Complete Plastid Genome Sequences of Three Rosids (*Castanea*, *Prunus*, *Theobroma*): Evidence for At Least Two Independent Transfers of *rpl22* to the Nucleus

Robert K. Jansen[1] Christopher Saski,[2] Seung-Bum Lee,[3] Anne K. Hansen,[1] and Henry Daniell*[,3]

[1]Section of Integrative Biology and Institute of Cellular and Molecular Biology, The University of Texas at Austin

[2]Clemson University, Genomics Institute

[3]Department of Molecular Biology and Microbiology, University of Central Florida

***Corresponding author:** E-mail: daniell@mail.ucf.edu.

Associate editor: Charles Delwiche

Research article

Abstract

Functional gene transfer from the plastid to the nucleus is rare among land plants despite evidence that DNA transfer to the nucleus is relatively frequent. During the course of sequencing plastid genomes from representative species from three rosid genera (*Castanea*, *Prunus*, *Theobroma*) and ongoing projects focusing on the Fagaceae and Passifloraceae, we identified putative losses of *rpl22* in these two angiosperm families. We further characterized *rpl22* from three species of *Passiflora* and one species of *Quercus* and identified sequences that likely represent pseudogenes. In *Castanea* and *Quercus*, both members of the Fagaceae, we identified a nuclear copy of *rpl22*, which consisted of two exons separated by an intron. Exon 1 encodes a transit peptide that likely targets the protein product back to the plastid and exon 2 encodes *rpl22*. We performed phylogenetic analyses of 97 taxa, including 93 angiosperms and four gymnosperm outgroups using alignments of 81 plastid genes to examine the phylogenetic distribution of *rpl22* loss and transfer to the nucleus. Our results indicate that within rosids there have been independent transfers of *rpl22* to the nucleus in Fabaceae and Fagaceae and a putative third transfer in *Passiflora*. The high level of sequence divergence between the transit peptides in Fabaceae and Fagaceae strongly suggest that these represent independent transfers. Furthermore, Blast searches did not identify the "donor" genes of the transit peptides, suggesting a de novo origin. We also performed phylogenetic analyses of *rpl22* for 87 angiosperms and four gymnosperms, including nuclear-encoded copies for five species of Fabaceae and Fagaceae. The resulting trees indicated that the transfer of *rpl22* to the nucleus does not predate the origin of angiosperms as suggested in an earlier study. Using previously published angiosperm divergence time estimates, we suggest that these transfers occurred approximately 56–58, 34–37, and 26–27 Ma for the Fabaceae, Fagaceae, and Passifloraceae, respectively.

Key words: plastid genome, *rpl22*, gene transfer, rosids.

Introduction

Subsequent to the endosymbiotic origin of plastids from a cyanobacterial ancestor, there was a massive transfer of genes to the nucleus (reviewed in Timmis et al. 2004). Cyanobacteria encode 5,000–7,000 genes, and only 20–200 of these have been retained in plastid genomes. This early mass transfer of genes suggests that a large proportion of nuclear genes originated in the plastid; estimates in *Arabidopsis* indicate that 18% of its nuclear genes originate from the ancestral plastid genome (Martin et al. 2002). The streamlining of the ancestral plastid genome has resulted in a very compact genome, highly conserved with respect to organization, gene content, and gene order. Examination of the 125 land plant plastid genomes currently on GenBank shows that genome size, gene content, and gene order are for the most part highly conserved, with substantial variation in intergenic spacer regions (Daniell et al. 2006; Saski et al. 2007) and regulatory sequences (Ruhlman et al. 2010). Most genomes have a quadripartite structure with two copies of a large inverted repeat separating two unequally sized single-copy regions termed the large and small single-copy regions. Land plant plastid genomes are 108–217 kb, with the vast majority in the 150–170 kb range. Most plastid genomes contain 110–130 distinct genes; the majority of these genes (about 80) code for proteins and are mostly involved in photosynthesis or gene expression with the remainder being transfer RNA (about 30) or ribosomal RNA (4) genes (Raubeson and Jansen 2005; Bock 2007).

Although most functional gene transfers to the nucleus occurred during early stages of plastid evolution, nonfunctional DNA transfers to the nucleus continue at a high rate (Martin et al. 2002; Timmis et al. 2004; Matsuo et al. 2005; Noutsos et al. 2005). Recent examinations of plant nuclear genomes demonstrated the presence of a large number of nuclear-localized plastid DNA fragments (nupDNAs). An extensive analysis of rice (Matsuo et al. 2005) estimated that there were 701 insertions of plastid DNA into the nucleus, for a total of 0.9 mb of nupDNAs representing 0.12% of the nuclear genome. The inserted fragment sizes varied in length from 38 bp to 131 kb, and the largest one included almost the entire plastid genome. Two studies examined the rate of plastid DNA transfer to the nucleus in tobacco

© The Author 2010. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

using plastid transformation. Using transgenic tobacco plants, Huang et al. (2003) found high frequency of transfer with 1 transfer in 16,000 gametes. A similar experiment with somatic cells also showed a high level of transfer but a 300-fold reduction relative to gametes (Stegemann et al. 2003).

