Complete chloroplast genome sequences of *Drimys*, *Liriodendron*, and *Piper*: Implications for the phylogeny of magnoliids and the evolution of GC content

C. Zhengqiu [1], C. Penaflor[2], J.V. Kuehl[3], J. Leebens-Mack[4], J. Carlson[4], C. W. dePamphilis[4,]J. L. Boore[3], and R. K. Jansen[1*]

[1]Section of Integrative Biology and Institute of Cellular and Molecular Biology, Patterson Laboratories 141, University of Texas, Austin, TX 78712, USA; [2]Biology Department, 373 WIDB, Brigham Young University, Provo, Utah 84602, USA; [3]DOE Joint Genome Institute and Lawrence Berkeley National Laboratory, Walnut Creek, CA; [4]Department of Biology, Huck Institutes of Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA.

*Corresponding author:

Robert K. Jansen

Section of Integrative Biology and Institute of Cellular and Molecular Biology

Patterson Laboratories 141

University of Texas at Austin

Austin, TX 78712

Phone: 512-471-8827

Fax: 512-232-9529

E-mail: jansen@mail.utexas.edu

Abstract

Background: The magnoliids represent the largest basal angiosperm clade with four orders, 19 families and 8,500 species. Although several recent angiosperm molecular phylogenies have supported the monophyly of magnoliids and suggested relationships among the orders, the limited number of genes examined resulted in only weak support, and these issues remain controversial. Furthermore, considerable incongruence has resulted in phylogenies supporting three different sets of relationships among magnoliids and the two large angiosperm clades, monocots and eudicots. This is one of the most important remaining issues concerning relationships among basal angiosperms. We sequenced the chloroplast genomes of three magnoliids, *Drimys* (Canellales), *Liriodendron* (Magnoliales), and *Piper* (Piperales), and used these data in combination with 32 other completed angiosperm chloroplast genomes to assess phylogenetic relationships among magnoliids.

Results: The *Drimys* and *Piper* chloroplast genomes are nearly identical in size at 160,606 and 160,624 bp, respectively. The genomes include a pair of inverted repeats of 26,649 bp (*Drimys*) and 27,039 (*Piper*), separated by a small single copy region of 18,621 (*Drimys*) and 18,878 (*Piper*) and a large single copy region of 88,685 bp (*Drimys*) and 87,666 bp (*Piper*). The gene order of both taxa is nearly identical to many other unrearranged angiosperm chloroplast genomes, including *Calycanthus*, the other published magnoliid genome. Comparisons of angiosperm chloroplast genomes indicate that GC content is not uniformly distributed across the genome. Overall GC content ranges from 34-39%, and coding regions have a substantially higher GC content than non-coding regions (both intergenic spacers and introns). Among protein-coding genes, GC content varies by codon position with 1st codon > 2nd codon > 3rd codon, and it varies by functional group with photosynthetic genes having the highest percentage and NADH genes the lowest. Across the genome, GC content is highest in the inverted repeat due to the presence of rRNA genes and lowest in the small single copy region where most NADH genes are located. Phylogenetic analyses using maximum parsimony and maximum likelihood methods were performed on DNA sequences of 61 protein-coding genes. Trees from both analyses provided strong support for the monophyly of magnoliids and two strongly supported groups were identified, the Canellales/Piperales and the Laurales/ Magnoliales. The phylogenies also provided moderate to strong support for the basal position of *Amborella*, and a sister relationship of magnoliids to a clade that includes monocots and eudicots.

Conclusions: The complete sequences of three magnoliid chloroplast genomes provide new data from the largest basal angiosperm clade. Evolutionary comparisons of these new genome sequences, combined with other published angiosperm genome, confirm that GC content is unevenly distributed across the genome by location, codon position, and functional group. Furthermore, phylogenetic analyses provide the strongest support so far for the hypothesis that the magnoliids are sister to a large clade that includes both monocots and eudicots.

Background

Phylogenetic relationships among basal angiosperms have been debated for over 100 years since Darwin [1903] identified this issue as an "abominable mystery". The difficulty of resolving these relationships is likely due to the rapid radiation of angiosperms immediately after their origin and subsequent extinction. During the past decade there has been considerable interest in developing phylogenies based on single or multiple genes to resolve the basal radiation of flowering plants [Donoghue and Mathews 1998, Mathews and Donoghue 1999, Barkman *et al.* 2000, Doyle and Endress 2000, Graham and Olmstead 2000, Zanis *et al.* 2002, Borsch *et al.* 2003, Goremykin *et al.* 2003, 2004, 2005; Davies *et al.* 2004, Soltis and Soltis 2004, Qiu *et al.* 1999, 2000, 2005, 2006, Leebens-Mack *et al.* 2005]. Recently, completely sequenced chloroplast genomes have been used to estimate relationships among basal angiosperms [Goremykin et al. 2003, 2004, 2005, Leebens-Mack et al. 2005] and these studies have converged on a consensus that the most basal lineage of flowering plants are either *Amborella* or *Amborella* plus the Nymphaelales. Although the whole chloroplast genome approach has enhanced our understanding of basal angiosperm relationships, issues of taxon sampling and methods of phylogenetic analysis have generated considerable controversy regarding the efficacy of this approach [Soltis and Soltis 2004, Stephanovic et al. 2004, Martin et al. 2005, Soltis et al. 2004, Leebens-Mack et al. 2005, Lockhart and Penny 2005]. One of the major limitations of this approach is the paucity of chloroplast genome sequences from basal lineages, especially the magnoliids, which are currently represented only by *Calycanthus* [Goremykin et al. 2004].

With four orders, 19 families, and approximately 8,500 species, the magnoliids comprise the largest clade of basal angiosperms [Soltis et al. 2005]. Recent single and multigene phylogenies have provided only weak to moderate support for the monophyly of this clade [Soltis et al. 1999, 2000, Qiu et al. 1999, 2000, Zanis et al. 2002, 2003]. One notable exception is the phylogeny of Qiu et al. [2006] based on eight chloroplast, mitochondrial, and nuclear genes and 162 taxa representing all of the major lineages of gymnosperms and angiosperms. This eight-gene phylogeny provides the first, strong support for the monophyly of the four orders of magnoliids, and for sister group relationships between the Canellales/Piperales and Laurales/Magnoliales.

One of the most important, remaining issues regarding the basal angiosperms is the relationship of magnoliids to the other major clades, including the monocots and eudicots. All three possible relationships among these major angiosperm clades have been generated based on single and multigene phylogenies, but none receive strong support in any of these studies [reviewed in Soltis et al. 2005]. The three-gene phylogeny of Soltis *et al*. [2000] placed the magnoliids and monocots in the same clade and this group was sister to the eudicots. Support for the monophyly of this group, which was referred to as eumagnoliids, was weak with a jackknife value of only 56%. This same topology was generated in Bayesian analyses using the chloroplast gene *matK*, with a parsimony bootstrap value of 78% and a Bayesian posterior probability of 0.73 [Hilu et al. 2003]. Two multi-gene molecular phylogenies identified the magnoliids sister to the eudicots. The 11-gene MP trees in Zanis *et al.* [2002] provided only weak support (56% bootstrap value) for the sister group relationship between magnoliids and eudicots, and in the 9-gene analyses of Qiu *et al.* [2005] support for this same relationship increased to 78% in ML trees. Finally, a third possible resolution of relationships among magnoliids, monocots, and eudicots was recovered in two other studies based on phytochrome genes [Mathews and Donoghue 1999] and 17 chloroplast genes [Graham and Olmstead 2000]. These phylogenies suggested that magnoliids were sister to a clade that included monocots and eudicots, however, support for this relationship had only weak to moderate bootstrap support (< 50% in Mathews and Donoghue [1999] and 76 or 83% in Graham and Olmstead [2000]). Thus, despite intensive efforts during the past 10 years, relationships among magnoliids, monocots, and eudicots remain unresolved.

In this paper, we report on the complete sequences of three magnoliid chloroplast genomes (*Drimys*, *Liriodendron*, and *Piper*). We characterize the organization of two of these genomes, including the most comprehensive comparisons of GC content among completely sequenced chloroplast genomes. Furthermore, the results of phylogenetic analyses of DNA sequences for 61 genes for 35 taxa, including 33 angiosperms and two gymnosperm outgroups provide new evidence for resolving relationships among basal angiosperms with an emphasis on relative positions of magnoliids, monocots and eudicots.

Results

Size, gene content, order and organization of the *Drimys* and *Piper* chloroplast genomes

We have sequenced the chloroplast genomes of three genera of magnoliids, *Drimys*, *Liriodendron*, and *Piper*. In this paper, we only characterize the genome of two of these, *Drimys* and *Piper*, and we use 61 protein-coding genes from all three for the phylogenetic analyses. The genome sequence for *Liriodendron* will be described in another paper on the application of 454 sequencing for chloroplast genomes [Leebens-Mack et al. in progress].

The sizes of the *Drimys* and *Piper* chloroplast genomes are nearly identical at 160,604 and 160,624 bp, respectively (Figs. 1-2). The genomes include a pair of inverted repeats of 26,649 bp (*Drimys*) and 27,039 bp (*Piper*), separated by a small single copy region of 18,621 bp (*Drimys*) and 18,878 bp (*Piper*) and a large single copy region 88,685 bp (*Drimys*) and 87,666 bp (*Piper*). The *Drimys* IR has expanded on the IRa side to duplicate *trnH-gug*. This expansion has not increased the overall size of the IR in *Drimys* because two of the genes in the IR of *Drimys* are shorter than they are in *Piper* (*ycf2* is 6909 and 6945 and *ndhB* is 1533 and 1686 in *Drimys* and *Piper*, respectively).

The *Drimys* and *Piper* chloroplast genomes contain 113 unique genes, and 18 (*Drimys*) and 17 (*Piper*) of these are duplicated in the IR, giving a total of 130-131 genes (Figs. 1-2, Table 1). There are only two differences in gene content between these two magnoliid genomes; one is due to the duplication of *trnH-gug* in *Drimys* and the second is that *ycf1* appears to be a pseudogene in *Piper*, since it has internal stop codons that result in a truncated gene that is only 927 bp long (versus 5,574 bp in *Drimys*). Eighteen genes contain introns, 14 of which contain one intron and three (*clpP*, *rps12*, and *ycf3*) of which contain two introns (Table 1). There are 30 distinct tRNAs, and 8 and 7 of these are duplicated in the IR of *Drimys* and *Piper,* respectively. The genomes consist of 50.12% (*Drimys*) and 48.36% (*Piper*) protein coding genes, 7.38% (*Drimys*) and 7.34% (*Piper*) RNA genes, and 42.5% (*Drimys*) and 44.3% (*Piper*) non-coding regions (intergenic spacers and introns). Gene order is identical in *Drimys* and *Piper* and both genomes have the same gene order as most angiosperms, including tobacco.

GC content

The overall GC content of *Piper* and *Drimys* chloroplast genomes is very similar, 38.31% and 38.79% respectively. These values are within the range of 34-39% GC content but slightly higher than that of the average for 35 chloroplast genomes representing all currently available angiosperms and one gymnosperm *Pinus* (Fig. 3A). GC content is not uniformly distributed across the chloroplast genome (Figs. 3-7). In general, GC content is higher in coding regions than the average GC content for the entire genome and it is lower in non-coding regions (i.e., intergenic spacers and introns) (Fig. 4). This pattern is also supported by the observation that GC content of protein-coding genes is higher than the overall GC content for the complete genomes (compare Figs 3A-B). GC content also varies by codon position with the 1st codon > 2nd codon > 3rd codon (Figs. 5A-B, 6A). GC content was compared by partitioning protein-coding genes into three functional groups (Figs 5A-B, 6A). This comparison demonstrates that the percent of GC for all the three codon positions is highest in photosynthesis genes, followed by genetic system genes, and lowest in NADH genes. Examination of GC content across the chloroplast genomes indicates that GC content is not evenly distributed (Fig. 7), however, the distribution of GC content is similar among all genomes even in taxa that have different gene orders (*i.e.* grasses and legumes). The IR regions have higher GC content and the SSC has the lowest. The much higher GC content in the IR is due to the presence of the rRNA genes (Fig. 6B). The lower GC content in the SSC is due to the presence of eight of the 11 NADH genes, which have a lower GC content than photosynthetic and genetic system genes (Figs. 5A-B, 6A). This genome wide pattern of GC content is maintained even when one copy of the IR is lost (bottom right panel in Fig. 7).

Phylogenetic analyses

Our phylogenetic data set included 61 protein-coding genes for 35 taxa (Table 1), including 33 angiosperms and two gymnosperm outgroups (*Pinus* and *Ginkgo*). The data set comprised 45,879 nucleotide positions but when the gaps were excluded there were 39,378 characters.

