U S A **103**:15511-15516.
Simmons, M. P., L. B. Zhang, C. T. Webb, and A. Reeves. 2006. How can third codon positions outperform first and second codon positions in phylogenetic inference? An empirical example from the seed plants. Syst Biol **55**:245-258.
Soltis, P. S., D. E. Soltis, and M. W. Chase. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature **402**:402-404.
Stefanovic, S., D. W. Rice, and J. D. Palmer. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: Amborella or monocots? BMC Evol Biol **4**:35.
Van de Peer, Y., T. Frickey, J. Taylor, and A. Meyer. 2002. Dealing with saturation at the amino acid level: a case study based on anciently duplicated zebrafish genes. Gene **295**:205-211.
Williams, J. H., and W. E. Friedman. 2002. Identification of diploid endosperm in an early angiosperm lineage. Nature **415**:522-526.
Zanis, M. J., D. E. Soltis, P. S. Soltis, S. Mathews, and M. J. Donoghue. 2002. The root of the angiosperms revisited. Proceedings of the National Academy of Sciences of the United States of America **99**:6848-6853.
Zanis, M. J., P. S. Soltis, Y. L. Qiu, E. Zimmer, and D. E. Soltis. 2003. Phylogenetic analyses and perianth evolution in basal angiosperms. Annals of the Missouri Botanical Garden **90**:129-150.

Acknowledgments

We thank S. D. Wullschleger, F. Chen and L. Gunter for critical comments on the manuscript. This work was supported by grants from the National Science Foundation to GAT and ZMC, a DOE-ORNL subcontract to ZMC, and an ORNL LDRD grant to TJT. Research sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U. S. Department of Energy under Contract No. DE-AC05-00OR22725.

Supplementary Materials

Figure S1. Maximum parsimony (MP) tree based on transversions plus transitions of the 3rd codon position of the whole chloroplast genome coding sequences of 36 land plant taxa compiled by Qiu *et al*. (2006). The filled circles represent conflicts with the MP tree (Fig. 2) based on P_3Tv.