Although there is a remarkably high rate of plastid DNA transfer to the nucleus, very few examples of functional gene transfers in land plants have been documented. Merely, transferring plastid DNA into the nuclear genome is not sufficient for functional gene transfer; for this, plastid genes must also acquire nuclear regulatory elements as well as transit peptides. Among land plants, there have been many proposed plastid gene losses (summarized in Raubeson and Jansen 2005; Jansen et al. 2007), but subsequent molecular characterizations of these events have been limited. Successful gene transfers in land plants have been documented for only four genes, including multiple transfers of *infA* in rosids (Millen et al. 2001), *rpl22* in *Pisum* (Gantt et al. 1991), *rpl32* in some Salicaceae (Cusack and Wolfe 2007; Ueda et al. 2007), and *rpoA* in mosses (Sugiura et al. 2003). The loss of *rps16* in *Medicago truncatula* and *Populus alba* was identified as a gene substitution rather than transfer to the nucleus because in these two species a nuclear-encoded mitochondrial-targeted copy is also targeted to the plastid (Ueda et al. 2008). Acetyl-CoA carboxylase subunit D (*accD*) gene has been lost six times from angiosperm plastid genomes (Jansen et al. 2007), but its fate has only been determined for the grasses. In this case, the prokaryotic multisubunit carboxylase has been replaced by plastid-targeted eukaryotic carboxylase (Konishi et al. 1996; Gornicki et al. 1997). A similar situation has been documented for *rpl23* in spinach; the prokaryotic plastid *rpl23* has been replaced by a eukaryotic cytosolic copy of this ribosomal protein (Bubunenko et al. 1994). Thus far three distinct pathways of recent plastid gene loss have been identified in land plants: transfer to the nucleus (*infA*, *rpl22*, *rpl32*, and *rpoA*), substitution of a nuclear-encoded mitochondrial targeted gene (*rps16*), and substitution of a nuclear gene for a plastid gene (*accD*, *rpl23*).

In the course of sequencing plastid genomes for additional rosids (*Castanea*, *Prunus*, and *Theobroma*) as well as from ongoing projects focusing on the Fagaceae and Passifloraceae, we identified putative losses of *rpl22* in these two angiosperm families. In this paper, we report the complete plastid genome sequence of three of these rosids, characterize the nonfunctional *rpl22* in the plastids of *Castanea*, *Quercus*, and *Passiflora*, document the transfer of *rpl22* to the nucleus in *Castanea* and *Quercus*, and discuss the phylogenetic distribution and timing of this gene transfer in angiosperm evolution.

Materials and Methods

DNA Sources

Publicly available bacterial artificial chromosome (BAC) libraries (http://www.genome. clemson.edu) of *Castanea mollissima* cultivar Vanuxem, *Theobroma cacao* received from United States Department of Agriculture-Agriculture Research Service-Subtropical Horticulture Research Station, Miami, FL, and *Prunus persica* cultivar Nemared were screened for plastid inserts using a *Glycine max* plastid DNA probe, and the first 96 positive clones were pulled from the library, arrayed in a 96-well microtitre plate, copied, and archived. Selected clones were then subjected to *Hind*III fingerprinting and *Not*I digests. End sequences were determined and localized on the plastid genome of *Arabidopsis thaliana* to deduce the relative positions of the clones; then a single clone that covered the entire genome was chosen for sequencing. *Passiflora* plastid DNA was isolated from fresh leaf tissue using methods described in Jansen et al. (2005), and vouchers are deposited at TEX.

DNA Sequencing and Genome Assembly

The nucleotide sequences of the selected BACs were determined by the bridging shotgun method. The purified BAC DNA was subjected to hydroshearing, end repair, and then size-fractionated by agarose gel electrophoresis. Fractions of approximately 3.0–5.0 kb were eluted and ligated into the vector pBlueskript IIKS+. The shotgun libraries were plated and then arrayed into forty 96-well microtitre plates for the sequencing reactions. Sequencing was performed using the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA). Sequence data from the forward and reverse priming sites of the shotgun clones were accumulated until eight times the size of the genome and assembled using the Phred–Phrap programs (Ewing and Green 1998). The plastid genomes of *Passiflora biflora* and *P. quadrangularis* were sequenced using methods described in Jansen et al. (2005), whereas *P. cirrhiflora* was sequenced using the 454 method as described in Moore et al. (2006).

Annotations of all plastid genomes were done with DOGMA (Wyman et al. 2004). Sequences of the plastid genomes of *Castanea*, *Prunus*, and *Theobroma* and gene sequences of *Passiflora* and *Quercus* have been deposited in GenBank (accession numbers HQ336404 - HQ336412).