Maximum Parsimony (MP) analyses resulted in a single, fully resolved tree with a length of 61,095, a consistency index of 0.41 (excluding uninformative characters) and a retention index of 0.57 (Fig. 8). Bootstrap analyses indicated that 22 of the 32 nodes were supported by values ≥ 95% and 18 of these had a bootstrap value of 100%. Of the remaining 10 nodes, five had bootstrap values between 80-95%. Maximum likelihood (ML) analysis resulted in a single tree with – lnL = 342478.92 (Fig. 9). ML bootstrap values also were also high, with values of ≥ 95% for 28 of the 32 nodes and 100% for 23 nodes. The ML and MP trees had very similar topologies. Both trees indicate that *Amborella* alone forms the earliest diverging angiosperm; however, support for this placement is much higher in MP tree (100%) than the ML tree (63%). The next most basal clade includes the Nymphaeales (*Nuphar* and *Nymphaea*) and support for this relationship is 100% in both MP and ML trees. Recent phylogenies based on complete chloroplast genome sequences [Goremykin et al. 2003, 2004, 2005, Leebens-Mack et al. 2005] have highlighted the difficulty of resolving the relative position of *Amborella* and the Nymphaelales. Two alternative hypotheses that have received the most support are: *Amborella* alone forms the earliest diverging lineage, or *Amborella* and the Nymphaeales form a monophyletic group at the base of angiosperms. In a previous study using 61chloroplast genes (see Fig. 4 in Leebens-Mack et al. 2005) the first hypothesis was strongly supported (100%) in parsimony trees and the second hypothesis received only moderate support (63%) in ML trees. Both MP and ML trees support the basal position of *Amborella* alone in our expanded taxon sampling. We performed a SH test to determine if the *Amborella*/Nymphaeales basal hypothesis is a reasonable alternative to the ML and MP trees that support the *Amborella* basal topology. The ML score for the alternative topology was -ln L = 339758.53918 versus 339758.53918 for the best ML tree. The difference in the -ln L was 5.21714 with a p = 0.303. Thus, the *Amborella*/Nymphaeales basal hypothesis could not be rejected by the SH test, indicating that identification of the most basal angiosperm lineage remains unresolved even with the addition of three magnoliid genomes.

Monophyly of the magnoliids is also strongly supported with 92 (MP) or 100% (ML) bootstrap values. Within magnoliids there a two well-supported clades, one including the Canellales/ Piperales with 74 (MP) or 99 (ML)% bootstrap support, and a second including the Laurales/ Magnoliales with 82 (MP) or 99 (ML)% bootstrap support. The magnoliid clade forms a sister group to a large clade that includes the monocots and eudicots. Support for the sister relationship of magnoliids to the remaining angiosperms is moderate (72% in MP tree, Fig. 8) to strong (99% in ML tree, Fig. 9).

Relationships among most other major angiosperm clades are congruent in the MP and ML trees. Support for the monophyly of monocots and eudicots is strong with 100% bootstrap values, and relationships among the monocots is identical in both analyses. The Ranunculales occupy the earliest diverging lineage among eudicots, and they are sister to two major, strongly supported clades, the rosids and Caryophyllales/asterids. There is strong support for the sister group relationship between the Caryophyllales and asterids (92% in MP and 100% in ML). Within rosids, there is strong support that *Vitis* is the earliest diverging lineage in both MP and ML trees.

The only remaining incongruence between MP and ML trees is found within the rosids. In both analyses, eurosids I are not monophyletic, although support for relationships among the five representatives of this clade to the eurosid II and Myrtales clades is not strong.

Discussion

Genome organization and evolution of GC content

The organization of the *Drimys* and *Piper* genomes with two copies of an IR separating the SSC and LSC regions is identical to most sequenced angiosperm chloroplast genomes [reviewed in Raubeson and Jansen 2005]. The sizes of the genomes at 160,604 and 160,624 bp, respectively are very similar to the each other but substantially larger than the only other sequenced magnoliid genome (*Calycanthus*, 153,337, Goremykin et al. 2003, Table 1). Most of this size increase is due to the larger size of the IR in *Drimys* (26,649 bp) and *Piper* (27,039 bp) relative to *Calycanthus* (23,295 bp), although some is also due to the larger LSC region (Table 1). Expansion and contraction of the IR is a common phenomenon in land plant chloroplast genomes [Goulding et al. 1996] with the IR ranging in size from 9,589 bp in the moss *Physcomitrella* [Sugiura et al. 2003] to 75,741 bp in the highly rearranged angiosperm genome of *Pelargonium* [Palmer et al. 1987, Chumley et al. 2006]. Among angiosperms the IR generally ranges in size between 20-27 kb, and the magnoliid genomes except for *Calycanthus* are at the high end of that range.

Gene order of the magnoliid chloroplast genomes is identical to tobacco and many other unrearranged angiosperm chloroplast genomes. There are a few differences in gene content and these can be explained by two phenomena. The first concerns differences in the annotation two genes in these genomes. Two putative genes (*ACRS* and *ycf15*) in *Calycanthus* were not annotated in *Drimys* and *Piper* due to uncertainty about whether they are functional. These putative genes have been identified in several angiosperm chloroplast genomes but several recent studies raised serious doubts about their functionality. The sequence of *ycf15* has been shown to be highly variable among angiosperm chloroplast genomes, with conserved motifs at the 5' and 3' ends and an intervening sequence that makes it a pseudogene [Schmitz-Linneweber *et al.*, 2001, Steane, 2005]. An examination of *ycf15* transcripts in spinach suggests that this may not be a functional gene [Schmitz-Linneweber et al., 2001]. Although conserved sequences for *ycf15* have also been located in *Drimys* and *Piper*, we decided not to annotate it because of the lack of evidence that they are functional. The ACRS gene was identified by Goremykin et al. [2003] in *Calycanthus* based on its very high sequence identity with the mitochondrial ACR-toxin sensitivity (*ACRS*) gene of *Citrus jambhiri* [Ohtani et al. 2002]. This conserved sequence has been identified (as *ycf68*) in a number of chloroplast genomes, however, there is no evidence that it is a functional gene. The second explanation for gene content differences among the three magnoliid genomes is caused by the expansion of the IR in *Piper*, which results in the duplication of *trnH*. Small expansions of the IR boundary are common in chloroplast genomes [Goulding et al. 1996] resulting in duplications of genes at the IR/SC boundaries. The duplication of *trnH* in *Piper* is interesting because this event has also occurred in *Nuphar*, a member of the Nymphaeales [L. Raubeson et al. unpublished]. This expansion of the IR to duplicate *trnH* has clearly happened independently in *Piper* and *Nuphar* since none of the other basal angiosperms, including *Nymphaea*, *Amborella*, *Drimys*, and *Calycanthus*, have this duplication.

Examination of GC content in 34 seed plant chloroplast genomes reveals several interesting patterns. GC content for the complete genomes ranges between 34-39% (Fig. 3A), confirming previous observations that chloroplast genomes are in general AT rich [Shimada and Sugiura 1991, Maier et al. 1995, Sato et al. 1999, Kato et al. 2000, Goremykin et al. 2003, Kim and Lee 2004, Steane 2005, Daniell et al. 2006]. The uneven distribution of GC content over the chloroplast genome is also very evident, and there are several explanations for this pattern. First, there is a clear bias for the coding regions to have a substantially higher GC content than non-coding regions (Figs. 3-4), which again confirms previous observations based on comparisons of many fewer genomes [Goremykin et al. 2003]. Second, there is an uneven distribution of GC content by regions of the genome with the highest GC content in the IR and the lowest in the SSC (Figs. 4, 7). The higher GC content in the IR can be attributed to the presence of the four rRNA genes in this region, which have the highest GC content of any coding regions (Fig. 6B). This higher GC content in the IR region is maintained even when one copy of the IR is lost as in *Medicago* and *Pinus* (Fig 7, bottom right panel). The lower GC content in the SSC region is due to the presence of 8 of the 11 NADH genes, which have the lowest GC content of any of the classes of genes compared (Figs. 5-6). Third, GC content varies by functional groups of genes. Among protein genes, GC content is highest for photosynthetic genes, lowest for NADH genes, with genetic system genes having intermediate values. This same pattern was observed by Shimada and Sugiura [1991] in comparisons of the first three sequenced land plant chloroplast genomes.

Differences in GC content were also observed by codon position in protein-coding genes (Figs. 3B, 5A,B, 6A). For each of the three classes of genes (photosynthetic, genetic system, and

NADH) the third position in the codon has a substantial AT bias. This pattern has been observed previously [Shimada and Sugiura 1991, Kim and Lee 2004, Liu and Xue 2005], and it has been attributed to codon bias. Previous studies have demonstrated that there is a strong A+T bias in the third codon position for chloroplast genes [Kim and Lee 2004, Liu and Xue 2005]. This is in contrast to a GC bias in codon usage for nuclear genes in plants [Liu and Xue 2005]. Several studies have examined codon usage of chloroplast genes to attempt to determine if these biases can be attributed to nucleotide compositional bias, selection for translational efficiency, or a balance among mutational biases, natural selection, and genetic drift [Morton 1993, 1994, Morton and Levin 1997, Morton 1998, Wall and Herbeck 2003]. All of these studies have been limited to examining a single or few genes, and they have been constrained by the limited sampling of complete genome sequences for taking variation in GC content into account. Our comparisons of GC content variation for a wide diversity of angiosperm lineages provide a rich source of information for future investigations of the relationship between GC content and codon usage bias.

Phylogenetic implications

The debate concerning the identity of the most basal angiosperm lineage continues even though numerous molecular phylogenetic studies of angiosperms have been conducted over the past 15 years [Martin & Dowd 1991, Hamby and Zimmer 1992, Chase et al. 1993, Qiu et al. 1993, 1999, 2000, 2001, 2005, 2006, Soltis et al. 1997, 2000, Hoot et al. 1999, Mathews and Donoghue 1999, 2000, Parkinson et al. 1999, Soltis et al. 1999, Barkman et al. 2000, Graham and Olmstead 2000, Savolainen et al. 2000, Zanis et al. 2002, 2003, Borsch et al. 2003, Goremykin et al. 2003, 2004, 2005, Hilu et al. 2003, Aoki et al. 2004, Kim et al. 2004, Stefanovic et al. 2004; Leebens-Mack et al. 2005, Löhne and Borsch 2005]. Several issues have confounded the resolution of relationships among basal angiosperms, including long branch attraction, taxon sampling, and phylogenetic methodology [Barkman et al. 2000, Graham and Olmstead 2000, Zanis et al. 2002, Stefanovic et al. 2004, Soltis et al. 2004, Leebens-Mack et al. 2005, Goremykin et al. 2005, Jansen et al. 2006]. One consensus that is emerging from the most recent studies is that *Amborella* and the Nymphaeales represent earliest diverging angiosperm lineages [Leebens-Mack et al. 2005, Qiu et al 2006]. Some recent molecular phylogenies based on a single or a few gene sequences have provided moderate to strong support for the placement of *Amborella* alone as the earliest diverging angiosperm lineage, whereas other phylogenies have suggested that *Amborella* + Nymphaeales form a sister group at the base of the angiosperms tree. The most recent multi-gene phylogenies based on nine gene sequences from the chloroplast, mitochondrial, and nuclear genomes [Qiu et al. 2006] generate trees supporting each of these two hypotheses depending on the method of phylogenetic analysis and the genes included. Phylogenies generated from chloroplast genes supported the *Amborella* basal hypothesis, whereas mitochondrial genes supported the *Amborella* + Nymphaeales hypothesis. Furthermore, MP analyses tended to support the *Amborella* basal hypothesis and ML analyses supported *Amborella* + Nymphaeales. A similar set of relationships was also observed in recent phylogenetic studies using sequences of 61 genes from completely sequenced chloroplast genomes [Leebens-Mack et al. 2005, Jansen et al. 2006]. In these studies, MP trees placed *Amborella* alone as the basal most angiosperm with strong support and ML trees placed *Amborella* + Nymphaeales at the base with moderate support. These differences were attributed to limited taxon sampling and long branch attraction.

Our phylogenetic analyses include three additional basal angiosperms representing three different orders of magnoliids. Both MP and ML trees (Figs. 8-9) support *Amborella* alone as the earliest diverging lineage of angiosperms. Support for this relationship is very strong in MP trees (100% bootstrap) and weak (63%) in ML trees. However, a SH test that constrained *Amborella* + Nymphaeales in a basal position indicated that the two hypotheses of basal angiosperm relationships are not significantly different. Thus, although both MP and ML analyses including the three additional magnoliid taxa support the *Amborella* alone hypothesis, sampling of more taxa and genes is needed to resolve this issue.

Several earlier molecular phylogenetic studies based on one or a few genes [Chase et al. 1993, Savolainen et al. 2000, Soltis et al. 2000] did not support the monophyly of magnoliids. Furthermore, morphological studies of angiosperms failed to detect any synapomorphies for this group. The circumscription, monophyly, and relationships of magnoliids has only recently been established based on multigene molecular phylogenies [Qiu et al. 1999, 2000]. These earlier multigene phylogenies provided only weak to moderate support for the monophyly of magnoliids and the sister group relationships of the Canellales/Piperales and Laurales/Magnoliales. A recent study using eight chloroplast, mitochondrial, and nuclear genes [Qiu et al. 2006] provided the first strong support for both the monophyly and relationships among the four orders of magnoliids. Our phylogeny based on 61 chloroplast protein-coding genes also provide very strong support for the monophyly of magnoliids and the sister relationship between the Canellales/Piperales and Laurales/Magnoliales.

