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

Permalink
https://escholarship.org/uc/item/9st9z7c8

Authors
Zhengqiu, C.
Penaflor, C.
Kuehl, J.V.
et al.

Publication Date
2008-04-16

Peer reviewed
Complete chloroplast genome sequences of *Drimys*, *Liriodendron*, and *Piper*:

Implications for the phylogeny of magnoliids and the evolution of GC content

C. Zhengqi1, C. Penafior2, J.V. Kuehl3, J. Leebens-Mack4, J. Carlson4, C. W.
dePamphilis4, J. L. Boore3, and R. K. Jansen1*

1Section of Integrative Biology and Institute of Cellular and Molecular Biology, Patterson Laboratories 141, University of Texas, Austin, TX 78712, USA; 2Biology Department, 373 WIDB, Brigham Young University, Provo, Utah 84602, USA; 3DOE Joint Genome Institute and Lawrence Berkeley National Laboratory, Walnut Creek, CA; 4Department 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 Nymphaeales. 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 Nymphaeales. 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 61 chloroplast 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 are 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

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/Magnoliaceae. 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/Magnoliaceae.
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](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, 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


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.


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


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.


Sugiura M: **The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression.** *EMBO J* 1986, **5**:2043-2049.


<table>
<thead>
<tr>
<th></th>
<th>Calycanthus</th>
<th>Drimys</th>
<th>Piper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (bp)</td>
<td>153,337</td>
<td>160,604</td>
<td>160,624</td>
</tr>
<tr>
<td>LSC length (bp)</td>
<td>86,948</td>
<td>88,685</td>
<td>87,666</td>
</tr>
<tr>
<td>SSC length (bp)</td>
<td>19,799</td>
<td>18,621</td>
<td>18,878</td>
</tr>
<tr>
<td>IR length (bp)</td>
<td>23,295</td>
<td>26,649</td>
<td>27,039</td>
</tr>
<tr>
<td>Number of genes</td>
<td>133 (115)</td>
<td>131 (113)</td>
<td>130 (113)</td>
</tr>
<tr>
<td>Number of gene</td>
<td>18</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>duplicated in IR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of genes</td>
<td>18 (3)</td>
<td>18 (3)</td>
<td>18 (3)</td>
</tr>
<tr>
<td>with introns</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Taxa included in phylogenetic analyses with GenBank accession numbers and references.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>GenBank Accession Numbers</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gymnosperms – Outgroups</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Pinus thunbergii</em></td>
<td>NC_001631</td>
<td>Wakasugi et al. 1994 []</td>
</tr>
<tr>
<td><em>Ginkgo biloba</em></td>
<td>DQ069337-DQ069702</td>
<td>Leebens-Mack et al 2005 []</td>
</tr>
<tr>
<td><strong>Basal Angiosperms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Amborella trichopoda</em></td>
<td>NC_005086</td>
<td>Goremykin et al. 2003 []</td>
</tr>
<tr>
<td><em>Nuphar advena</em></td>
<td>DQ069337-DQ069702</td>
<td>Leebens-Mack et al 2005 []</td>
</tr>
<tr>
<td><em>Nymphaea alba</em></td>
<td>NC_006050</td>
<td>Goremykin et al. 2004 []</td>
</tr>
<tr>
<td><strong>Magnoliids</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Calycanthus floridus</em></td>
<td>NC_004993</td>
<td>Goremykin et al. 2003 []</td>
</tr>
<tr>
<td><em>Drimys granatensis</em></td>
<td>To be submitted to GenBank</td>
<td>Current study</td>
</tr>
<tr>
<td><em>Liriodendron tulipifera</em></td>
<td>To be submitted to GenBank</td>
<td>Current study</td>
</tr>
<tr>
<td><em>Piper coenoclatum</em></td>
<td>To be submitted to GenBank</td>
<td>Current study</td>
</tr>
<tr>
<td>Monocots</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>----------</td>