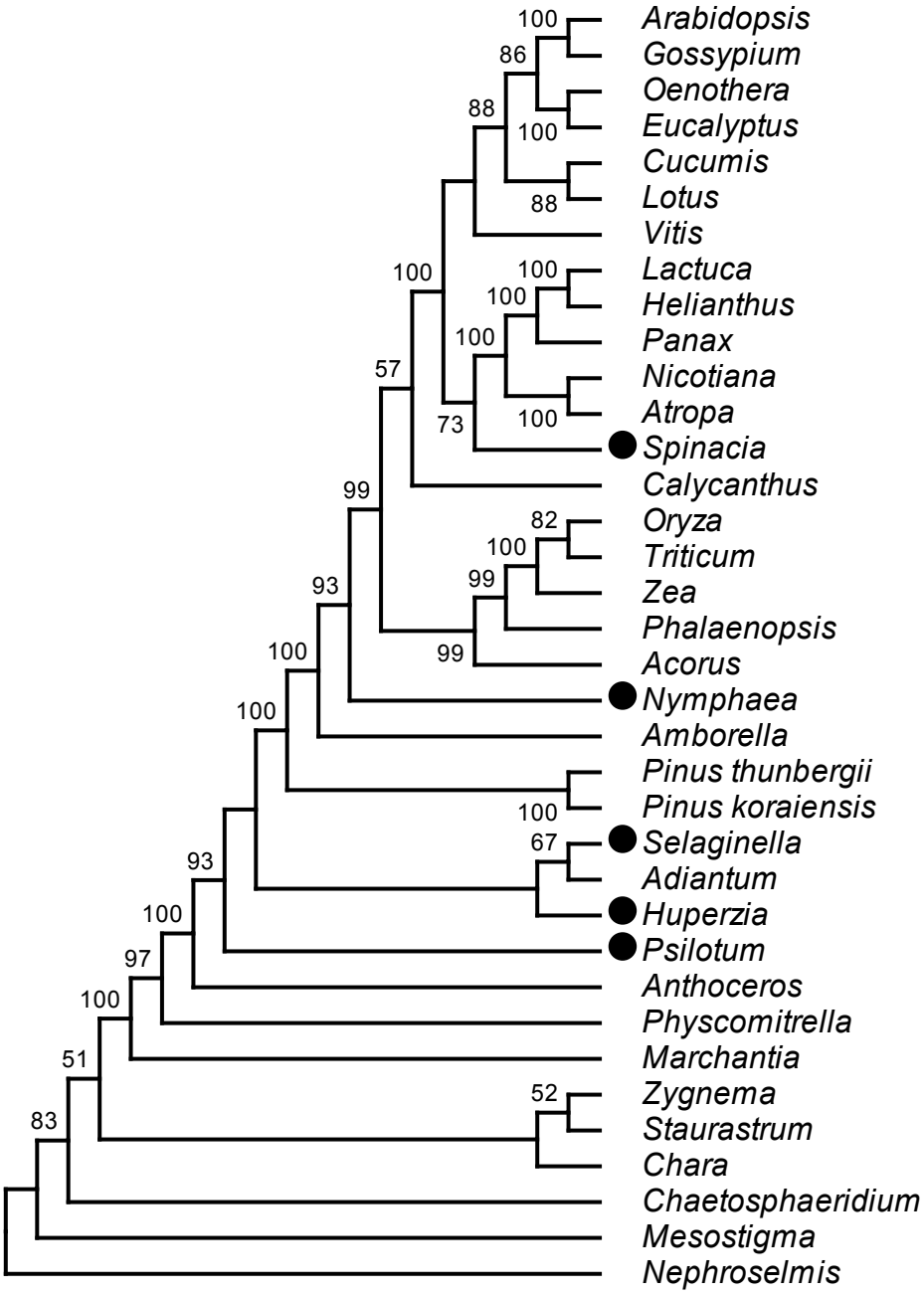

Figure S2. Maximum likelihood (ML) tree based on transversions plus transitions of 3rd codon positions of the whole chloroplast genome coding sequences of 36 land plant taxa compiled by Qiu *et al*. (2006). The filled circles represent conflicts with the maximum parsimony tree (Fig. 2) based on P_3Tv.

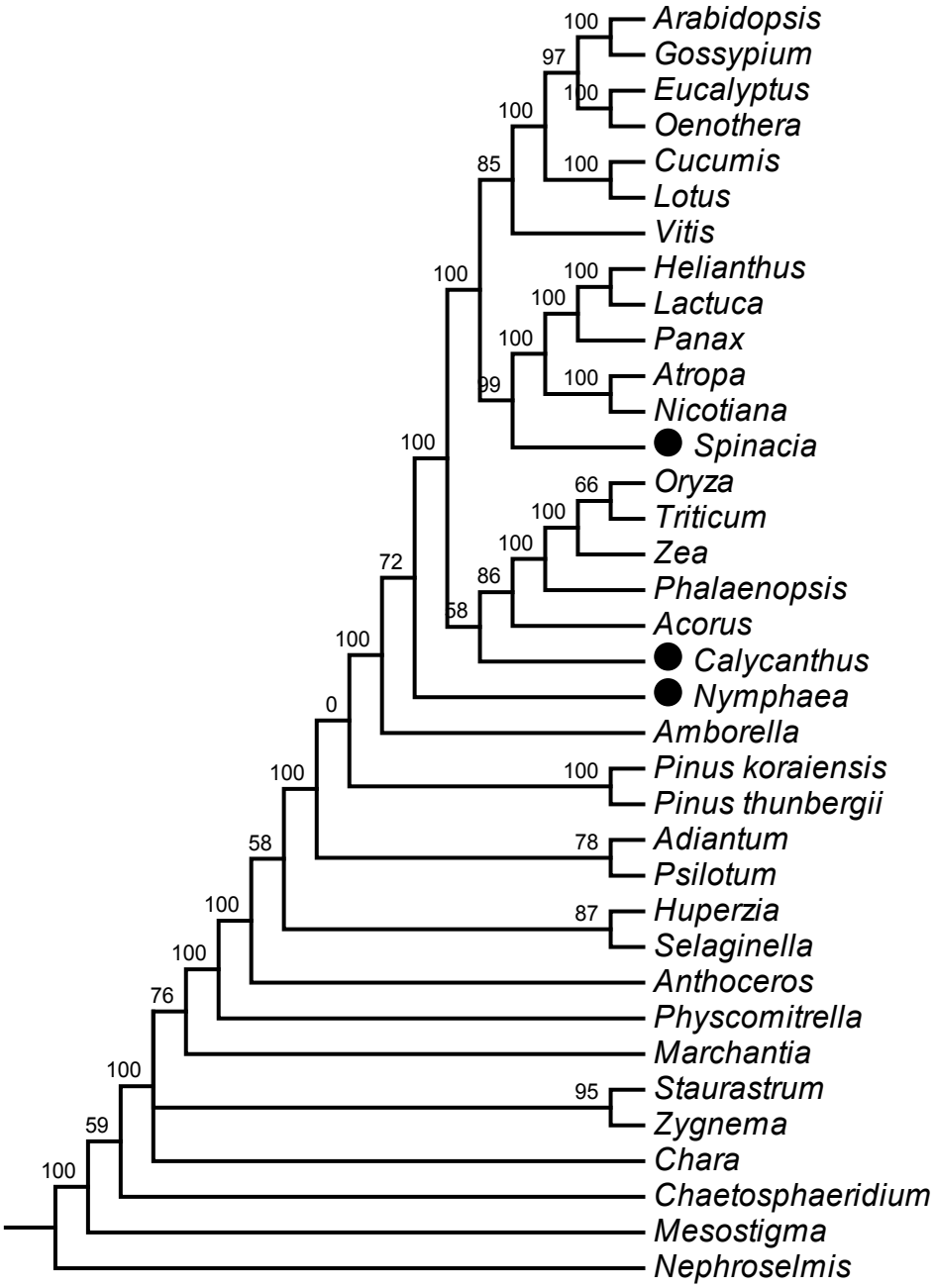

Figure S3. Maximum parsimony (MP) tree based on transversions of $1^{st}+2^{nd}$ codon positions of the whole chloroplast genome coding sequences of 36 land plant taxa compiled by Qiu *et al*. (2006). The filled circles represent conflicts with the MP tree (Fig. 2) based on P_3Tv.