One of the most controversial remaining issues regarding relationships among angiosperms concerns the resolution of relationships among the magnoliids, monocots and eudicots. Previous phylogenetic studies have supported three different hypotheses of relationships among these lineages: (1) (magnoliids (monocots, eudicots)), (2) (monocots (magnoliids, eudicots)), and (3) (eudicots (magnoliids, monocots)). The first hypothesis was supported in phylogenies based on phytochrome genes [Mathews and Donoghue 1999] and 17 chloroplast genes [Graham and Olmstead] but bootstrap support for a sister relationship of monocots and eudicots was only 67%. Several studies supported the second hypothesis [Nickrent et al. 2002, Zanis et al. 2002, Qiu et al. 2005], however, bootstrap support was again weak ranging from 55 - 78%. The three-gene phylogeny of Soltis et al. [2000] supported the third hypothesis with only 56% jackknife support. This relationship was also recovered in a *matK* phylogeny with a parsimony bootstrap value of 78% and a posterior probability of 0.73 [Hilu et al. 2003]. Both MP and ML phylogenies based on 61 chloroplast-encoded protein genes support hypothesis 1 (Figs. 8-9). Branch support for this hypothesis is moderate (MP, Fig. 8) to strong (ML, Fig. 9). Congruence of the results from both MP and ML analyses is notable because our previous molecular phylogenies using whole chloroplast genomes that included only one member of the magnoliid clade (*Calycanthus*, [Leebens-Mack et al. 2005, Jansen et al. 2006]) were incongruent. In these earlier studies, MP trees supported hypothesis 2 (monocots sister to a clade that included magnoliids and eudicots), whereas ML trees supported hypothesis 1 (magnoliids sister to a clade that included monocots and dicots). These differences provide yet another example of the importance of expanded taxon sampling in phylogenetic studies using sequences from whole chloroplast genomes [Leebens-Mack et al. 2005, Jansen et al. 2006]. The addition of other basal angiosperm lineages, especially members of the Chloranthales, Certatophyllaceae, and Illiciales will be critical for providing additional resolution of relationships among the major clades.

Methods

Chloroplast isolation, amplification, and sequencing

10 -20 g of fresh leaf material of *Drimys granatenis* and *Piper coenoclatum* was used for the chloroplast isolation. Leaf material was obtained from the University of Connecticut Greenhouses (accession numbers 200100052 for *Drimys* and 199600027 for *Piper*). Chloroplasts were isolated from fresh leaves using the sucrose-gradient method [Palmer 1986]. They were then lysed and the entire chloroplast genome was amplified using Rolling Circular Amplification (RCA, using the REPLI-g™ whole genome amplification kit, Molecular Staging) following the methods outlined in Jansen *et al.* [2005]. The RCA product was then digested with the restriction enzymes *Eco*RI and *Bst*BI and the resulting fragments were separated by agarose gel electrophoresis to determine the quality of chloroplast DNA. The RCA product was sheared by serial passage through a narrow aperture using a Hydroshear device (Gene Machines), and the resulting fragments were enzymatically repaired to blunt ends and gel purified, then ligated into pUC18 plasmids. The clones were introduced into *E. coli* by electroporation, plated onto nutrient agar with antibiotic selection, and grown overnight. Colonies were randomly selected and robotically processed through RCA of plasmid clones, sequencing reactions using BigDye chemistry (Applied Biosystems), reaction cleanup using solid-phase reversible immobilization, and sequencing determination using an ABI 3730 XL automated DNA sequencer. Detailed protocols are available at http://www.jgi.doe.gov/sequencing/protocols/protsproduction.html.

Genome assembly and annotation

Sequences from randomly chosen clones were processed using PHRED and assembled based on overlapping sequence into a draft genome sequence using PHRAP [Ewing and Green 1998]. Quality of the sequence and assembly was verified using Consed [Gordon et al. 1998]. In most regions of the genomes we had 6-12-fold coverage but there were a few areas with gaps or low depth of coverage. PCR and sequencing at the University of Texas at Austin were used to bridge gaps and fill in areas of low coverage in the genome. Additional sequences were added until a completely contiguous consensus was created representing the entire chloroplast genome with a minimum of 2X coverage and a consensus quality score of Q40 or greater.

Genome Annotation

The coordinate of each genome was standardized for gene annotation to be the first bp after IRa on the *psbA* side. The genomes of *Piper* and *Drimys* were annotated using the program DOGMA (Dual Organellar GenoMe Annotator, Wyman et al. 2004). All genes, rRNAs, and tRNAs were identified using the plastid/bacterial genetic code.

Examination of GC content

GC content was calculated for 34 seed plant chloroplast genomes, including the gymnosperm *Pinus* and 33 angiosperms. GC content was also determined for 66 protein-coding genes. These genes were partitioned into three functional groups (photosynthesis (33), genetic system genes (22), and NADH (11) genes), and GC content was calculated for the entire gene and the first, second, and third codon positions. The genes included in each of these three groups are: (1) photosynthetic genes (*atpA*, *atpB*, *atpE*, *atpF*, *atpH*, *atpI*, *psbZ*, *petA*, *petB*, *petD*, *petG*, *petL*, *petN*, *psaA*, *psaB*, *psaC*, *psaI*, *psaJ*, *psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *rbcL*), genetic system genes (*rpl14*, *rpl16*, *rpl2*, *rpl20*, *rpl32*, *rpl33*, *rpl36*, *rpoA*, *rpoB*, *rpoC1*, *rpoC2*, *rps11*, *rps12*, *rps14*, *rps15*, *rps18*, *rps19*, *rps2*, *rps3*, *rps4*, *rps7*, *rps8*), and NADH genes (*ndhA*, *ndhB*, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*). GC content was also plotted over the entire genome for all 34 taxa, which were classified into 10 groups based on gene order and phylogenetic placement (Figs. 8-9).

Phylogenetic analysis

Alignment

The 61 protein-coding genes included in the analyses of Goremykin et al. [2003, 2004] and Leebens-Mack et al. [2005] were extracted from *Drimys*. *Liriodendron* and *Piper* using the organellar genome annotation program DOGMA [Wyman et al. 2004]. The same set of 61 genes was extracted from chloroplast genome sequences of 32 other sequenced chloroplast genomes (see Table 1 for complete list of genomes examined). All 61 protein-coding genes of the 34 taxa were translated into amino acid sequences, which were aligned using MUSCLE [Edgar 2004] followed by manual adjustments, and then nucleotide sequences of these genes were aligned by constraining them to the aligned amino acid sequences. A Nexus file with character sets for phylogenetic analyses was generated after nucleotide sequence alignment was completed.

Tree reconstruction

Phylogenetic analyses using maximum parsimony (MP) and maximum likelihood (ML) were performed using PAUP* version 4.10 [Swofford 2003] on a data including 34 taxa (Table 1). Phylogenetic analyses excluded gap regions. All MP searches included 100 random addition replicates and TBR branch swapping with the Multrees option. Modeltest 3.7 [Posada and Crandall 1998] was used to determine the most appropriate model of DNA sequence evolution for the combined 61-gene dataset. Hierarchical likelihood ratio tests and the Akaikle information criterion were used to assess which of the 56 models best fit the data, which was determined to be GTR + I + Γ by both criteria. For ML analyses we performed an initial parsimony search with 100 random addition sequence replicates and TBR branch swapping, which resulted in a single tree. Model parameters were optimized onto the parsimony tree. We fixed these parameters and performed a ML analysis with three random addition sequence replicates and TBR branch swapping. The resulting ML tree was used to re-optimize model parameters, which then were fixed for another ML search with three random addition sequence replicates and TBR branch swapping. This successive approximation procedure was repeated until the same tree topology and model parameters were recovered in multiple, consecutive iterations. This tree was accepted as the final ML tree (Fig. 8). Successive approximation has been shown to perform as well as full-optimization analyses for a number of empirical and simulated datasets [Sullivan et al. 2005]. Non-parametric bootstrap analyses [Felsenstein 1985] were performed for MP analyses with 1000 replicates with TBR branch swapping, 1 random addition replicate, and the Multrees option and for ML analyses with 100 replicates with NNI branch swapping, 1 random addition replicate, and the Multrees option.

Test of alternate topology

A Shimodaira-Hasegawa (SH) test [Shimodaira and Hasegawa 1999] was performed to determine if the alternative topology with *Amborella* + Nymphaeales basal was significantly worse than the ML tree that places *Amborella* alone as the basal angiosperm lineage. A constraint topology with this alternative tree topology was used and the SH test was conducted using RELL optimization [Goldman et al. 2000] as implemented in PAUP* version 4.10 [Swofford 2003].

Abbreviations

IR inverted repeat; SSC, small single copy; LSC, large single copy, bp, base pair; ycf, hypothetical chloroplast reading frame; rrn, ribosomal RNA; MP, maximum parsimony; ML, maximum likelihood.

Authors' contributions

CZ finished and annotated the *Drimys* and *Piper* chloroplast genomes, wrote perl scripts to automate the alignment of both amino acid and nucleotide sequences, performed phylogenetic analyses and comparisons of GC content, and wrote several sections of the manuscript; CP assisted in chloroplast isolations and rolling circular amplification of the *Drimys* and *Piper* genomes, assisted in finishing the genome sequences, and extracting gene sequences; JVK and JLB generated libraries, performed the draft genome sequencing and assembly; RKJ assisted with chloroplast isolations, finishing, and annotation of the *Drimys* and *Piper* genomes, assisted in phylogenetic analyses, and wrote several portions of the manuscript; JLM, JC, and CWP sequenced, annotated, and extracted gene sequences from the *Liriodendron* chloroplast genome. All authors have read and approved the final manuscript.

Acknowledgements

We thank Clinton Morse from the University of Connecticut greenhouses for providing leaf material of *Drimys* and *Piper*. We also thank Andrew Alverson for assistance with the S-H tests.

This research was supported in part by a grant from the National Science Foundation (DEB 0120709) and the Sidney F. and Doris Blake Centennial Professorship in Systematic Botany to RKJ. Part of this work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Berkeley National Laboratory, under contract No. DE-AC02-05CH11231.

References

Darwin, C. **Letter to J. D. Hooker**. In More letters of Charles Darwin, Edited by Darwin F Seward AC, vol. 2, London: John Murray; 1903.

Soltis, DE, Soltis PS: ***Amborella* not a "basal angiosperm"? Not so fast**. *Amer J Bot* 2004, **91**:997-1001.

Stefanovic S, Rice DW, Palmer JD: **Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots?** *BMC Evol Biol* 2004, **4**:35.

Martin W, Deusch O, Stawski N, Grunheit N, Goremykin V: **Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution**. *Trends Plant Sci* 2005, **10**:203-209.

Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu Y-Q, Chase MW, Farris JS, Stefanović S, Rice DW, Palmer JD, Soltis PS: **Genome-scale data, angiosperm relationships, and ‘ending incongruence’: a cautionary tale in phylogenetics**. *Trends Plant Sci.* 2004, **9**:477–483.

Lockhart PJ, Penny D: **The place of *Amborella* within the radiation of angiosperms**. *Trends Plant Sci* 2005, **10**:201-202.

Donoghue D J, Mathews S: **Duplicate genes and the root of angiosperms, with an example using phytochrome sequences**. *Mol Phylogen Evol* 1998, **9**:489-500.

Barkman TJ, Chenery G, McNeal JR, Lyons-Weiler J, Ellisens WJ, Moore G, Wolfe AD, dePamphilis CW: **Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny**. *Proc Natl Acad Sci USA* 2000, **97**:13166-13171.

Borsch T, Hilu KW, Quandt DV, Wilde Neinhuis C, Barthlott W: **Noncoding plastid trnT-*trnF* sequences reveal a well-resolved phylogeny of basal angiosperms**. *J Evol Biol* 2003, **16**:558-576.

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

Qiu Y-L, Li L, Hendry T, Li R, Taylor DW, Issa MJ, Ronen AJ, Vekaria ML, White AM: **Reconstructing the Basal Angiosperm Phylogeny: Evaluating Information Content of the Mitochondrial Genes**. *Taxon* 2006, in press.

Soltis DE, Soltis PS, Endress PK, Chase MW: *Phylogeny and evolution of Angiosperms*. Sunderland Massachusetts: Sinauer Associates Inc.: 2005.

Soltis PS, Soltis DE, Chase MW: Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* 1999, **402**:402-404.

Soltis DE, Soltis PS, Chase MW, Mort ME, Albach DC, Zanis M, Savolainen V, Hahn WJ, Hoot SB, Fay MF, Axtell M, Swensen SM, Prince LM, Kress WJ, Nixon KC, Farris JS: **Angiosperm phylogeny inferred from 18S rDNA, *rbcL,* and *atpB* sequences**. *Bot J Linn Soc* 2000, **133**:381-461.

Qiu Y-L, Lee J, Whitlock BA, Bernasconi-Quadroni F, Dombrovska O: **Was the ANITA rooting of the angiosperm phylogeny affected by long branch attraction?** *Mol Biol Evol* 2001, **18**:1745-1753.

Hoot SB, Magallon S, Crane PR: **Phylogeny of basal tricolpates based on three molecular data sets: *atpB, rbcL,* and 18S nuclear ribosomal DNA sequences**. *Ann Missouri Bot Gard* 1999, **86**:1-32.

Zanis MJ, Soltis PS, Qiu Y-L, Zimmer EA, Soltis DE: **Phylogenetic analyses and perianth evolution in basal angiosperms**. *Ann Missouri Bot Gard* 2003, **90**:129-150.

Doyle JA, Endress PK: **Morphological phylogenetic analysis of basal angiosperms: Comparison and combination with molecular data**. *Int J Plt Sci* 2000, **161**: S121-S153.

Qiu Y-L, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainen V, Chase MW: **Phylogeny of basal angiosperms: analyses of five genes from three genomes**. *Int J Plt Sci* 2000, **161**:S3-S27.