Identification and Isolation of the *rpl22* Gene in the *Castanea* and *Quercus* Nuclear Genome

A *C. mollissima* unigene assembly containing 48,335 contigs (~850,000 454 sequences from various tissue sources, http://www.fagaceae.org) was searched using Blast with the *Pisum sativum rpl22* gene sequence (gi169065). This resulted in the identification of contig5108, which is a unigene assembly of 17 individual 454 sequences and displays homology to the conserved region of the *rpl22* coding sequence. Forward and reverse polymerase chain reaction (PCR) primers were designed from the conserved coding sequence using the Primer3 software (http://frodo.wi.mit.edu/primer3/) (Contig5108F 5′-GGCGTTCCTATG-AGGAGTCA-3′ and Contig5108R 5′-ATATGACACG-AGCGCCTTCT-3′), and confirmed amplification products were radiolabeled using the DECAprimeII random primer labeling kit (Ambion) and used to probe two (10× coverage each) genomic *Castanea* BAC libraries. The

hybridization conditions were as follows: ^{32}P-labeled probe was denatured for 10 min at 95 °C and cooled on ice for 1 min and then added to hybridization tube containing four filters in 50 ml hybridization buffer (0.5 M phosphate buffer, 7% sodium dodecyl sulfate [SDS], and 1 mM ethylenediaminetetraacetic acid). Hybridization was performed at 60 °C overnight; filters were washed two times with 1× standard saline citrate, 0.1% SDS at 60 °C for 30 min and exposed to phosphor screens overnight and the images recorded by a Typhoon 9400 Imager (GE Healthcare, Bio-Sciences).

A *Quercus rubra* unigene assembly consisting of 28,041 contigs (277,154 454 sequences from various tissues, http://www.fagaceae.org) was searched using Blast with the *C. mollissima* plastid-encoded *rpl22* pseudogene and the *P. sativum rpl22* nuclear-encoded gene (gi169065). This search resulted in the identification of two distinct contigs; RO454_contig27007_v2 (ID3244801) and RO454_contig15690_v2 (ID3133835).

Alignments of both DNA and protein sequence data were performed using MUSCLE (Edgar 2004) in the Geneious Pro 4.8.4 (Drummond et al. 2009). Two bioinformatic tools were used to identify putative transit peptides and predict their target, TargetP version 1.1 (Emanuelsson et al. 2000) and Predotar version 1.03 (Small et al. 2004).

Phylogenetic Analyses

Phylogenetic analyses were performed on two data sets. The first included 97 species whose plastid genomes are completely sequenced (Supplementary table 1, Supplementary Material online), including 93 angiosperms from all the major clades and four gymnosperm outgroups. For each species, nucleotide sequence of the 81 included genes from Jansen et al. (2007) were extracted from the plastid genome, sequences were translated, amino acid sequences were aligned in MSWAT (http://mswat.ccbb.utexas.edu), manually adjusted, and this alignment was used to constrain the nucleotide alignment. The second data set included the *rpl22* gene sequence for 94 taxa. This matrix includes the same taxa as the 81-gene data set except that those taxa missing *rpl22* were deleted (six Fabaceae, *Castanea*, and *Passiflora*) and the nuclear copies for three Fabaceae (*Glycine*, *Medicago*, *Pisum*) and two Fagaceae (*Castanea*, *Quercus*) were included. The aligned data matrices are at http://chloroplast.psu.edu/supplement.html.

Phylogenetic analyses using maximum parsimony (MP) and maximum likelihood (ML) were performed with PAUP*version 4.10b10 (Swofford 2003) and GARLI version 0.942 (Zwickl 2006), respectively. Gap regions were treated as missing data and not excluded or recorded. MP searches included 100 random addition replicates and tree bisection reconnection (TBR) branch swapping with the Multrees option. Nonparametric bootstrap analyses (Felsenstein 1985) were performed for MP analyses with 1,000 replicates with TBR branch swapping, one random addition replicate, and the Multrees option. MrModeltest 2 (Nylander 2004) was used to determine the most appropriate model of DNA sequence evolution. Hierarchical likelihood ratio tests and the Akaike information criterion were used to assess the models that best fit the data, which was determined to be GTR+I+gamma. For ML analyses in GARLI, two independent runs were performed using the default settings (see GARLI manual at http://www.bio.utexas.edu/faculty/antisense/garli/Garli.html). Nonparametric bootstrap analyses (Felsenstein 1985) were performed in GARLI for ML analyses using default settings and 1,000 replicates.