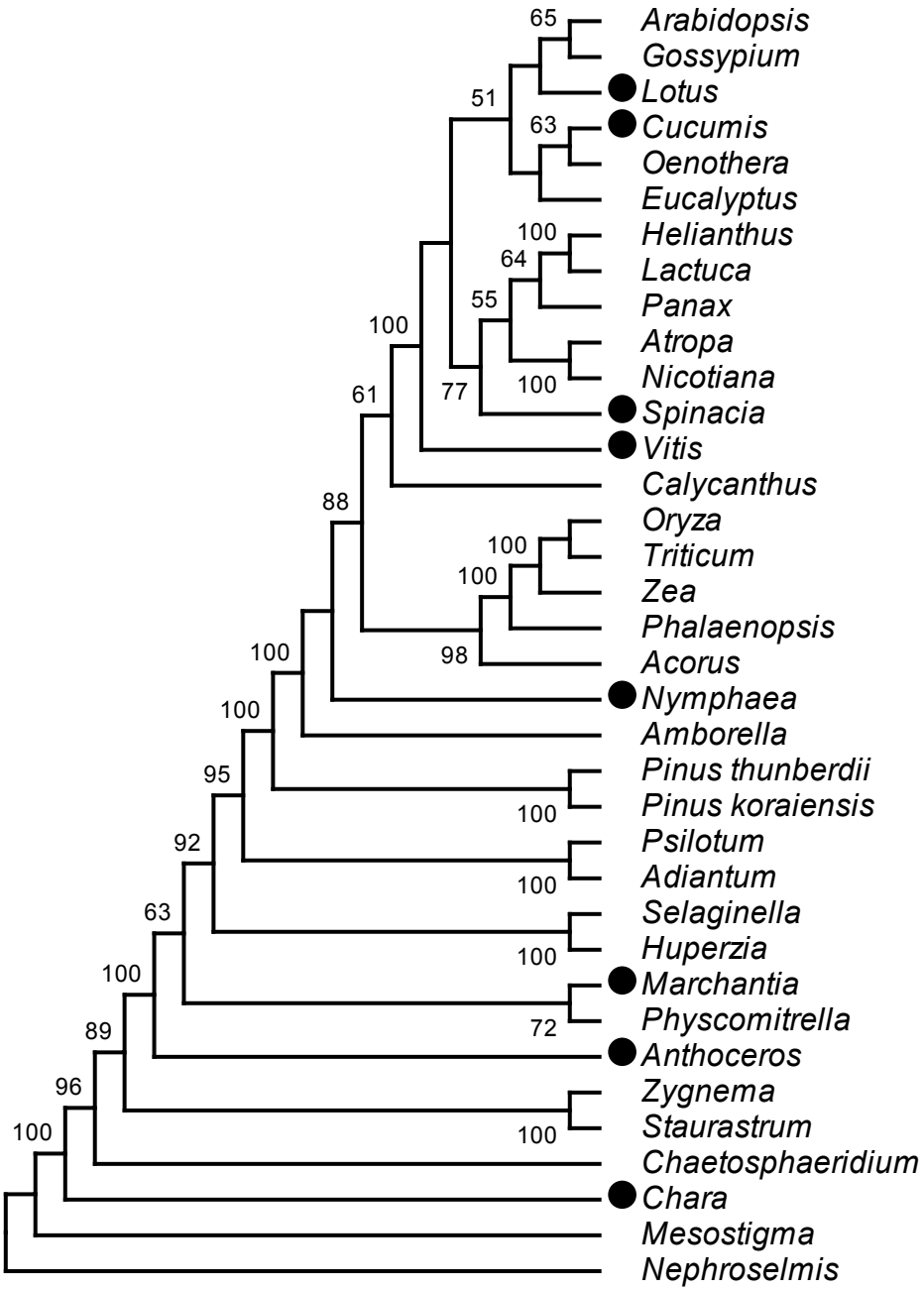

Figure S4. Maximum likelihood (ML) tree based on transversions of 1st+2nd codon positions of the whole chloroplast genome coding sequences of 36 land plant taxa compiled by Qiu *et al*. (2006). The filled circles represent conflicts with the maximum parsimony tree (Fig. 2) based on P_3Tv.