Hamby RK, Zimmer EA: **Ribosomal RNA as a phylogenetic tool in plant systematics**. In *Molecular Systematics of Plants.* Edited by: Soltis PS, Soltis DE, Doyle JJ. Chapman and Hall; 1992: 50-91.

Soltis DE, Soltis PS, Nickrent DL, Johnson LA, Hahn WJ, Hoot SB, Sweere JA, Kuzoff RK, Kron KA, Chase M, Swensen SM, Zimmer E, Shaw SM, Gillespie LJ, Kress WJ, Sytsma K: **Angiosperm phylogeny inferred from 18S ribosomal DNA sequences**. *Ann Missouri Bot Gard* 1997, 84:1-49.

Qiu Y-L, Chase MW, Les DH, Parks CR: **Molecular phylogenetics of the Magnoliidae: cladistic analyses of nucleotide sequences of the plastid gene *rbcL***. *Ann Missouri Bot Gard* 1993, **80**:587-606.

Savolainen V, Chase MW, Morton CM, Soltis DE, Bayer C, Fay MF, De Bruijn A, Sullivan S, Qiu Y-L: **Phylogenetics of flowering plants based upon a combined analysis of plastid *atpB* and *rbcL* gene sequences**. *Syst Biol* 2000, **49**:306-362.

Hilu KW, Borsch T, Muller K, Soltis DE, Soltis PS, Savolainen V, Chase M, Powell M, Alice L, Evans R, Sauquet H, Neinhuis C, Slotta T, Rohwer J, Chatrou L**: Inference of angiosperm phylogeny based on *matK* sequence information**. *Amer J Bot* 2003, **90**:1758-1776.

Chase M, Soltis D, Olmstead R, Morgan D, Les D, Mishler B, Duvall M, Price R, Hills H, Qui Y-L, Kron K, Rettig J, Conti E, Palmer J, Manhart J, Sytsma K, Michaels H, Kress J, Karol K, Clark D, Hedren M, Gaut B, Jansen R, Kim K-J, Wimpee C, Smith J, Furnier G, Straus S, Xiang Q-Y, Plunkett G, Soltis P, Swensen S, Williams S, Gadek P, Quinn C, Equiarte L, Golenberg E, Learn G, Graham S, Barrett S, Dayanandan S, Albert V: **Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbc*L**. *Ann Missouri Bot Gard* 1993, **80**:528-580.

Raubeson LA, Jansen RK: **Chloroplast genomes of plants**. In *Diversity and Evolution of Plants-Genotypic and Phenotypic Variation in Higher Plants*. Edited by: Henry H. Wallingford: CABI Publishing; 2005:45–68.

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: **Phylogenetic analysis of basal angiosperms based on nine plastid, mitochondrial, and nuclear genes**. *Int J Plt Sci* 2005, **166**: 815-842.

Martin PG, Dowd JM: **Studies of angiosperm phylogeny using protein sequences**. *Ann Misssouri Bot Gard* 1991, **78**:296-337.

Davies TJ, Barraclough TG, Chase MW, Soltis PS, Soltis DE, Savolainen V: **Darwin's abominable mystery: insights from a supertree of the angiosperms.** *Proc Natl Acad. Sci USA* 2004, **101**:1904–1909.

Zanis MJ, Soltis DE, Soltis PS, Mathews S, Donoghue MJ: **The root of the angiosperms revisited**. *Proc Natl Acad Sci* USA 2002, **99**:6848-6853.

Graham SW, Olmstead RG: **Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms**. *Am J Bot*, 2000, 87:1712-1730.

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

Mathews S, Donoghue MJ: **Basal angiosperm phylogeny inferred from duplicated phytochromes A and C**. *Int J Plt Sci* 2000, **161**:S41-S55.

Parkinson CL, Adams KL and Palmer JD: **Multigene analyses identify the three earliest lineages of extant flowering plants**. *Curr Biol* 1999, **9**:1485-1488.

Aoki S, Uehara K, Imafuku M, Hasebe M, Ito M: **Phylogeny and divergence of basal angiosperms inferred from *APETALA3*- and *PISTILLATA*-like MADS-box genes**. *J Plt Res* 2004, **117**:229 – 244.

Kim ST, Yoo MJ, Albert VA, Farris JS, Soltis PS, Soltis DE: **Phylogeny and diversification of B-function MADS-box genes in angiosperms: Evolutionary and functional implications of a 260-million-year-old duplication**. *Amer J Bot* 2004, 91:2102-2118.

Löhne C, Borsch T: **Molecular evolution and phylogenetic utility of the *petD* group II intron: a case study in basal angiosperms**. *Mol Biol Evol* 2005, **22**:317-332.

Nickrent DL, Blarer A, Qiu Y-L, Soltis, DE, Soltis PS, Zanis M: **Molecular data place Hydnoraceae with Aristolochiaceae**. *Amer J Bot* 2002, **89**:1809-1817.

Wyman SK, Boore JL, Jansen RK: **Automatic annotation of organellar genomes with DOGMA**. *Bioinformatics* 2004, **20**:3252–3255 [http://evogen.jgi.psf.org/dogma].

Higgins DG, Thompson JD, Gibson TJ: **Using CLUSTAL for multiple sequence alignments**. *Meth Enzy* 1996, **266**:383–402.

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

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH: **Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm**. *Mol Biol Evol* 2003, **20**:1499-1505.

Leebens-Mack J, Raubeson LA, Cui L, Kuehl J, Fourcade M, Chumley T, Boore JL, Jansen RK, and dePamphilis CW: **Identifying the basal angiosperms in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone**. *Mol Biol Evol* 2005, **22**:1948-1963.

Goremykin VV, Holland B, Hirsch-Ernst KI, Hellwig FH: **Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications**. *Mol Biol Evol* 2005, 22:1813-1822.

Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucl Acids Res* 2004, **32**: 1792-1797.

Shimodaira H, Hasegawa M: **Multiple comparisons of log-likelihoods with applications to phylogenetic inference**. *Mol Biol Evol* 1999, **16**: 1114-1116.

Palmer JD: **Isolation and structural analysis of chloroplast DNA**. In: *Methods Enzymol.* Edited by Weissbach A, Weissbach H, vol. 118. New York: Academic Press; 1986: 167-186.

Jansen RK, Raubeson LA, Boore JL, dePamphilis CW, Chumley TW, Haberle RC, Wyman SK, Alverson AJ, Peery R, Herman SJ, Fourcade HM, Kuehl JV, McNeal JR, Leebens-Mack J, Cui L: **Methods for obtaining and analyzing chloroplast genome sequences**. *Meth Enzym* 2005, **395**:348 - 384.

Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities**. *Genome Res* 1998, **8**:186-194.

Gordon D, Abajian C, Green P: **Consed: a graphical tool for sequence finishing**. *Genome Res* 1998, **8**:195-202.

Swofford DL: PAUP*: *Phylogenetic analysis using parsimony (*and other methods)*, ver. 4.0 Sunderland MA: Sinauer Associates; 2003.

Goldman N, Anderson JP, Rodrigo AG: **Likelihood-based tests of topologies in phylogenetics**. *Syst Biol* 2000, **49**: 652-670.

Posada D., Crandall KA: **MODELTEST: testing the model of DNA substitution**. *Bioinformatics* 1998, **14**:817–818.

Sullivan J, Abdo Z, Joyce P, Swofford DL: **Evaluating the performance of a successive-approximations approach to parameter optimization in maximum-likelihood phylogeny estimation**. *Mol Biol Evol* 2005, **22**:1386-1392.

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

Wakasugi T, Tsudzuki J, Ito S, Nakashima K, Tsudzuki T, Sugiura M: **Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii***. *Proc Natl Acad Sci USA* 1994, **91**:9794-9798.

Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun CR, Meng BY, Li YQ, Kanno A, Nishizawa Y, Hirai A, Shinozaki K, Sugiura M: **The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals**. *Mol Gen Genet* 1989, **217**:185-194.

Asano T, Tsudzuki T, Takahashi S, Shimada H, Kadowaki K: **Complete nucleotide sequence of the sugarcane (*Saccharum officinarum*) chloroplast genome: a comparative analysis of four monocot chloroplast genomes**. *DNA Res* 2004, **11**:93-99.

Maier RM, Neckermann K, Igloi GL, Kossel H: **Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing**. *J Mol Biol* 1995, **251**:614–628.

Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S: **Complete structure of the chloroplast genome of *Arabidopsis thaliana***. *DNA Res* 1999, **6**:283–290.

Steane DA: **Complete nucleotide sequence of the chloroplast genome from the Tasmanian blue gum, *Eucalyptus globulus* (Myrtaceae)**. *DNA Res* 2005, **12**:215-220.

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H, Sugiura M: **The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression**. *EMBO J* 1986, **5**:2043-2049.

Kim K-J, Lee H-L: **Complete chloroplast genome sequence from Korean Ginseng (*Panax schiseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants**. *DNA Res* 2004, **11**:247-261.

Daniell H, Lee S-B, Grevich J, Saksi C, Quesada-Vargas T, Guda C, Tomkins J, Jansen RK: **Complete chloroplast genome sequences of *Solanum*** *bulbocastanum, Solanum lycopersicum* and comparative analyses with other Solanaceae genomes. *Theor Appl Genet* 2006 **112**:1503-1518.

Schmitz-Linneweber C, Maier RM, Alcaraz JP, Cottet A, Herrmann RG, Mache R: **The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization**. *Plt Mol Biol* 2001, **45**:307–315.

Chang C-C, Lin H-C, Lin I-P, Chow T-Y, Chen H-H, Chen W-H, Cheng C-H, Lin C-Y, Liu S-M, Chang C-C, Chaw S-M: **The chloroplast genome of *Phalaenopsis Aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications**. *Mol Biol Evol* 2006, **23**:279-291.

Lee S-B, Kaittanis C, Jansen RK, Hostetler JB, Tallon LJ, Town CD, Daniell H: **The complete chloroplast genome sequence of *Gossypium hirsutum*: organization and phylogenetic relationships to other angiosperms**. *BMC Genomics* 2006, **7**: 61.

Hupfer H, Swaitek M, Hornung S, Herrmann RG, Maier RM, Chiu WL Sears B: **Complete nucleotide sequence of the *Oenothera elata* plastid chromosome, representing plastome 1 of the five distinguishable *Euoenthera* plastomes**. *Mol Gen Genet* 2000, **263**:581–585.

Schmitz-Linneweber C, Regel R, Du TG, Hupfer H, Herrmann RG, Maier RM: **The plastid chromosome of *Atropa belladonna* and its comparison with that of *Nicotiana tabacum*: the role of RNA editing in generating divergence in the process of plant speciation**. *Mol Biol Evol* 2002, **19**:1602-1612.

APG II 2002. **An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II**. *Bot J Linn Soc* 2003, **141**:399-436.

Jansen RK, Kaittanis C, Saski C, Lee SB, Tomkins J, Alverson AJ, Daniell H: **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* 2006, **6**:32.

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH: **The chloroplast genome of the "basal" angiosperm *Calycanthus fertilis* - structural and phylogenetic analyses**. *Plt Syst Evol* 2003, **242**:119-135.

Goulding SE, Olmstead RG, Morden CW, Wolfe KH: **Ebb and flow of the chloroplast inverted repeat**. *Mol Gen Gen* 1996, **252**: 195–206.

Sugiura C, Kobayashi Y, Aoki S, Sugita C, Sugita M: **Complete chloroplast DNA sequence of the moss *Physcomitrella patens*: evidence for the loss and relocation of *rpoA* from the chloroplast to the nucleus**. Nucl Acids Res 2003, **31**:5324-5331.

Palmer JD, Nugent JM, Herbon LA: **Unusual structure of geranium chloroplast DNA - A triple-sized inverted repeat, extensive gene duplications, multiple inversions, and 2 repeat families**. Proc Natl Acad Sci USA 1987, **84**:769-773.

Chumley TW, Palmer JD, Mower JP, Fourcade HM, Caile PJ, Boore JL, Jansen RK: **The complete chloroplast genome sequence of *Pelargonium* × *hortorum*: Organization and evolution of the largest and most highly rearranged chloroplast genome of land plants**. *Mol Biol Evol*, in review.

Ohtani K, Yamamoto H, Akimitsu K: **Sensitivity to *Alternaria alternata* toxin in citrus because of altered mitochondrial RNA processing**. *Proc Natl Acad Sci USA* 2002, **99**:2439-2444.

Shimada H, Sugiura M: **Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes**. *Nucl Acids Res* 1991, **19**:983-995.

Kato T, Kaneko T, Sato S, Nakamura Y, Tabata S: **Complete structure of the chloroplast genome of a legume, *Lotus japonicus***. *DNA Res* 2000, **7**:323–330.

Liu Q, Xue Q: **Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species**. *J Genet* 2005, **84**:55-62.

Morton BR: **Chloroplast DNA codon use: Evidence for selection at the *psbA* locus based on tRNA availability**. *J Mol Evol* 1993, **37**:273-280.

Morton BR: **Codon use and the rate of divergence of land plant chloroplast genes**. *Mol Biol Evol* 1994, **11**:231-238.

Morton BR: **Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages**. *J Mol Evol* 1998, **46**:449-459.

Morton BR, Levin JA: **The atypical codon usage of the plant *psbA* gene maybe the remnant of an ancestral bias**. *Proc Natl Acad Sci USA* 1997, **94**:11434-11438.

Wall DP, Herbeck JT: **Evolutionary patterns of codon usage in the chloroplast gene *rbcL***. *J Mol Evol* 2003, **56**:673-688.