Results

Genome Organization of Three New Rosid Sequences

The three newly sequenced rosid plastid genomes are quite similar to each other in terms of overall organization, gene/intron content, gene order, and GC content (fig. 1, table 1, with accession numbers), and they fall within the typical size range for photosynthetic angiosperm plastid genomes that have not been rearranged (Raubeson and Jansen 2005; Bock 2007). The only exceptional feature is the putative loss of one ribosomal protein gene, *rpl22*, in *Castanea*. There is a pseudogene with 16 internal stop codons remaining in the plastid genome at the correct location within the highly conserved S10 operon (fig. 2). Another genus of Fagaceae, *Quercus*, also has a pseudogene of *rpl22* in the plastid that has six internal stop codons (fig. 2). Alignment of the *rpl22* pseudogene sequence (fig. 3) of *Castanea* and *Quercus* with copies of the gene for 28 other eudicots shows that sequence divergence of the *Castanea* pseudogene is not substantially higher than levels of divergence between functional copies of other rosids (69.2 vs. 70.6%). It also shows that *rpl22* is incomplete in *P. biflora*, *P. cirrhiflora*, and *P. quadrangularis* with seven or eight internal stop codons, suggesting that in these three species the plastid-encoded copy may also not be functional (fig. 4). Overall average sequence identity among all 33 taxa (28 functional copies, *Castanea*, *Quercus*, and three *Passiflora* pseudogenes) is 62.9% with the 11 asterids having 83.6% identity and the 22 rosids with 66.5% identity. Identity between the 17 intact copies in rosids not including *Castanea*, *Quercus*, and *Passiflora* is only slightly higher at 70.6%.

Characterization of Nuclear-Encoded *rpl22* Gene in *Castanea* and *Quercus*

Blast searches against the *C. mollissima* unigene expressed sequence tag (EST) assembly containing 48,335 contigs with the nuclear-encoded *P. sativum rpl22* gene sequence (M60952) identified a contig (5108) with high sequence identity to the *rpl22* coding sequence. This EST sequence did not contain the internal stop codons observed in the plastid-encoded pseudogene or homology to either of the genes (*rps19* and *rps3*) that normally flank *rpl22* in the plastid. A PCR product generated from within this sequence was radioactively labeled and used to probe two

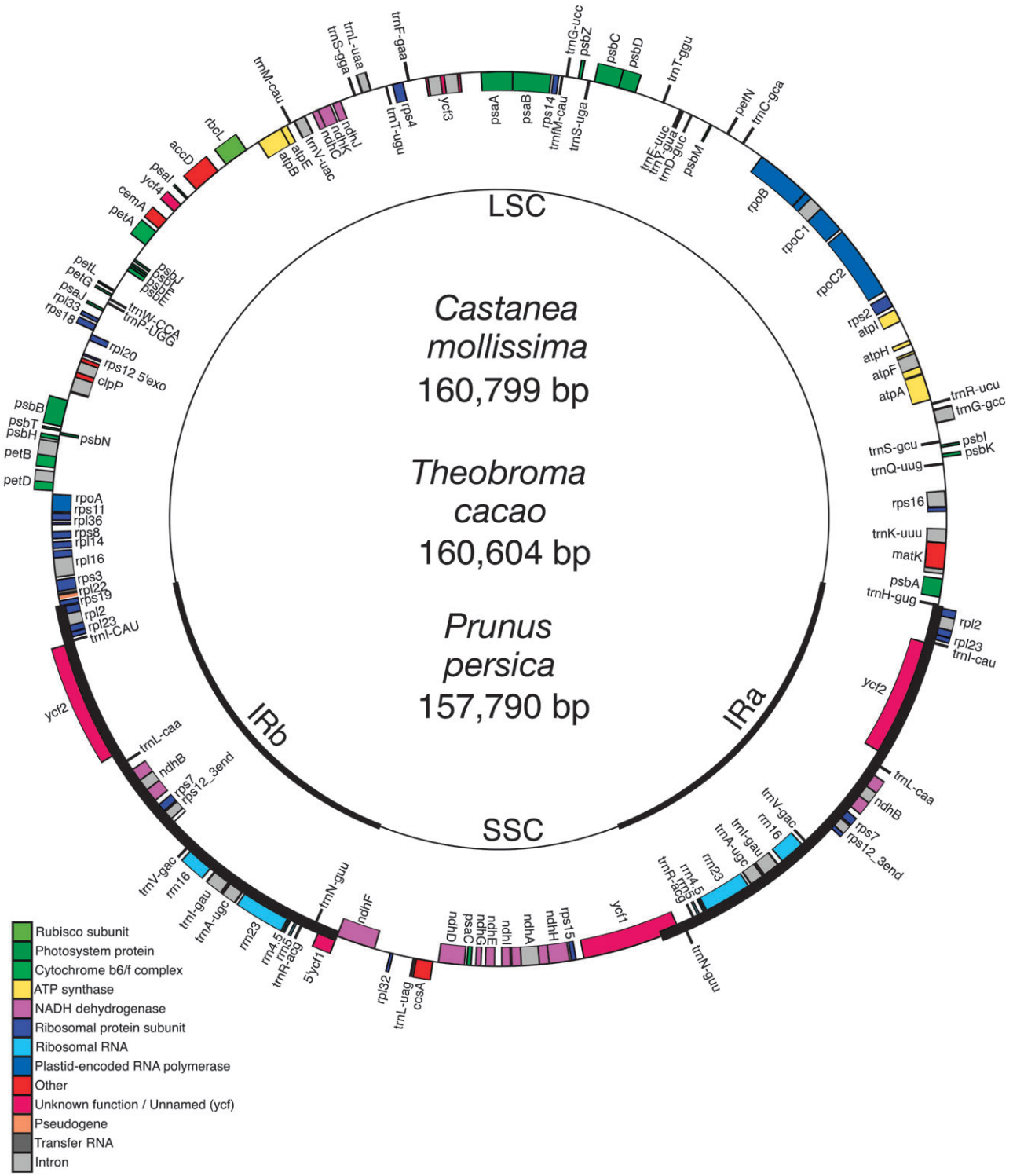

Fig. 1. Circularized gene map of the plastid genomes of three rosids. The thick lines indicate the extent of the inverted repeats (IRa and IRb), which separate the genomes into small (SSC) and large (LSC) single-copy regions. Genes on the outside of map are transcribed in the counterclockwise direction, and genes on the inside of the map are transcribed in the clockwise direction.