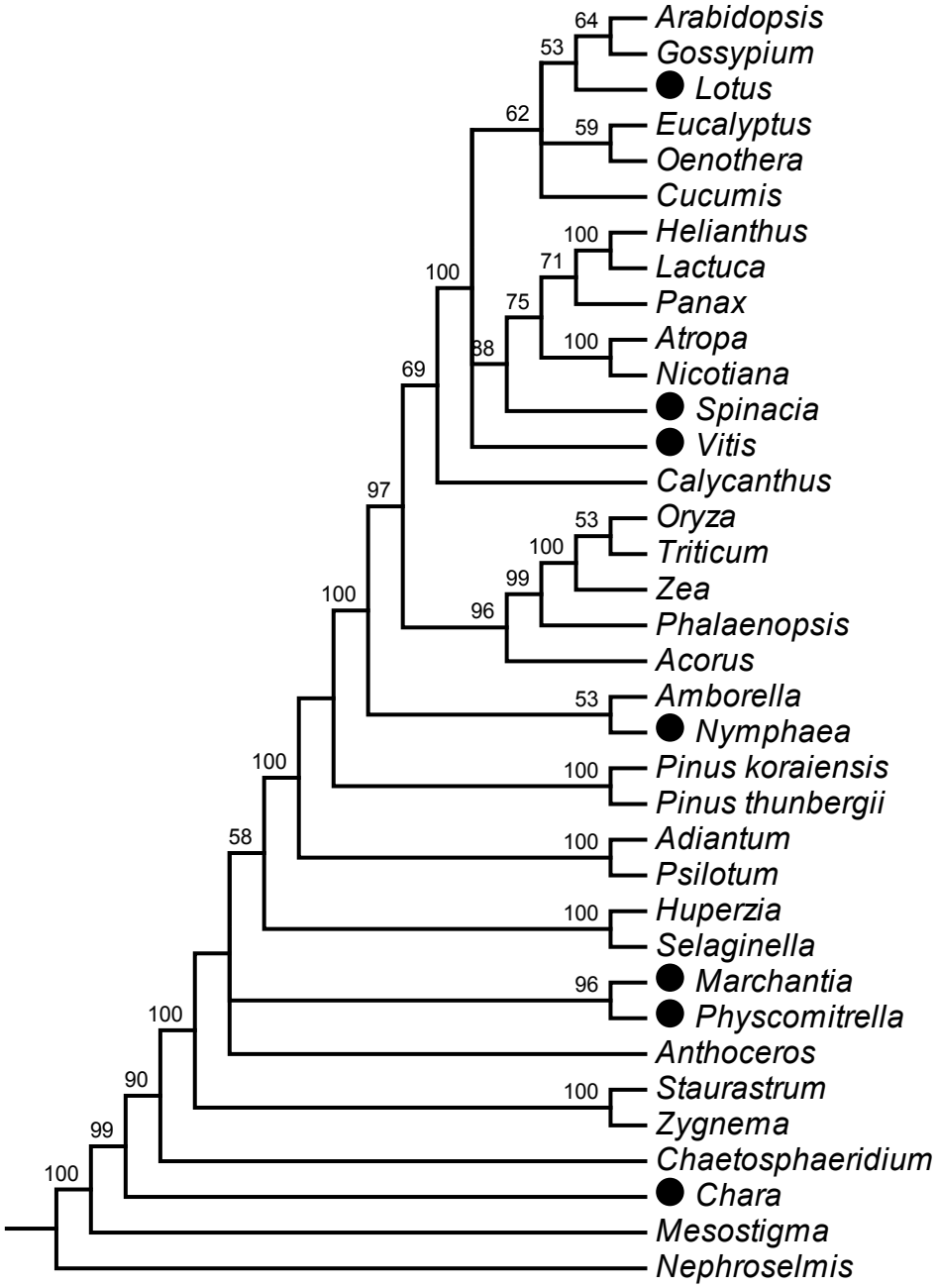

Figure S5. Maximum parsimony (MP) tree based on transversions plus transitions of $1^{st}+2^{nd}$ codon positions of the whole chloroplast genome coding sequences of 36 land plant taxa compiled by Qiu *et al*. (2006). The filled circles represent conflicts with the MP tree (Fig. 2) based on P_3Tv.