Table 1. Comparison of major features of magnoliid chloroplast genomes.

	Calycanthus	*Drimys*	*Piper*
Size (bp)	153,337	160,604	160,624
LSC length (bp)	86,948	88,685	87,666
SSC length (bp)	19,799	18,621	18,878
IR length (bp)	23,295	26,649	27,039
Number of genes	133 (115)	131 (113)	130 (113)
Number of gene duplicated in IR	18	18	17
Number of genes with introns (with 2 introns)	18 (3)	18 (3)	18 (3)

Table 2. Taxa included in phylogenetic analyses with GenBank accession numbers and references.

Taxon	GenBank Accession Numbers	Reference
Gymnosperms – Outgroups		
Pinus thunbergii	NC_001631	Wakasugi et al. 1994 []
Ginkgo biloba	DQ069337-DQ069702	Leebens-Mack et al 2005 []
Basal Angiosperms		
Amborella trichopoda	NC_005086	Goremykin et al. 2003 []
Nuphar advena	DQ069337-DQ069702	Leebens-Mack et al 2005 []
Nymphaea alba	NC_006050	Goremykin et al. 2004 []
Magnoliids		
Calycanthus floridus	NC_004993	Goremykin et al. 2003 []
Drimys granatensis	To be submitted to GenBank	Current study
Liriodendron tulipifera	To be submitted to GenBank	Current study
Piper coenoclatum	To be submitted to GenBank	Current study

Monocots		
Acorus americanus	DQ069337-DQ069702	Leebens-Mack et al 2005 []
Oryza sativa	NC_001320	Hiratsuka et al. 1989 []
Saccharum officinarum	NC_006084	Asano et al. 2004 []
Triticum aestivum	NC_002762	Ikeo and Ogihara, unpublished
Phaleanopsis aphrodite	AY916449	Chang et al. 2006 []
Typha latifolia	DQ069337-DQ069702	Leebens-Mack et al 2005 []
Yucca schidigera	DQ069337-DQ069702	Leebens-Mack et al 2005 []
Zea mays	NC_001666	Maier et al. 1995 []
Eudicots		
Arabidopsis thaliana	NC_000932	Sato et al. 1999 []
Atropa belladonna	NC_004561	Schmitz-

		Linneweber et al. 2002 []
Citrus sinensis	XXXXXXX	Bausher et al. unpublished
Cucumis sativus	NC_007144	Plader et al. unpublished
Eucalyptus globulus	AY780259	Steane 2005 []
Glycine max	DQ317523	Saski et al. 2005 []
Gossypium hirsutum	DQ345959	Lee et al. 2006 []
Lotus corniculatus	NC_002694	Kato et al. 2000 []
Medicago truncatula	NC_003119	Lin et al., unpublished
Nicotiana tabacum	NC_001879	Shinozaki et al. 1986 []
Oenothera elata	NC_002693	Hupfer et al. 2000 []
Panax schinseng	NC_006290	Kim and Lee 2004 []
Populus trichocarpa	http://genome.ornl.gov/poplar_chloroplast/	unpublished
Ranunculus	DQ069337-DQ069702	Leebens-Mack et

macranthus		al 2005 []
Solanum lycopersicum	DQ347959	Daniell et al. 2006 []
Solanum bulbocastanum	DQ347958	Daniell et al. 2006 []
Spinacia oleracea	NC_002202	Schmitz-Linneweber et al. 2001 []
Vitis vinifera	DQ424856	Jansen et al. 2006 []

Figure 1. Gene map of the *Drimys granatensis* chloroplast genome. The thick lines indicate the extent of the inverted repeats (IRa and IRb), which separate the genome into small (SSC) and large (LSC) single copy regions. Genes on the outside of the map are transcribed in the clockwise direction and genes on the inside of the map are transcribed in the counterclockwise direction.

Figure 2. Gene map of the *Piper coenoclatum* chloroplast genome. The thick lines indicate the extent of the inverted repeats (IRa and IRb), which separate the genome into small (SSC) and large (LSC) single copy regions. Genes on the outside of the map are transcribed in the clockwise direction and genes on the inside of the map are transcribed in the counterclockwise direction.

Figure 3. Histogram of GC content for 34 seed plant chloroplast genomes, including the gymnosperm *Pinus* and 33 angiosperms (see Table 2 for list of genomes). **A.** Overall GC content of complete genomes. **B.** GC content for 66 protein-coding genes, including average value for all codon positions, followed by values for the 1st, 2nd, and 3rd codon positions, respectively.

Figure 4. Graphs of GC content plotted over the entire chloroplast genomes of *Drimys* and *Piper*. X axis represents the proportion of GC content between 0 and 1 and the Y axis gives the coordinates in kb for the genomes. Coding and non-coding regions are indicated in blue and red, respectively. The green dashed line indicates that average GC content for the entire genome.

Figure 5. Histogram of GC content for photosynthetic and genetic system genes for 34 seed plant chloroplast genomes (see Table 2 for list of genomes). GC content includes average value for all codon positions, followed by values for the 1st, 2nd, and 3rd codon positions, respectively. **A.** GC content for 33 photosynthetic genes. **B.** GC content for 22 genetic system genes.

Figure 6. Histogram of GC content for NADH and rRNA genes for 34 seed plant chloroplast genomes (see Table 2 for list of genomes). **A.** GC content for 33 photosynthetic genes, which includes average value for all codon positions, followed by values for the 1st, 2nd, and 3rd codon positions, respectively. **B.** GC content for four rRNA genes.

Figure 7. Graphs of GC content plotted over the entire chloroplast genomes of 34 seed plants. The graphs are organized by genomes with the same gene order and by clade in the phylogenies in Figures 8 and 9. X axis represents the proportion of GC content between 0 and 1 and the Y axis gives the coordinates in kb for the genomes.

Figure 8. Phylogenetic tree of 35-taxon data set based on 61 chloroplast protein-coding genes using maximum parsimony. The tree has a length of 61,095, a consistency index of 0.41 (excluding uninformative characters) and a retention index of 0.57. Numbers at each node are bootstrap support values. Numbers above node indicate number of changes along each branch and numbers below nodes are bootstrap support values. Ordinal and higher level group names follow APG II [2002]. Taxa in red are the three new genomes reported in this paper.

Figure 9. Phylogenetic tree of 35-taxon data set based on 61 chloroplast protein-coding genes using maximum likelihood. The single ML tree has an ML value of – lnL = 342478.92. Numbers at nodes are bootstrap support values ≥ 50%. Scale at base of tree indicates the number of base substitutions. Ordinal and higher level group names follow APG II [2002]. Taxa in red are the three new genomes reported in this paper.

Figure 1

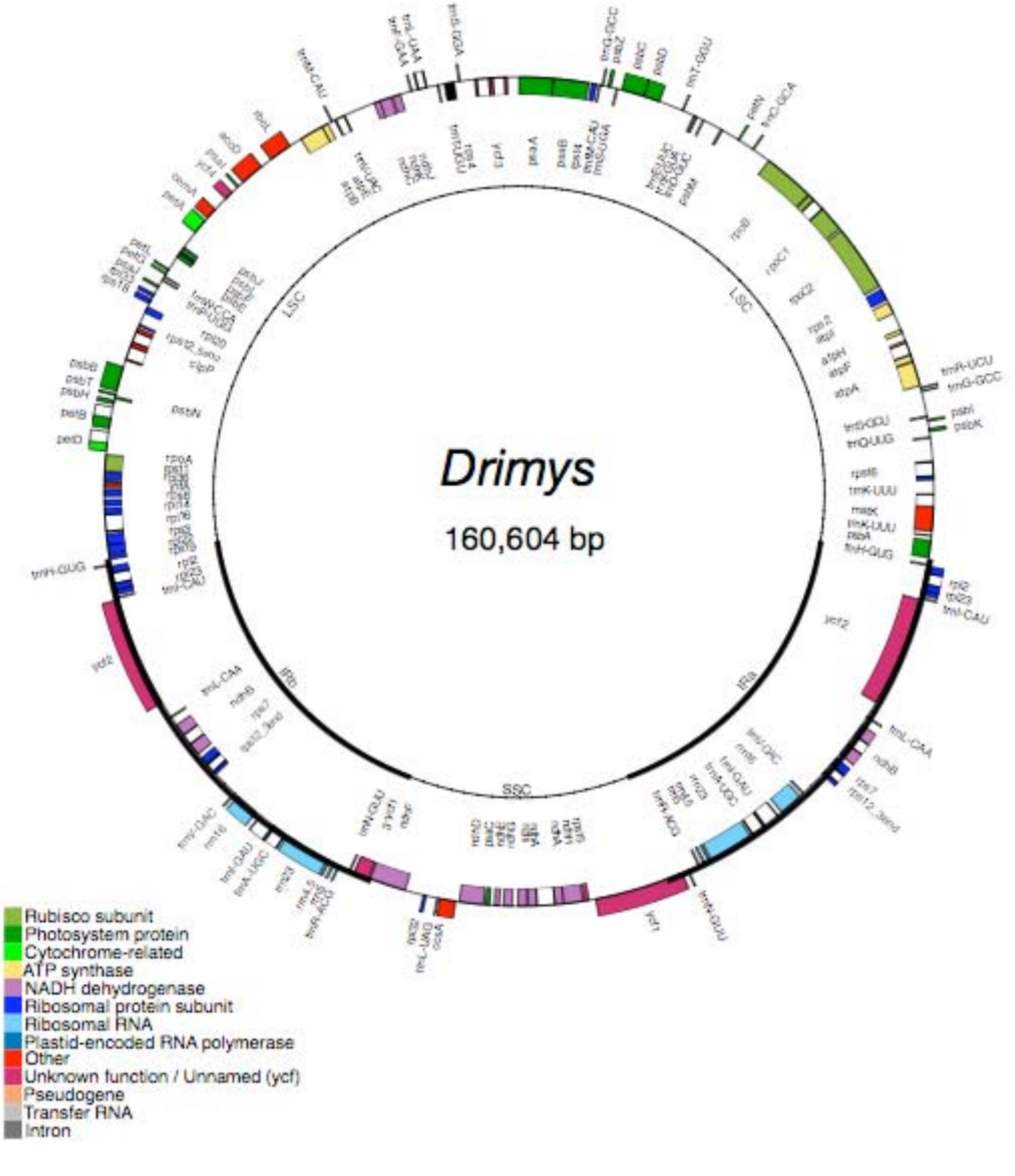

Drimys
160,604 bp
LSC
LSC
IRa
IRb
SSC
trnM-CAU
trnL-UAA
trnF-GAA
trnS-GGA
trnG-GCC
psbZ
psbC
psbD
trnT-GGU
petN
trnC-GCA
trnV-UAC
atpE
atpB
ndhJ
ndhK
ndhC
trnT-UGU
rps4
ycf3
psaA
psaB
rps14
trnfM-CAU
trnS-UGA
trnE-UUC
trnY-GUA
trnD-GUC
psbM
rpoB
rpoC1
rpoC2
rps2
atpI
atpH
atpF
atpA
trnR-UCU
trnG-GCC
psbI
psbK
trnS-GCU
trnQ-UUG
rps16
trnK-UUU
matK
trnK-UUU
psbA
trnH-GUG
rpl2
rpl23
trnI-CAU
ycf2
trnL-CAA
ndhB
rps7
rps12_3end
trnV-GAC
rrn16
trnI-GAU
trnA-UGC
rrn23
rrn4.5
rrn5
trnR-ACG
trnN-GUU
ycf1
ndhD
psaC
ndhE
ndhG
ndhI
ndhA
ndhH
rps15
trnN-GUU
ycf1
ndhF
rpl32
trnL-UAG
ccsA
trnR-ACG
rrn5
rrn4.5
rrn23
trnA-UGC
trnI-GAU
rrn16
trnV-GAC
trnL-CAA
ndhB
rps7
rps12_3end
ycf2
trnH-GUG
trnI-CAU
rpl23
rpl2
rps19
rpl22
rps3
rpl16
rpl14
rps8
infA
rpl36
rps11
rpoA
petD
petB
psbH
psbT
psbB
psbN
clpP
rps12_5end
rpl20
trnW-CCA
trnP-UGG
psbE
psbF
psbL
psbJ
rps18
rpl33
psaJ
petG
petL
petA
cemA
ycf4
psaI
accD
rbcL
Rubisco subunit
Photosystem protein
Cytochrome-related
ATP synthase
NADH dehydrogenase
Ribosomal protein subunit
Ribosomal RNA
Plastid-encoded RNA polymerase
Other
Unknown function / Unnamed (ycf)
Pseudogene
Transfer RNA
Intron