(10 × coverage each) genomic BAC libraries of *Castanea*. The hybridization resulted in 12 positively identified BAC clones, indicating a single to low copy number sequence. To determine nuclear integration of the *rpl22* gene, direct BAC sequencing was performed to generate sequence data in both the 5′ and 3′ direction from each of the 12 BAC clones. Of the 12 BACs, only four contain *rpl22* and sequence reads assembled together to form a 3,250 bp consensus sequence, 2,039 bp of which represents the nuclear copy of the *rpl22* gene (fig. 5). The *Castanea rpl22* gene contains two exons totaling 609 bp separated by a 1,430 bp intron. The exon/intron boundaries have the

Table 1. Comparison of Major Features of Three Newly Sequenced Rosid Plastid Genomes.

	Castanea mollissima (HQ336406)	*Prunus persica* (HQ336405)	*Theobroma cacao* (HQ336404)
Size (bp)	**160,799**	**157,790**	**160,604**
LSC length (bp)	**90,432**	**85,968**	**89,395**
SSC length (bp)	**18,995**	**19,060**	**20,187**
IR length (bp)	**25,686**	**26,381**	**25,511**
Number of genes	**127**	**128**	**128**
Number of gene duplicated in IR[a]	**16**	**16**	**16**
Number of genes with introns (with 2 introns)	**18 (3)**	**18 (3)**	**18 (3)**
GC content	**36.8%**	**36.8%**	**36.9%**

Note.—Genbank accession numbers are provided below each species. IR, inverted repeat.
[a] *rps12* is not included in this number; it has one exon in SSC and two exons in IR; only genes completely duplicated are included.

highly conserved sequences (gt at 5′ end and ag at 3′ end) that are required for intron splicing (fig. 5).

Blast searches of the *Q. rubra* unigene assembly consisting of 28,041 contigs with the *C. mollissima* plastid-encoded *rpl22* pseudogene and the *P. sativum rpl22* nuclear-encoded gene sequences identified two distinct contigs; RO454_contig27007_v2 (ID3244801) and RO454_contig15690_v2 (ID3133835), respectively. RO454_contig27007 is 3,807 bp and consists of 132, 454 EST sequences, and RO454_contig15690_v2 is 743 bp and consists of five 454 EST sequences. RO454_contig27007_v2 has high nucleotide sequence identity (95%) to the *Castanea* plastid-encoded pseudogene and contains the flanking plastid-encoded genes (*rps19* and *rps3*) and contains six internal stop codons (fig. 2). RO454_contig15690_v2 has high nucleotide sequence identity (95.9%) to the *Castanea* nuclear-encoded *rpl22* gene sequence.

```
Castanea mollissima
atg ata aag aag aat cca tat acc gaa ata tat atg cgc ttt aag taa cca tat atg tat
 M  I  K  K  N  P  Y  T  E  I  Y  M  R  F  K  *  P  Y  M  Y
gtc tgc taa taa agc aga aag agt aat tga aat tga tca gat ttg tgg acg ttt ata cga
 V  C  *  *  S  R  K  S  N  *  N  *  S  D  L  W  T  F  I  R
aga aag acg tat gag act cga act tat gcc tta tcg agt ggg tta tcc aat ttt taa att
 R  K  T  Y  E  T  R  T  Y  A  L  S  S  G  L  S  N  F  *  I
ggt tta ttc tgc agc aac aaa tgc tat tca caa tgt cgg ttt aaa cga agc aag ttt aat
 G  L  F  C  S  N  K  C  Y  S  Q  C  R  F  K  R  S  K  F  N
cat tag caa agc gga agt cgt gaa ggg gta cta ctg tga aaa aat taa aac ctc gag ctc
 H  *  Q  S  G  S  R  E  G  V  L  L  *  K  N  *  N  L  E  L
gag ggc gta gtt atc cga taa aaa gac ccg ctt ttc ata taa cta ttg gat taa aag ata
 E  G  V  V  I  R  *  K  D  P  L  F  I  *  L  L  D  *  K  I
tat ctg taa agg aag tat aaa gga gcc aaa tac gcc ata tta tac ttc aaa ggc aac gta
 Y  L  *  R  K  Y  K  G  A  K  Y  A  I  L  Y  F  K  G  N  V
tgg aga aat aaa aat aag taa gta cac gga tat gac gtg tca tga tat ata tag tat tgg
 W  R  N  K  N  K  *  V  H  G  Y  D  V  S  *  Y  I  *  Y  W
gag att atg gga caa aaa ata a
 E  I  M  G  Q  K  I

Quercus rubra
atg ata aag aag aat cca tat acc gaa cat gtg aaa taa tat ata tgc gct tta agt aaa
 M  I  K  K  N  P  Y  T  E  H  V  K  *  Y  I  C  A  L  S  K
cat ata tgt atg tct gct aat aaa gca gaa aga gta att gaa att gat cag att tgt gga
 H  I  C  M  S  A  N  K  A  E  R  V  I  E  I  D  Q  I  C  G
cgt tta tac gaa gaa aga cgt atg aga ctc gaa ctt atg cct tat cga gtg ggt tat cca
 R  L  Y  E  E  R  R  M  R  L  E  L  M  P  Y  R  V  G  Y  P
att tta aaa ttg gtt tat tct gca gca aca aat gct att cac aat gtc ggt tta aac gaa
 I  L  K  L  V  Y  S  A  A  T  N  A  I  H  N  V  G  L  N  E
gca agt tta atc att agc aaa gcg gaa gtc gtg aag ggg tac tac tgt gaa aaa att aaa
 A  S  L  I  I  S  K  A  E  V  V  K  G  Y  Y  C  E  K  I  K
acc tcg agc tcg agg gcg tag tta tcc gat aaa aag acc cgt ttt tca tat aac tat tgg
 T  S  S  S  R  A  *  L  S  D  K  K  T  R  F  S  Y  N  Y  W
att aaa aga tat atc tgt aaa gga agt ata aaa aag gaa gta taa agg agc caa ata cgc
 I  K  R  Y  I  C  K  G  S  I  K  K  E  V  *  R  S  Q  I  R
cat att ata ctt caa ctt caa agg caa cgt atg gag aaa taa aaa taa gta agt aca cgg
 H  I  I  L  Q  L  Q  R  Q  R  M  E  K  *  K  *  V  S  T  R
ata tga cgt gtc atg ata tat ata gta ttg ggn gat tat ggg aca aaa aat aa
 I  *  R  V  M  I  Y  I  V  L  X  D  Y  G  T  K  N
```