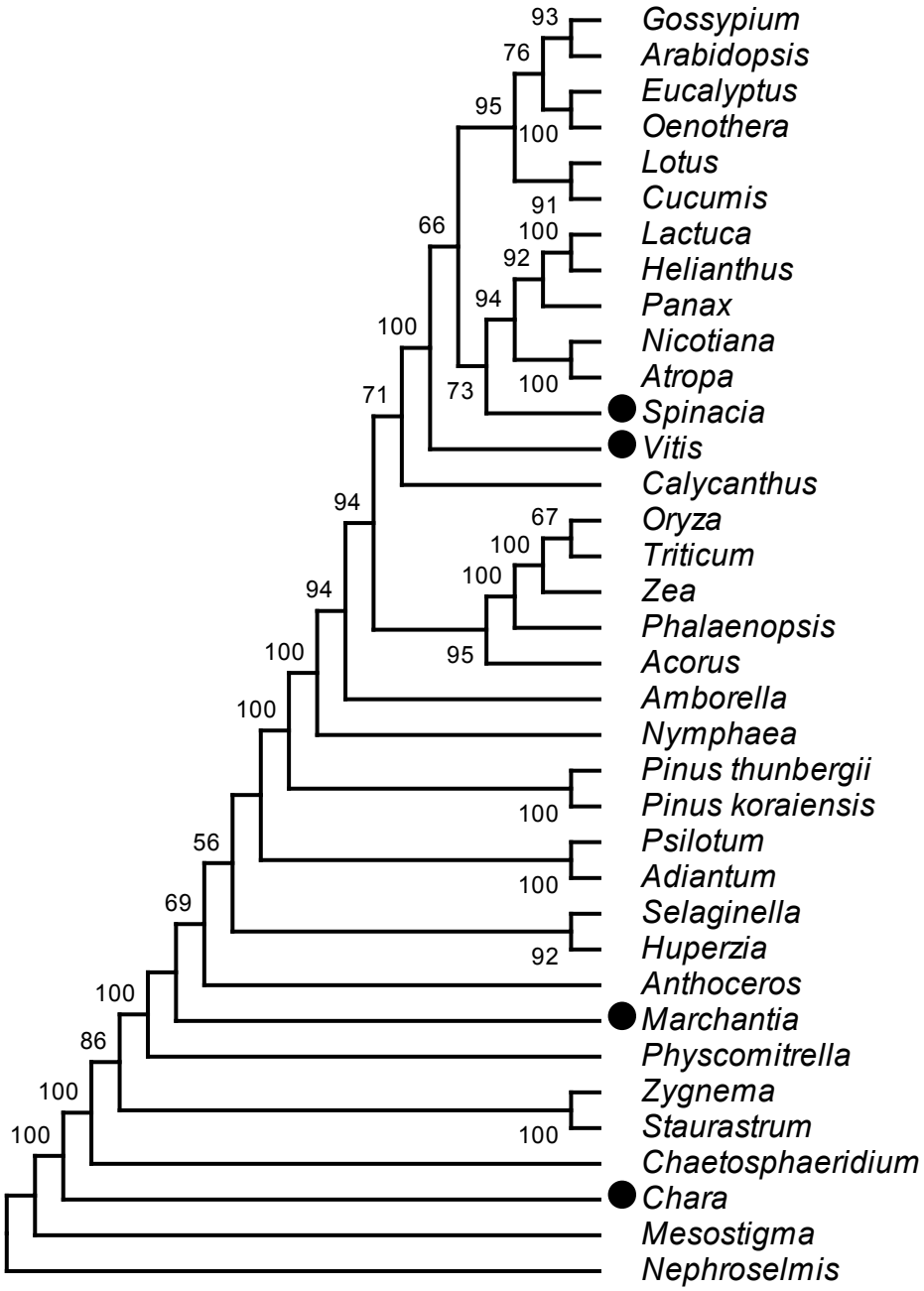

Figure S6. Maximum likelihood (ML) tree based on transversions plus transitions of $1^{st}+2^{nd}$ codon positions of the whole chloroplast genome coding sequences of 36 land plant taxa compiled by Qiu *et al*. (2006). The filled circles represent conflicts with the maximum parsimony tree (Fig. 2) based on P_3Tv.