Figure 2

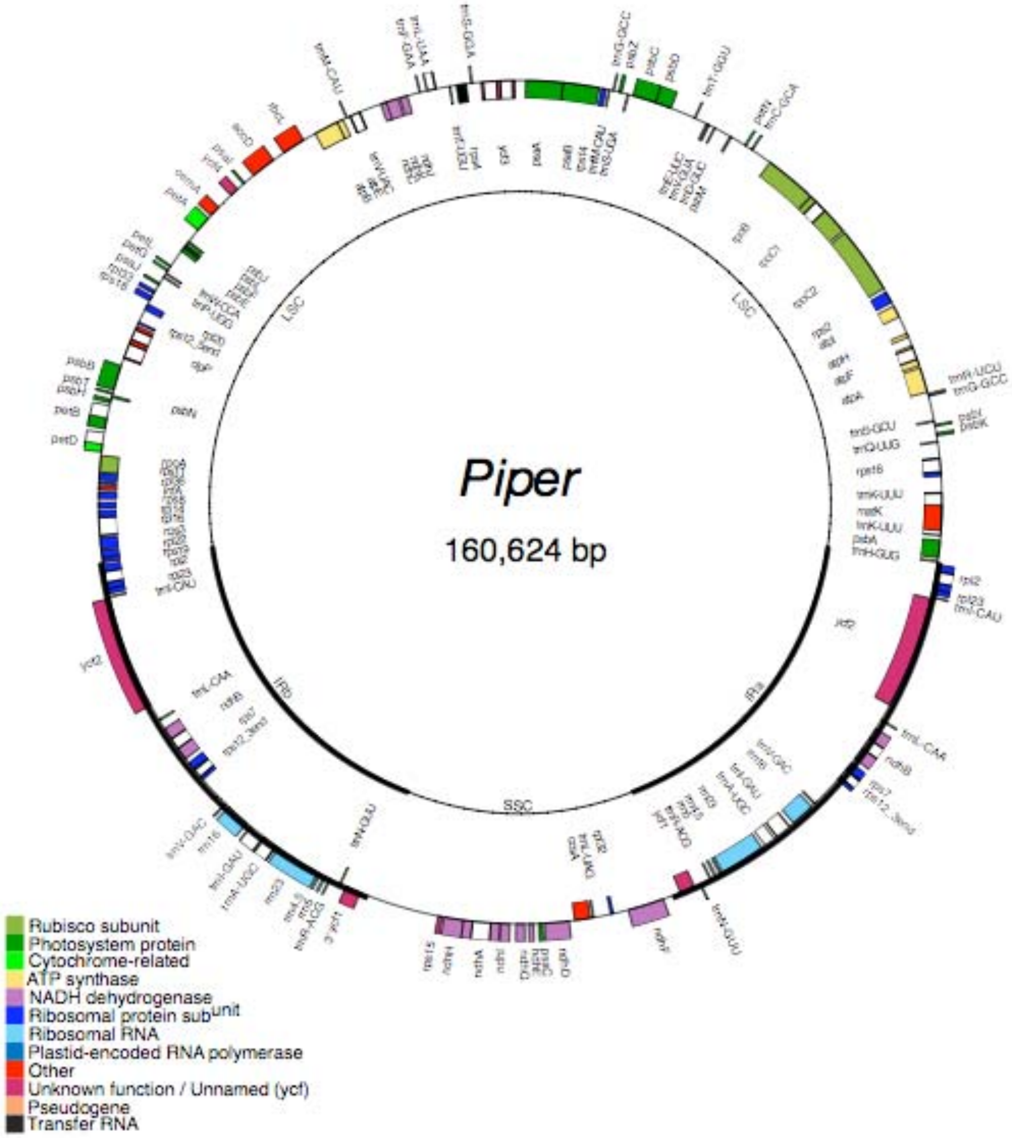

Piper
160,624 bp
LSC
LSC
IRb
IRa
SSC
trnS-GCA
trnL-UAA
trnF-GAA
trnM-CAU
rbcL
accD
psaI
ycf4
cemA
petA
psbJ
psbL
psbF
psbE
petL
petG
psaJ
rpl33
rps18
psbB
psbT
psbH
petB
petD
trnG-GCC
psbZ
psbC
psbD
trnT-GGU
petN
trnC-GCA
rpoB
rpoC1
rpoC2
rps2
atpI
atpH
atpF
atpA
trnR-UCU
trnG-GCC
psbI
psbK
trnS-GCU
trnQ-UUG
rps16
trnK-UUU
matK
trnK-UUU
psbA
trnH-GUG
rpl2
rpl23
trnI-CAU
ycf2
ycf2
trnL-CAA
ndhB
rps7
rps12_3end
trnL-CAA
ndhB
rps7
rps12_3end
trnV-GAC
rrn16
trnI-GAU
trnA-UGC
rrn23
rrn4.5
rrn5
trnR-ACG
trnV-GAC
rrn16
trnI-GAU
trnA-UGC
rrn23
rrn4.5
rrn5
trnR-ACG
ycf1
trnN-GUU
trnN-GUU
ndhF
rpl32
trnL-UAG
ccsA
ndhD
psaC
ndhE
ndhG
ndhI
ndhA
ndhH
rps15
trnE-UUC
trnY-GUA
trnD-GUC
psbM
psbN
clpP
rps12_5end
trnW-CCA
trnP-UGG
psaA
psaB
rps14
trnfM-CAU
trnS-UGA
trnI-CAU
rpl23
Rubisco subunit
Photosystem protein
Cytochrome-related
ATP synthase
NADH dehydrogenase
Ribosomal protein subunit
Ribosomal RNA
Plastid-encoded RNA polymerase
Other
Unknown function / Unnamed (ycf)
Pseudogene
Transfer RNA

Figure 3

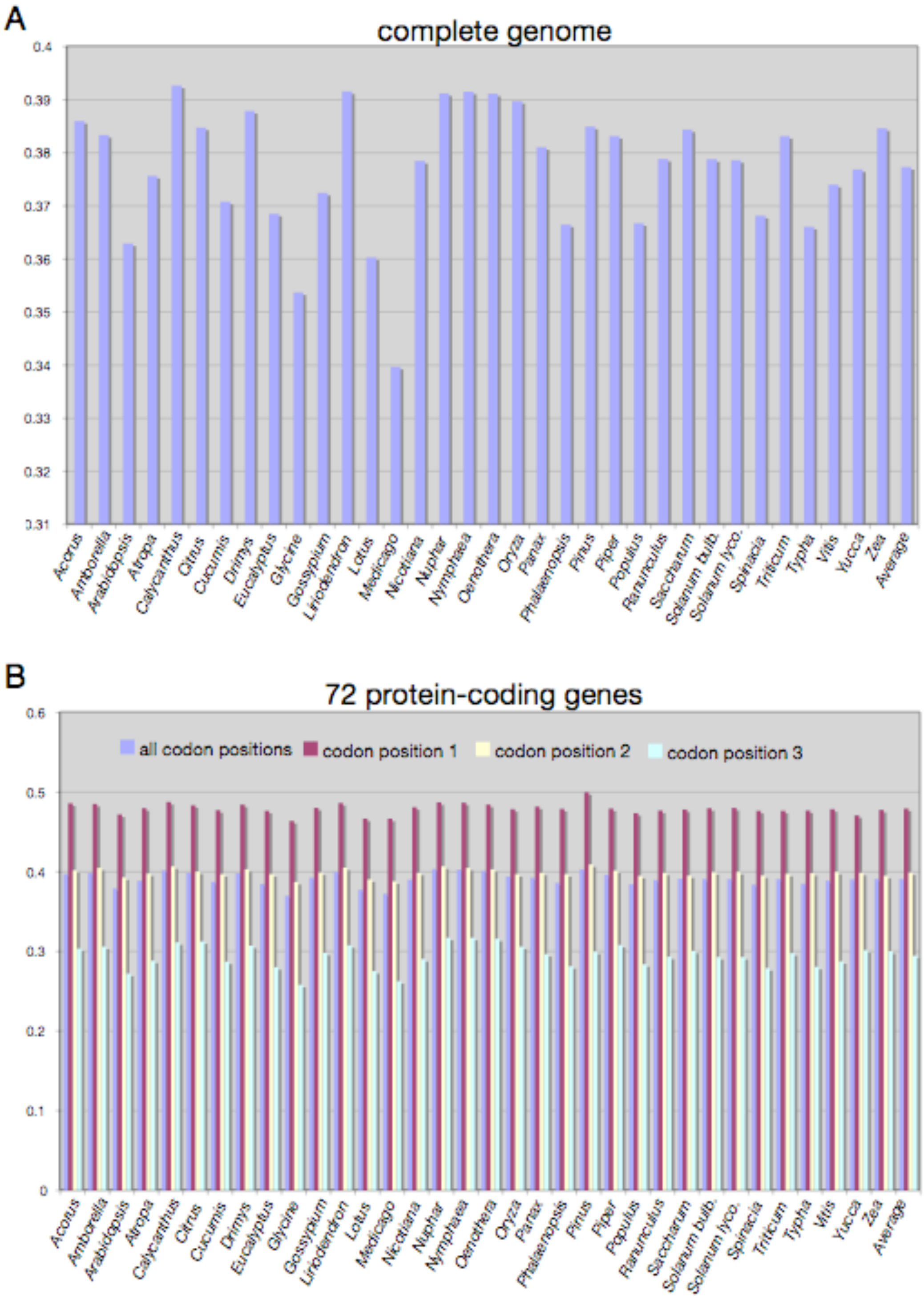
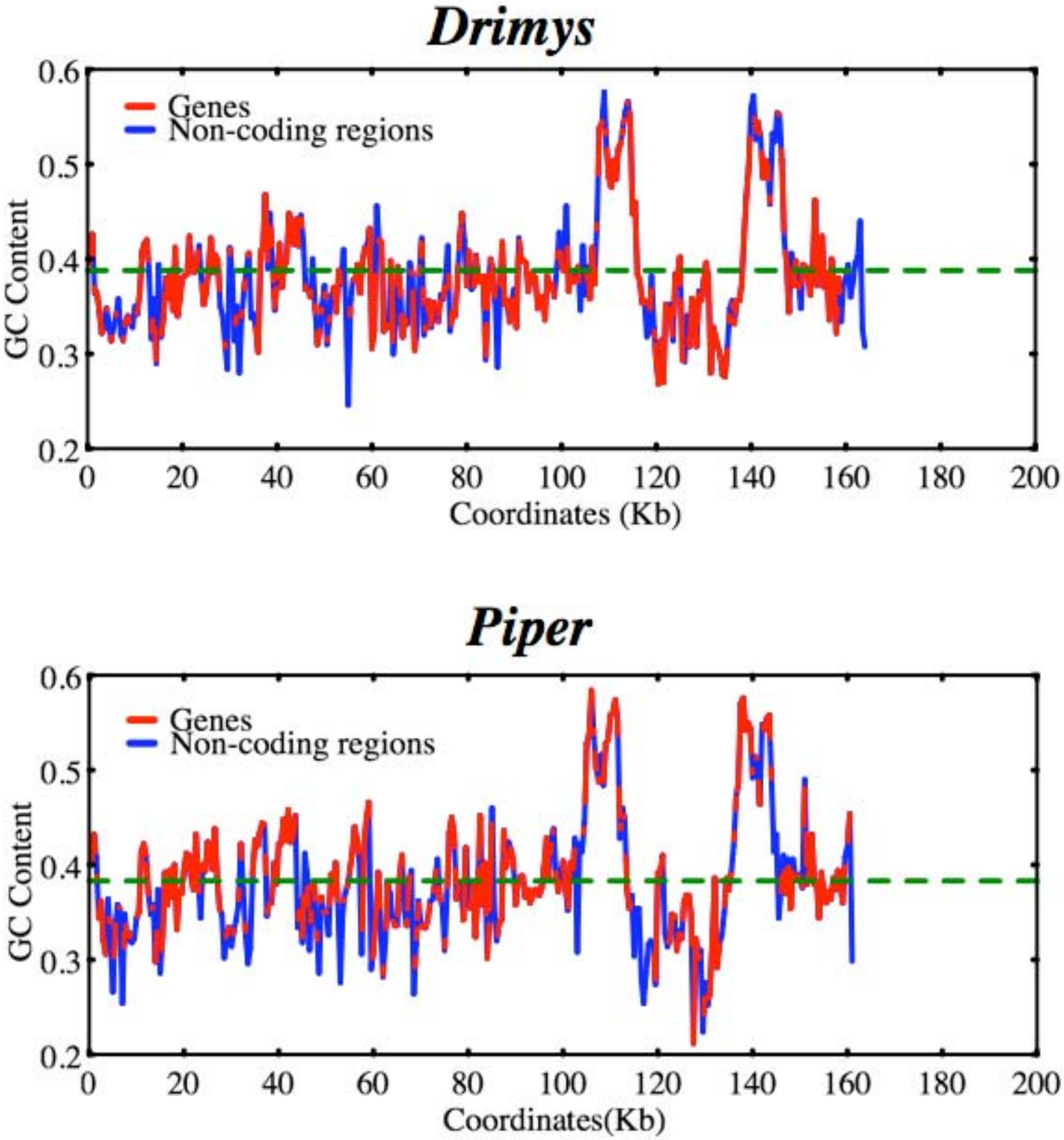

A
complete genome
0.4
0.39
0.38
0.37
0.36
0.35
0.34
0.33
0.32
0.31
Acorus
Amborella
Arabidopsis
Atropa
Calycanthus
Citrus
Cucumis
Drimys
Eucalyptus
Glycine
Gossypium
Liriodendron
Lotus
Medicago
Nicotiana
Nuphar
Nymphaea
Oenothera
Oryza
Panax
Phalaenopsis
Pinus
Piper
Populus
Ranunculus
Saccharum
Solanum bulb.
Solanum lyco.
Spinacia
Triticum
Typha
Vitis
Yucca
Zea
Average
B
72 protein-coding genes
all codon positions
codon position 1
codon position 2
codon position 3
0.6
0.5
0.4
0.3
0.2
0.1
0
Acorus
Amborella
Arabidopsis
Atropa
Calycanthus
Citrus
Cucumis
Drimys
Eucalyptus
Glycine
Gossypium
Liriodendron
Lotus
Medicago
Nicotiana
Nuphar
Nymphaea
Oenothera
Oryza
Panax
Phalaenopsis
Pinus
Piper
Populus
Ranunculus
Saccharum
Solanum bulb.
Solanum lyco.
Spinacia
Triticum
Typha
Vitis
Yucca
Zea
Average