Fig. 2. DNA and amino acid sequence of *rpl22* pseudogenes in the plastid genomes of *Castanea mollissima* and *Quercus rubra*. Asterisks indicate stop codons.

TargetP and Predotar were used to predict the target of the transit peptide of exon 1 for *Castanea* and *Quercus*. TargetP predicted this exon to be a plastid-targeted protein in both cases with a reliability class score of 2 (actual prediction values were plastid = 0.818/0.869, mitochondrion = 0.217/0.155, secretory pathway = 0.03/0.045, and other = 0.025/0.05 for *Castanea*/*Quercus*). Predotar also predicted that exon 1 is targeted to the plastid with the following prediction values (plastid = 0.50/0.57, mitochondrion = 0.23/0.09, endoplasmic reticulum = 0.03/0, and elsewhere = 0.37/0.39 for *Castanea*/*Quercus*). Thus, the 5′ exon of both *Castanea* and *Quercus* is predicted to encode the transit sequence that targets the plastid. The 3′ exon blasts to the plastid-encoded *rpl22* with a high sequence identity up to 85%.

Protein sequences for nuclear copies of *rpl22* from three Fabaceae (*G. max*—AK286885, *M. truncatula*—L00667, and *P. sativum*—M60952) and two Fagaceae (*C. mollissima*—HQ336407 and *Q. rubra*—HQ336408) were aligned to plastid-encoded sequences from five closely related rosids (fig. 6). The transit peptides in exon 1 for both Fabaceae and Fagaceae are highly divergent with a sequence identity of only 29.3% among the five species. This is in contrast to the much higher sequence identity of 67.9% among these five species for exon 2. Sequence identity among the transit peptides of the three Fabaceae is 46.1% and among the two Fagaceae is 78.3%. Overall sequence identity of exon 2 of Fabaceae and Fagaceae with the plastid-encoded *rpl22* from the five related rosids is 62.4%, indicating the high level of conservation of this gene regardless of where it is encoded.

The transit peptides (both DNA and protein) of exon 1 for *Castanea*, *Quercus*, and *Pisum* were subjected to Blast searches to attempt to identify the source of these sequences. In the case of *Castanea* and *Quercus*, there were no hits for the protein searches but the DNA sequence of *Castanea* has a 77.4% identity to a 62 bp transcription regulator from *Arabidopsis* (accession number NM_122687). In contrast, Blast results of both DNA and proteins for *Pisum* matched nuclear-encoded copies of *rpl22* in only two other legumes, *Medicago* (62.8% aa identity) and *Glycine* (42.4% aa identity) over 78 amino acids.

Fig. 3. Nucleotide alignment of plastid copies of *rpl22* for 28 eudicots and pseudogenes in *Castanea mollissima*, *Quercus rubra*, and three species of *Passiflora*. Identity across taxa is indicated by the histogram shown at top of alignment.