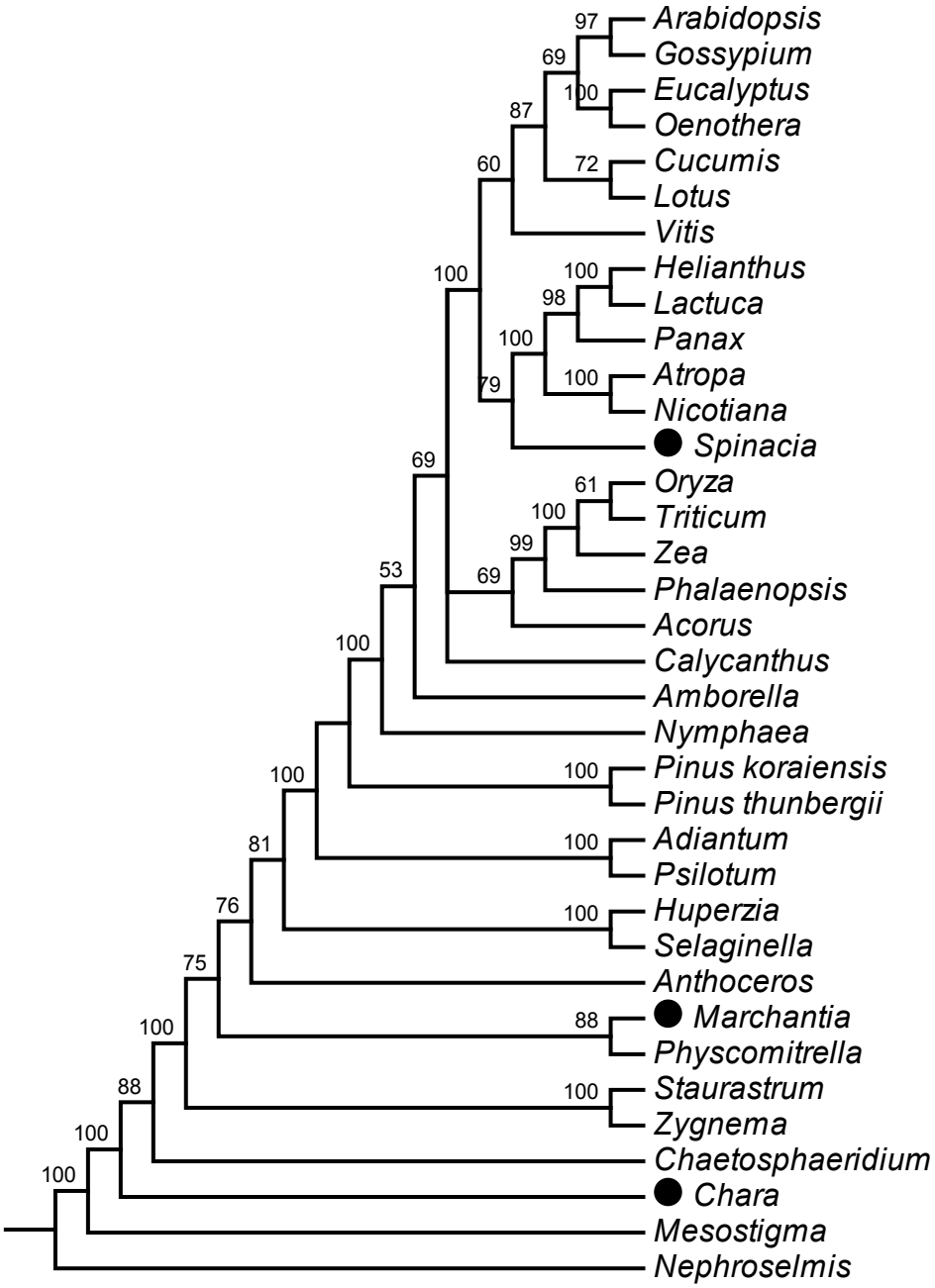

Figure S7. Maximum parsimony (MP) tree based on transversions of $1^{st}+2^{nd}+3^{rd}$ codon positions of the whole chloroplast genome coding sequences of 36 land plant taxa compiled by Qiu *et al*. (2006). The filled circles represent conflicts with the maximum parsimony tree (Fig. 2) based on P_3Tv.