Figure 4



Drimys
Genes
Non-coding regions
GC Content
0.2
0.3
0.4
0.5
0.6
0
20
40
60
80
100
120
140
160
180
200
Coordinates (Kb)
Piper
Genes
Non-coding regions
GC Content
0.2
0.3
0.4
0.5
0.6
0
20
40
60
80
100
120
140
160
180
200
Coordinates(Kb)

Figure 5

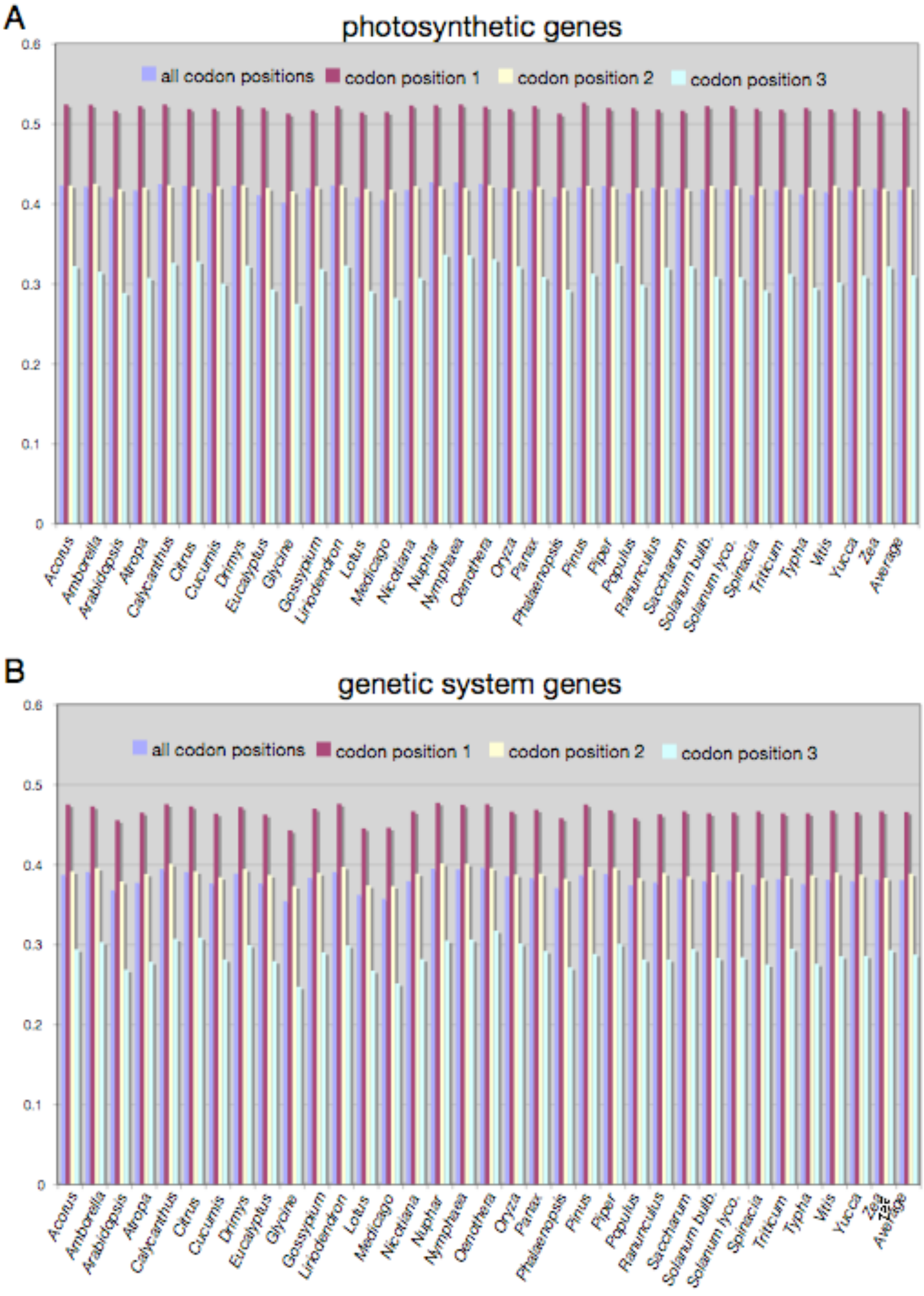

A
photosynthetic genes
all codon positions
codon position 1
codon position 2
codon position 3
0.6
0.5
0.4
0.3
0.2
0.1
0
Acorus
Amborella
Arabidopsis
Atropa
Calycanthus
Citrus
Cucumis
Drimys
Eucalyptus
Glycine
Gossypium
Liriodendron
Lotus
Medicago
Nicotiana
Nuphar
Nymphaea
Oenothera
Oryza
Panax
Phalaenopsis
Pinus
Piper
Populus
Ranunculus
Saccharum
Solanum bulb.
Solanum lyco.
Spinacia
Triticum
Typha
Vitis
Yucca
Zea
Average
B
genetic system genes
all codon positions
codon position 1
codon position 2
codon position 3
0.6
0.5
0.4
0.3
0.2
0.1
0
Acorus
Amborella
Arabidopsis
Atropa
Calycanthus
Citrus
Cucumis
Drimys
Eucalyptus
Glycine
Gossypium
Liriodendron
Lotus
Medicago
Nicotiana
Nuphar
Nymphaea
Oenothera
Oryza
Panax
Phalaenopsis
Pinus
Piper
Populus
Ranunculus
Saccharum
Solanum bulb.
Solanum lyco.
Spinacia
Triticum
Typha
Vitis
Yucca
Zea
Average

Figure 6

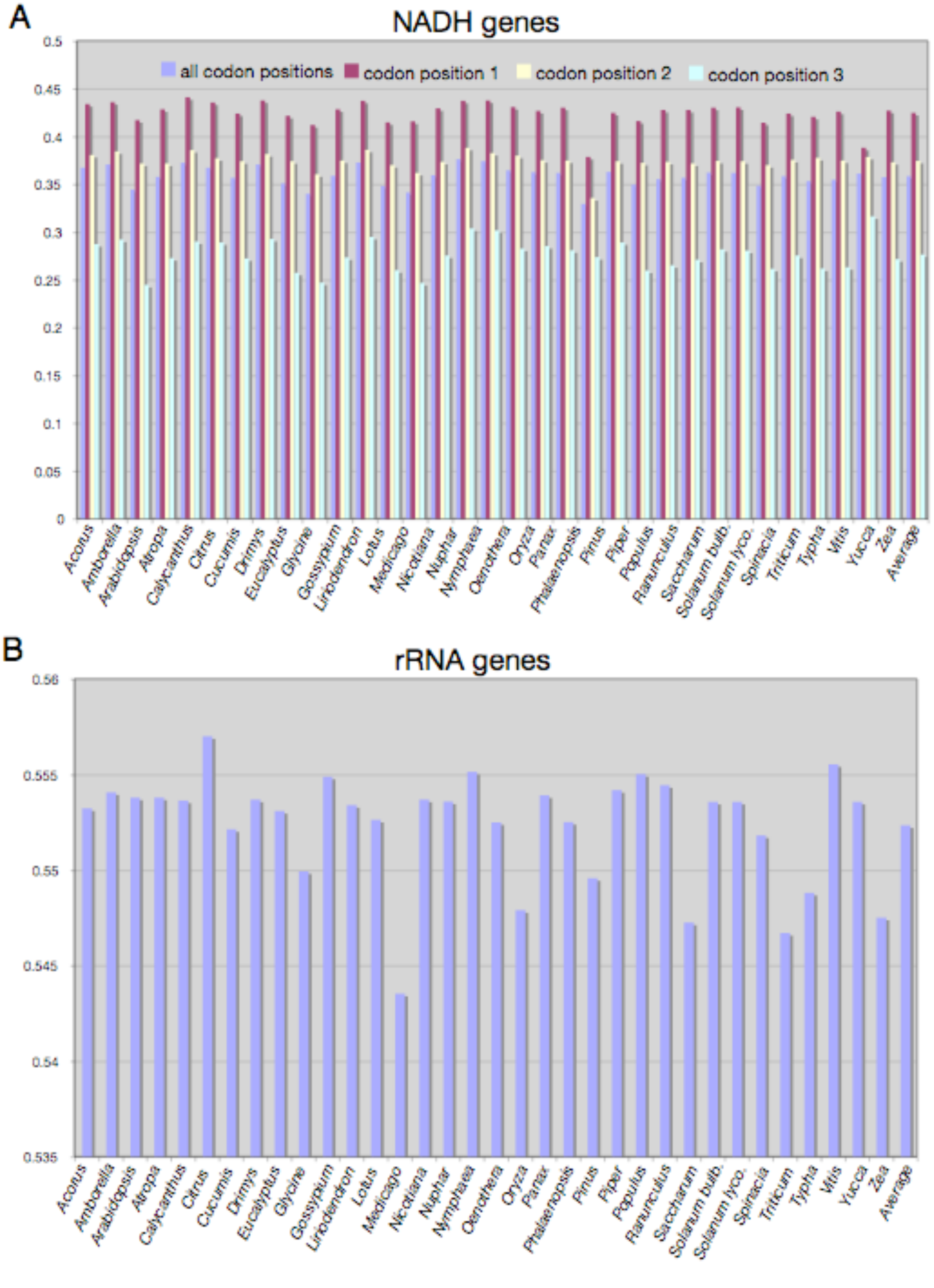

A
NADH genes
all codon positions
codon position 1
codon position 2
codon position 3
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Acorus
Amborella
Arabidopsis
Atropa
Calycanthus
Citrus
Cucumis
Drimys
Eucalyptus
Glycine
Gossypium
Liriodendron
Lotus
Medicago
Nicotiana
Nuphar
Nymphaea
Oenothera
Oryza
Panax
Phalaenopsis
Pinus
Piper
Populus
Ranunculus
Saccharum
Solanum bulb.
Solanum lyco.
Spinacia
Triticum
Typha
Vitis
Yucca
Zea
Average
B
rRNA genes
0.56
0.555
0.55
0.545
0.54
0.535
Acorus
Amborella
Arabidopsis
Atropa
Calycanthus
Citrus
Cucumis
Drimys
Eucalyptus
Glycine
Gossypium
Liriodendron
Lotus
Medicago
Nicotiana
Nuphar
Nymphaea
Oenothera
Oryza
Panax
Phalaenopsis
Pinus
Piper
Populus
Ranunculus
Saccharum
Solanum bulb.
Solanum lyco.
Spinacia
Triticum
Typha
Vitis
Yucca
Zea
Average