Phylogenetic Distribution of *rpl22* Loss/Transfer

Phylogenetic analyses were performed on a data set that included 81 protein-coding genes (78,765 nucleotide positions) for 97 taxa (supplementary table 1, Supplementary Material online). MP analyses resulted in a single fully resolved tree with a length of 205,523, a consistency index (CI) of 0.33 (excluding uninformative characters) and a retention index (RI) of 0.67 (tree not shown). Bootstrap analyses indicated that 82 of the 94 nodes were supported by values ≥95%, and 77 of these had a bootstrap value of 100%. Of the remaining 12 nodes, four had bootstrap values between 70% and 94%. ML analysis resulted in a single tree with −lnL = 1108190.01 (fig. 7). ML bootstrap values were also high, with values of ≥95% for 84 of the 94 nodes and 100% for 80 of these nodes. The ML and MP trees had similar topologies and one of the major differences concerned the position of *Piper* and *Ceratophyllum* as described in earlier papers (Jansen et al. 2007; Moore et al. 2007). These differences have no effect on the interpretation of *rpl22* loss/transfer across angiosperms so they will not be presented here. One other difference in the MP and ML trees, that is, relevant to the phylogenetic distribution of *rpl22* loss/transfer concerns the order of taxa in the eurosid I clade. The MP topology for these taxa is included in the inset in figure 7. The difference in the branching pattern between the MP and ML involves the position of *Cucumis* and *Castanea*. In the ML tree, these two genera are sister taxa with 91% bootstrap support, whereas in the MP tree, *Cucumis* is sister to the clade that includes the six genera of Fabaceae with weak support (64% bootstrap) and *Castanea*

P. biflora

g gaa gaa gaa gtt ata cga taa aaa gac taa ctt gtc ata taa ata ttg tat tga aat cta
 E E E V I R * K D * L V I * I L Y * N L
Tat gtt tat aag aag aat ata aga tat gct taa aag aat aag tca aaa aaa gtc cac aaa
 Y V Y K K N I R Y A * K N K S K K V H K
tat gac att tca tga tat ata tta tag taa agg ggg att atg gca caa aaa a
 Y D I S * Y I L * * R G I M A Q K

P. cirrhiflora

gag gac gaa gtt ata c ga tag aaa gac taa ctt gtc ata taa ata ttg tat tga aag ata
 E D E V I R * K D * L V I * I L Y * K I
tat cct tat at g aag aat ata aga tat gct taa aac ttc gaa taa gtc aaa aaa agt cca
 Y P Y M K N I R Y A * N F E * V K K S P
caa ata t ga cat ttc atg ata tat att ata gta atg gag gat tat ggc aca aaa aa
 Q I * H F M I Y I I V M E D Y G T K

P. quadrangularis

gag gac gag gtt cta cga taa aaa gac taa ctt gtc ata taa ata ttg tat tga aag ata
 E D E V L R * K D * L V I * I L Y * K I
tat cct tat att tat at g aag aat ata aga tat gct taa aac ttc gaa taa gta aaa aaa
 Y P Y I Y M K N I R Y A * N F E * V K K
agt cca caa ata tga cat ttc atg ata tat att ata gta atg ggg gat tat ggc aca aaa aa
 S P Q I * H F M I Y I I V M G D Y G T K