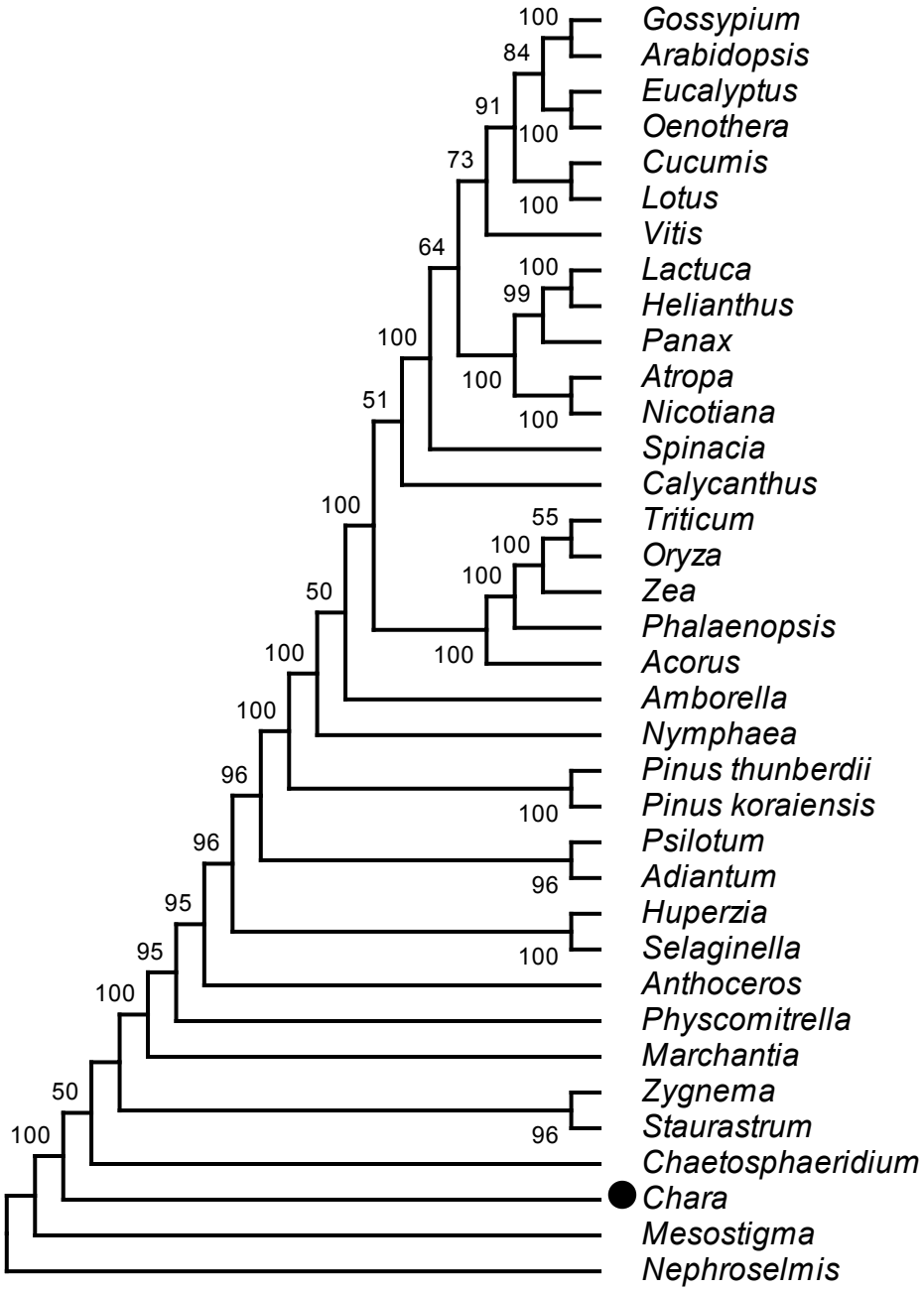

Figure S8. Maximum likelihood (ML) tree based on transversions of $1^{st}+2^{nd}+3^{rd}$ codon positions of the whole chloroplast genome coding sequences of 36 land plant taxa compiled by Qiu *et al*. (2006). The filled circles represent conflicts with the maximum parsimony tree (Fig. 2) based on P_3Tv.