Figure 7

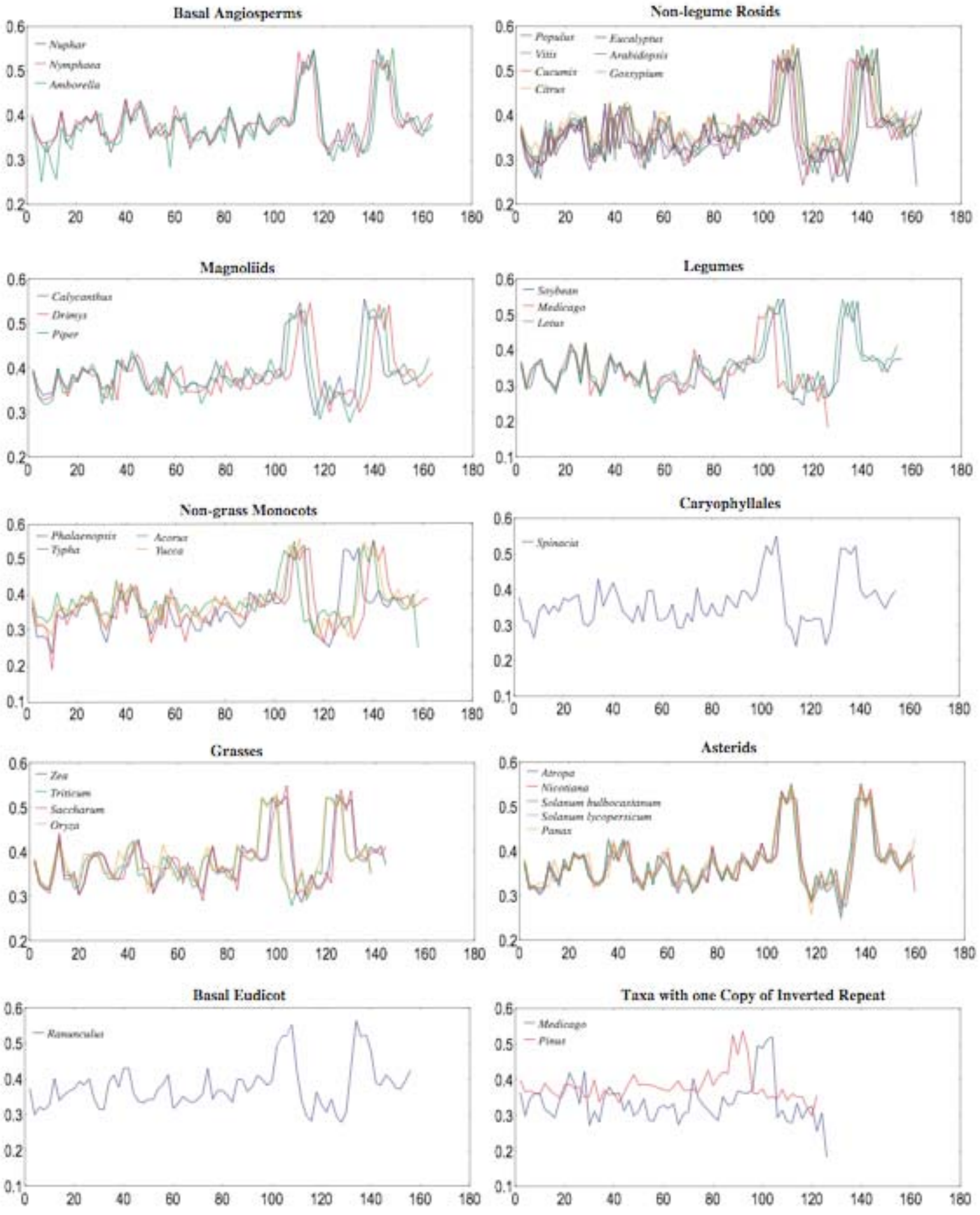
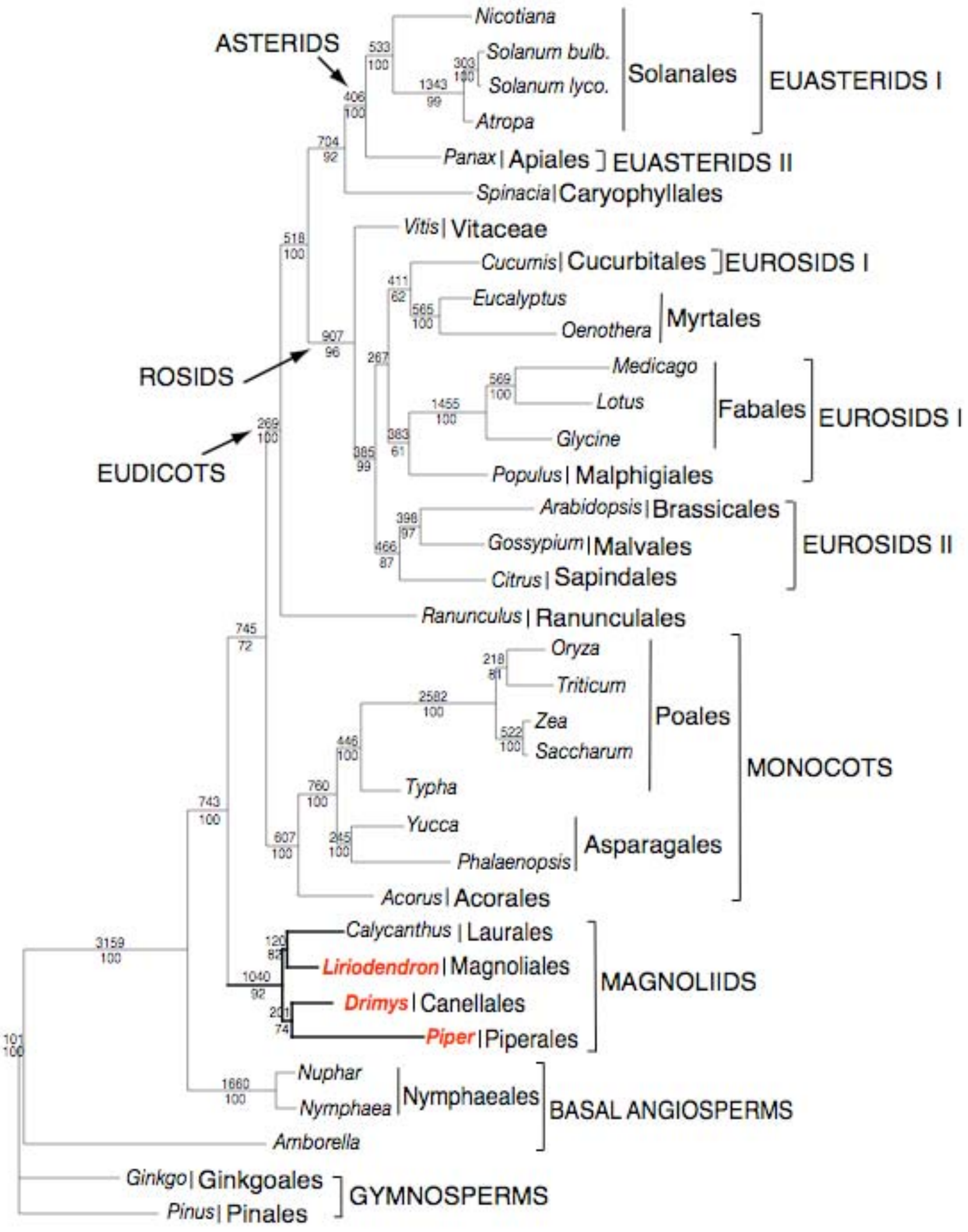
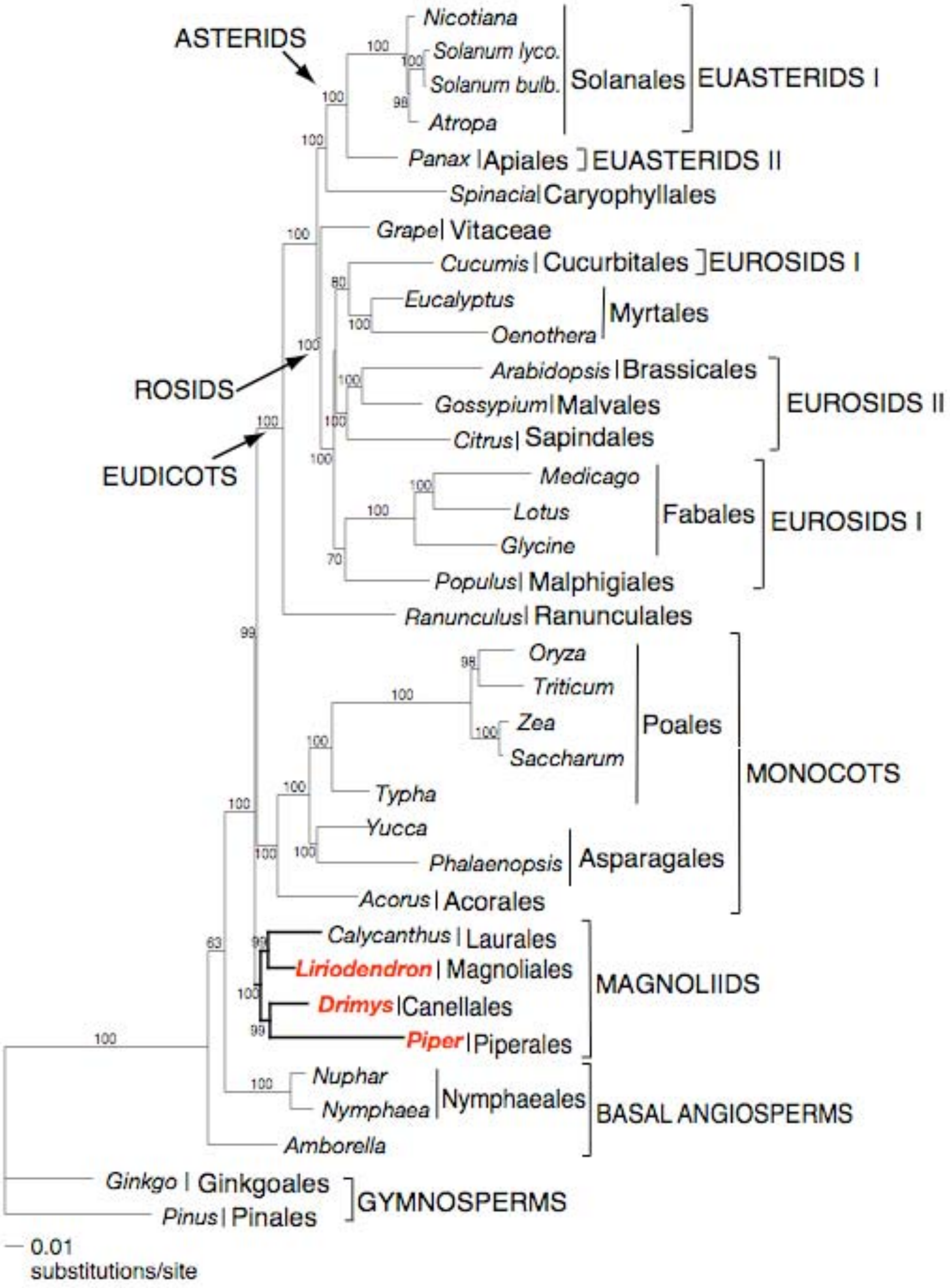

Basal Angiosperms
Nuphar
Nymphaea
Amborella
0.6
0.5
0.4
0.3
0.2
0
20
40
60
80
100
120
140
160
180
Non-legume Rosids
Populus
Vitis
Cucumis
Citrus
Eucalyptus
Arabidopsis
Gossypium
0.6
0.5
0.4
0.3
0.2
0
20
40
60
80
100
120
140
160
180
Magnoliids
Calycanthus
Drimys
Piper
0.6
0.5
0.4
0.3
0.2
0
20
40
60
80
100
120
140
160
180
Legumes
Soybean
Medicago
Lotus
0.6
0.5
0.4
0.3
0.2
0.1
0
20
40
60
80
100
120
140
160
180
Non-grass Monocots
Phalaenopsis
Typha
Acorus
Yucca
0.6
0.5
0.4
0.3
0.2
0.1
0
20
40
60
80
100
120
140
160
180
Caryophyllales
Spinacia
0.6
0.5
0.4
0.3
0.2
0.1
0
20
40
60
80
100
120
140
160
180
Grasses
Zea
Triticum
Saccharum
Oryza
0.6
0.5
0.4
0.3
0.2
0
20
40
60
80
100
120
140
160
180
Asterids
Atropa
Nicotiana
Solanum bulbocastanum
Solanum lycopersicum
Panax
0.6
0.5
0.4
0.3
0.2
0
20
40
60
80
100
120
140
160
180
Basal Eudicot
Ranunculus
0.6
0.5
0.4
0.3
0.2
0.1
0
20
40
60
80
100
120
140
160
180
Taxa with one Copy of Inverted Repeat
Medicago
Pinus
0.6
0.5
0.4
0.3
0.2
0.1
0
20
40
60
80
100
120
140
160
180

Figure 8



ASTERIDS
ROSIDS
EUDICOTS
Nicotiana
Solanum bulb.
Solanum lyco.
Atropa
Solanales
EUASTERIDS I
533 100
303 100
1343 99
406 100
704 92
Panax | Apiales
EUASTERIDS II
Spinacia | Caryophyllales
518 100
907 96
Vitis | Vitaceae
Cucumis | Cucurbitales
EUROSIDS I
411 62
Eucalyptus
Oenothera
565 100
Myrtales
267
Medicago
Lotus
Glycine
569 100
1455 100
Fabales
EUROSIDS I
383 61
Populus | Malphigiales
385 99
269 100
Arabidopsis | Brassicales
Gossypium | Malvales
Citrus | Sapindales
398 97
466 87
EUROSIDS II
Ranunculus | Ranunculales
745 72
Oryza
Triticum
218 81
Zea
Saccharum
522 100
2582 100
Poales
446 100
Typha
760 100
Yucca
Phalaenopsis
245 100
Asparagales
607 100
Acorus | Acorales
MONOCOTS
743 100
Calycanthus | Laurales
Liriodendron | Magnoliales
Drimys | Canellales
Piper | Piperales
120 82
1040 92
201 74
MAGNOLIIDS
3159 100
101 100
Nuphar
Nymphaea
1660 100
Nymphaeales
Amborella
BASAL ANGIOSPERMS
Ginkgo | Ginkgoales
Pinus | Pinales
GYMNOSPERMS

Figure 9


ASTERIDS
ROSIDS
EUDICOTS
Nicotiana
Solanum lyco.
Solanum bulb.
Atropa
Solanales
EUASTERIDS I
Panax | Apiales
EUASTERIDS II
Spinacia | Caryophyllales
Grape | Vitaceae
Cucumis | Cucurbitales
EUROSIDS I
Eucalyptus
Oenothera
Myrtales
Arabidopsis | Brassicales
Gossypium | Malvales
Citrus | Sapindales
EUROSIDS II
Medicago
Lotus
Glycine
Fabales
Populus | Malphigiales
EUROSIDS I
Ranunculus | Ranunculales
Oryza
Triticum
Zea
Saccharum
Poales
Typha
Yucca
Phalaenopsis
Asparagales
MONOCOTS
Acorus | Acorales
Calycanthus | Laurales
Liriodendron | Magnoliales
Drimys | Canellales
Piper | Piperales
MAGNOLIIDS
Nuphar
Nymphaea
Nymphaeales
Amborella
BASAL ANGIOSPERMS
Ginkgo | Ginkgoales
Pinus | Pinales
GYMNOSPERMS
0.01
substitutions/site
100
100
98
100
100
100
100
80
100
100
100
100
100
100
100
70
99
98
100
100
100
100
100
100
63
99
100
99
100
100