<td>--------------------------</td>
</tr>
<tr>
<td><em>Acorus americanus</em></td>
<td>DQ069337-DQ069702</td>
<td>Leebens-Mack et al. 2005</td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>NC_001320</td>
<td>Hiratsuka et al. 1989</td>
</tr>
<tr>
<td><em>Saccharum officinarum</em></td>
<td>NC_006084</td>
<td>Asano et al. 2004</td>
</tr>
<tr>
<td><em>Triticum aestivum</em></td>
<td>NC_002762</td>
<td>Ikeo and Ogihara,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>unpublished</td>
</tr>
<tr>
<td><em>Phaleanopsis aphrodite</em></td>
<td>AY916449</td>
<td>Chang et al. 2006</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Typha latifolia</em></td>
<td>DQ069337-DQ069702</td>
<td>Leebens-Mack et al. 2005</td>
</tr>
<tr>
<td><em>Yucca schidigera</em></td>
<td>DQ069337-DQ069702</td>
<td>Leebens-Mack et al. 2005</td>
</tr>
<tr>
<td><em>Zea mays</em></td>
<td>NC_001666</td>
<td>Maier et al. 1995</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Eudicots</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>NC_000932</td>
<td>Sato et al. 1999</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Atropa belladonna</em></td>
<td>NC_004561</td>
<td>Schmitz-</td>
</tr>
<tr>
<td>Species</td>
<td>Accession</td>
<td>Reference</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>------------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td><em>Citrus sinensis</em></td>
<td>XXXXXXXX</td>
<td>Linneweber et al. 2002</td>
</tr>
<tr>
<td><em>Cucumis sativus</em></td>
<td>NC_007144</td>
<td>Bausher et al. unpublished</td>
</tr>
<tr>
<td><em>Eucalyptus globulus</em></td>
<td>AY780259</td>
<td>Plader et al. unpublished</td>
</tr>
<tr>
<td><em>Glycine max</em></td>
<td>DQ317523</td>
<td>Saski et al. 2005</td>
</tr>
<tr>
<td><em>Gossypium hirsutum</em></td>
<td>DQ345959</td>
<td>Lee et al. 2006</td>
</tr>
<tr>
<td><em>Medicago truncatula</em></td>
<td>NC_003119</td>
<td>Lin et al., unpublished</td>
</tr>
<tr>
<td><em>Nicotiana tabacum</em></td>
<td>NC_001879</td>
<td>Shinozaki et al. 1986</td>
</tr>
<tr>
<td><em>Oenothera elata</em></td>
<td>NC_002693</td>
<td>Hupfer et al. 2000</td>
</tr>
<tr>
<td><em>Panax schinseng</em></td>
<td>NC_006290</td>
<td>Kim and Lee 2004</td>
</tr>
<tr>
<td><em>Populus trichocarpa</em></td>
<td></td>
<td>unpublished</td>
</tr>
<tr>
<td><em>Ranunculus</em></td>
<td>DQ069337-DQ069702</td>
<td>Leebens-Mack et</td>
</tr>
<tr>
<td>Species</td>
<td>Accession</td>
<td>Authors</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------</td>
<td>------------------</td>
</tr>
<tr>
<td><em>macranthus</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Solanum lycopersicum</em></td>
<td>DQ347959</td>
<td>Daniell et al.</td>
</tr>
<tr>
<td><em>Solanum bulbocastanum</em></td>
<td>DQ347958</td>
<td>Daniell et al.</td>
</tr>
<tr>
<td><em>Spinacia oleracea</em></td>
<td>NC_002202</td>
<td>Schmitz-Linnweber et al.</td>
</tr>
<tr>
<td><em>Vitis vinifera</em></td>
<td>DQ424856</td>
<td>Jansen et al.</td>
</tr>
</tbody>
</table>
**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 $-\ln L = 342478.92$. Numbers at nodes are bootstrap support values $\geq 50\%$. 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
Figure 3

A

complete genome

B

72 protein-coding genes
Figure 4

**Drimys**

![Graph showing GC content for Drimys genes and non-coding regions.]

**Piper**

![Graph showing GC content for Piper genes and non-coding regions.]

Figure 7

Basal Angiosperms
- Nephase
- Nymphphone
- Amborella

Non-legume Rosids
- Papulu
- Eucalyptus
- Vitis
- Arabidopsis
- Cucumis
- Citrus
- Coccygium

Magnoliids
- Calycanthus
- Dranys
- Piper

Legumes
- Sycamora
- Medicago
- Lotus

Non-grass Monocots
- Phalaenopsis
- Acroc
- Raxa

Caryophyllales
- Spinacea

Grasses
- Zea
- Triticum
- Saccharum
- Grena

Asterids
- Abroga
- Nereidina
- Solanum helenoxalbo
d- Solanum lycopersicum
- Pusa

Basal Eudicot
- Ranaeculus

Taxa with one Copy of Inverted Repeat
- Medicago
- Pima