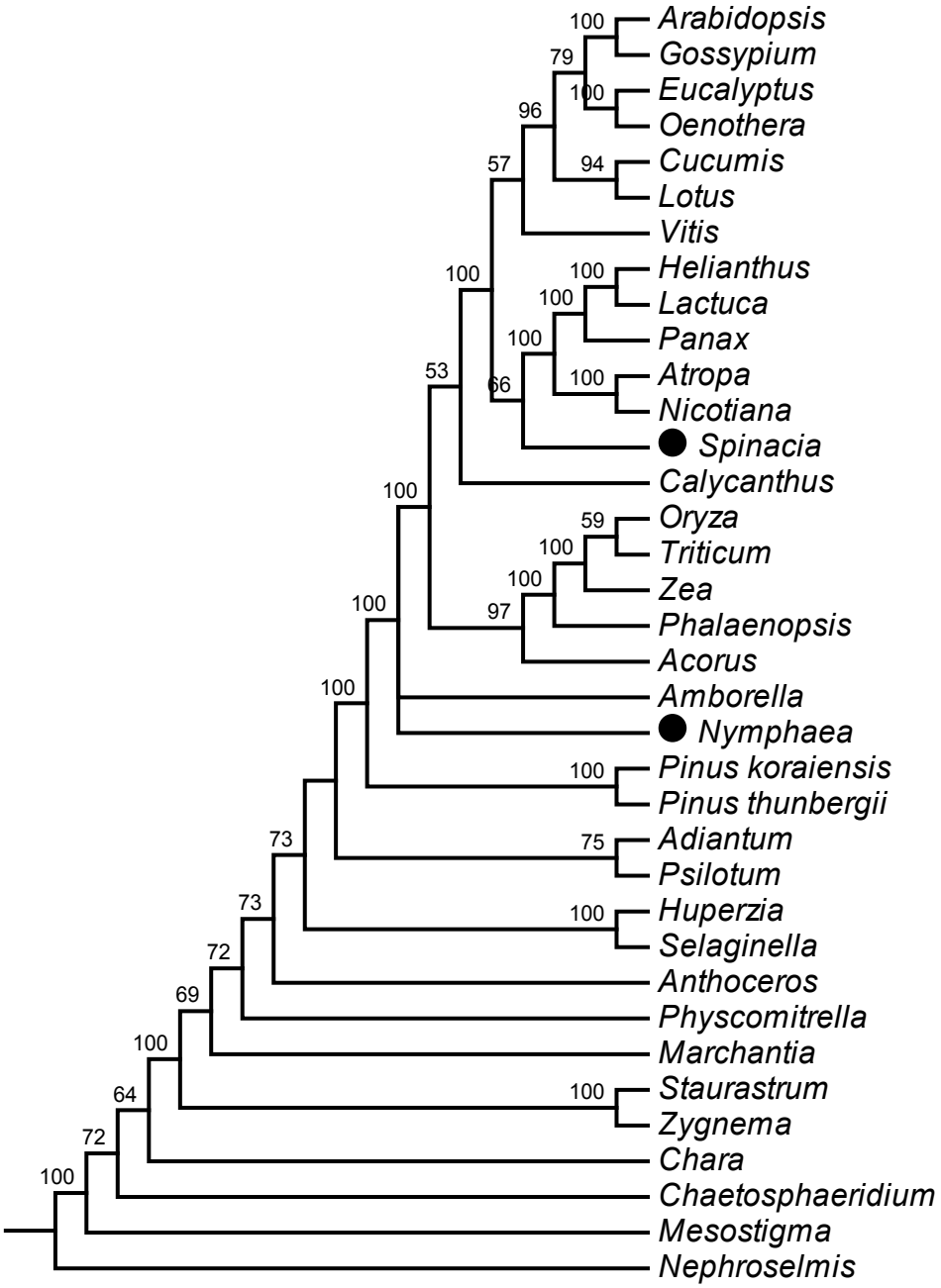

Figure S9. Maximum parsimony (MP) tree based on transversions plus transitions of $1^{st}+2^{nd}+3^{rd}$ codon positions of the whole chloroplast genome coding sequences of 36 land plant taxa compiled by Qiu *et al*. (2006). The filled circles represent conflicts with the MP tree (Fig. 2) based on P_3Tv.

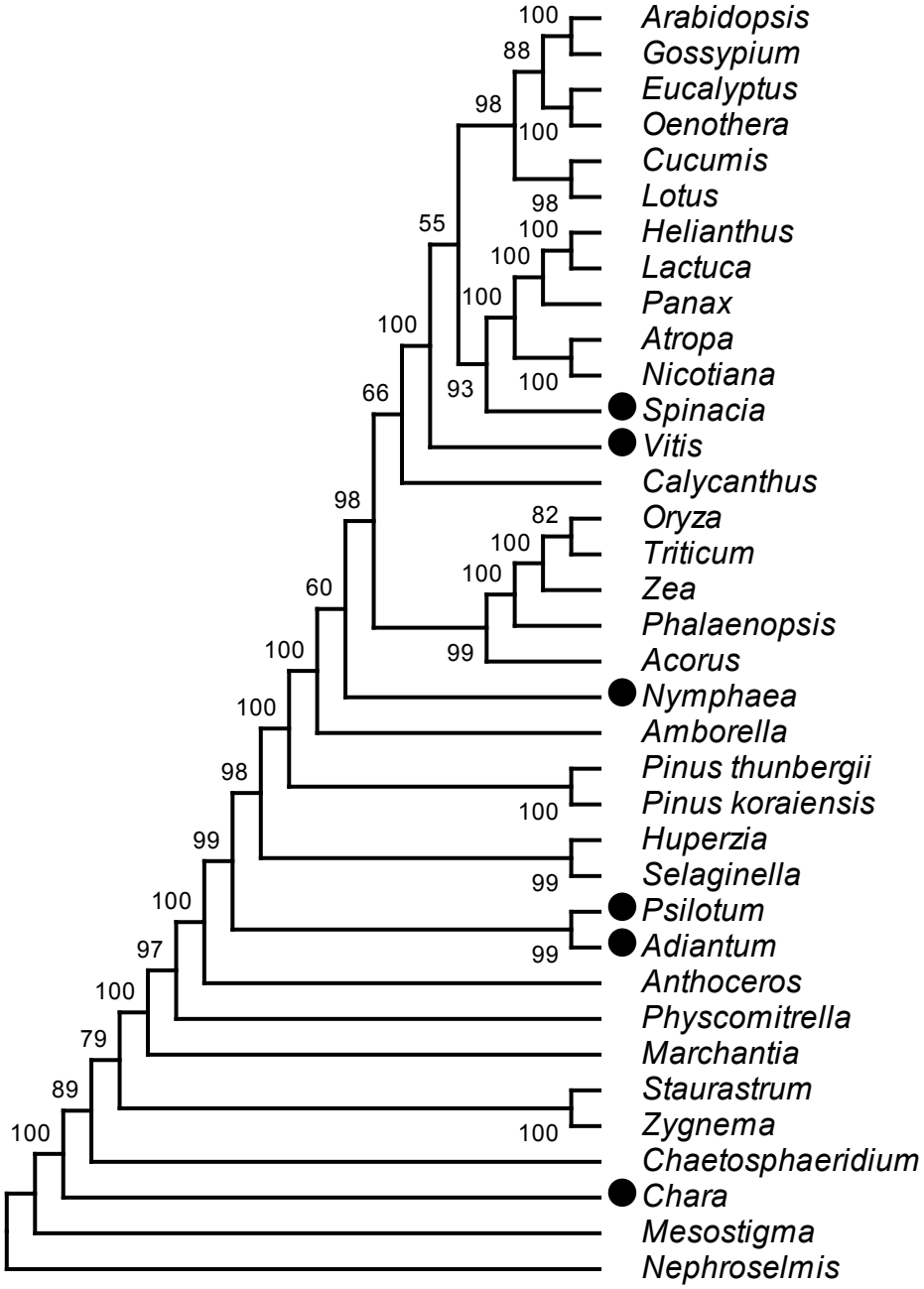

Figure S10. Maximum likelihood (ML) tree based on transversions plus transitions of $1^{st}+2^{nd}+3^{rd}$ codon positions of the whole chloroplast genome coding sequences of 36 land plant taxa compiled by Qiu *et al*. (2006). The filled circles represent conflicts with the maximum parsimony tree (Fig. 2) based on P_3Tv.