Fig. 4. DNA and amino acid sequence of putative *rpl22* pseudogenes in the plastid genomes of three species of *Passiflora*. Asterisks indicate stop codons.

```
cacatggaag gccaatggta gttatataat aggagaagta atatagtttc gcaataaggt      60
taagtgtaga aacaaattca atatctaaaa aaaaaattgg ttatagttag cactgctcca     120
tatccaaatg gctgttttta tattttatgc agtttatttt ttacaataat gttaccataa     180
attttgaaaa atttcacaag ttatttaaat ggtaaaatat ataatattgt acgtattttc     240
attaaagata atgctacctt cataacatat tttcaataat ttggtagatg gtaaattatt     300
attagtttta atttgaacgt acttaaatta tatttttgct acaaataata gtcaattaca     360
atttatcact taaaatatgt tgtgtttgtc cgttccgtaa gaccatcaat gtgtttttc      420
tataaaagaa taaatgaaat tgaaattatt ttgtcttatt atttttatc cagtaaccag      480
tcgccaccct aaatacacta gcggacacac ttttgtggat gttgcaaaat tgtaacccct     540
ctcaaccaaa agcaaaacct tatcctgcgc tgagtctgac tgagctcATG GCCATGGCTC     600
TCACTCTTTG TCACGCGATG TCTCGCTTGT CTCTCTTTCA TCGAAACTCT CAGATTCCTC     660
CCAACCCCAA CACCAACACC AACACTGCCA CATCCATTCA CCCACTGAGA TTCCCAAAGC     720
CTAATAATGA CCTTTTAACG CTCAAAACTA CTACCTTCAT TCCCAATAAT ACCCATTACG     780
CCCCTGTCCT AACTCGCCCT CGCGCCACAG CTCAACCTCA ATCTGAAGgt aggtacccct     840
tctcaatctg cttgctttga ttagatcaaa ataatttctt ttatttatta atgctccaga     900
tttatttctt ttttttgggt ttgcttcaga tacatcatat atgtgttctt gttacaccac     960
aggttttttt ttttttggc ttgaaatttt gtagtattaa ttgcgtttca ttggaatttt    1020
gattagctcc tcccatatgg aaaaaaccca cccaatgaaa gtggtcccca gtgattgtcg    1080
ataggcattc ataacacaat cattttgatt gtgatgagcc gttgcttctg ttcagtggac    1140
tatatggcct atagtggagg tgtgaatgtg atgaggggag tatcagtatc atgccgcaat    1200
ggaagttggg atggaatggt cagtccatga aggttgttag acaatatcct atctgggtag    1260
tggcgagaag ggtgatatgt tgtagcctag gggtggatga ttccattgcc aactgctaaa    1320
agattgataa cttttctgta agaatgggga ttgatggaat agtataaatt gaagaatagg    1380
gagagtgcaa attgctttgt ggatctacat gttaacatac tactctttgc ctgtccacct    1440
gtttgttcgt atgccaaaaa ctatgattat ttgaaaataa taagcactgg ttttagcttt    1500
attatgctag tagtgaccga aattgagaaa gaaaaatctt cacattcctc tctttggttg    1560
tacatcatca tttactaaaa ggaaaataag tcctgggttc ttagcattcg ggatattaga    1620
tgatcaacct gcaactaggt aagtaccctc attctgaaat taatgcagcc aataggtgat    1680
tccattatta ttaattttta aattgctgaa tacaaagtcc aggtttgtgg ttaaggaatg    1740
atgcttgatg aaagccactc aaaaaggtga tttgtgaggc atattataat ggggctgaag    1800
gtgtaaattt taactcctcg gaatgcattg aatgaaaatg gagaaatgta gaaggcttaa    1860
ttaatttagc acaacttatc cttgatttgt aatttattga ctgctatcat ctaaaagttt    1920
cctgagccaa tttagcatta atctgtattc actttgctta gatggactaa ttgcttagaa    1980
gctttcctgt gagaattgaa acagtggggt ttgattttgt ttaatagaag aggtatttat    2040
ggctgtgtgc ttatttatat gcagaacatg tctgagaatg ttgaagaaat catgattcat    2100
aattttcagc atgcccatat attactcttt aagtattcat gcttttgtgt atgataagtc    2160
tagttggata atctaatgga aataaaacag agaagctagt ctcaaatgat catttacttg    2220
taaaataaaa ggtttcatat gcttatgaaa tacgacagGC ATTGTGACAA AGAAGGGGAA    2280
GCAGGATTCT TATGCAGAAG CACGCGCAAT TGGTCGATAC ATACCTATGT CTGCTAACAA    2340
AGCACGAAGA GCAATTGATC AGATTCGTGG GCGTTCCTAT GAGGAGTCAC TTATGATACT    2400
GGAACTCATG CCTTACCGAG CATGTGATCC CATTCTCAAA TTGGTTTATT CTGCAGCAGC    2460
AAATGCTAGT AAAAACATGG GTTTGAATGA AGCAAGTTTA GTAGTTAGTA AAGCTGAAGT    2520
CAATGAGGGT CCTACTAGGA AAAAAGCTAG ACCTCAAGCT CGTGGAAGAG TTCATCCCAT    2580
AAGAAGGCGC TCGTGTCATA TTACTGTTGT ATTGAAAGAT ACATCTTTGT GAttatgaat    2640
ggcacgcaac taaactttta ataccagttc aacctgagat gttccgcttg taacgttgaa    2700
agacatgcaa ttgttgtaga ttctgtagga actttagaag aactataatc catgttttta    2760
aacctatgat atctcctgaa ttattttaca aagttgcatg accaaaatgt gattcaacaa    2820
aattgctcaa acttgtgtga cccaaaaggt aaagagtttg ctaacgtgta tattaaataa    2880
tatagccctc ccctttcata ttatgttggc actatttttt atgactcgtt ttaaaattgg    2940
ctttgccatt gattcaagtc catttacccc tctacataga ggaaattcga aattgtgtag    3000
aattctagcg gtaccatgaa tagtactgta tgttctgttg taggtctgaa tccctttgac    3060
cctgtatgtc ctgggactgt tgtgggtctg aatccctttg accctgtatg ttctgttgct    3120
catcctttga tgattggtgt agctgacaat cctaaacaat ctacgagttg tgcaattcag    3180
ccacctgaag tttttttgtt ttttattttt tggggtgggt ttgtatttct tgatgtgatg    3240
catattatgg                                                           3250
```

Fig. 5. Nucleotide sequence of 3,250 bp region that includes nuclear-encoded copy of *rpl22* in *Castanea mollissima*, which contains two exons totaling 609 bp separated by a 1,430 bp intron (Accession number HQ336407). The highly conserved sequences (gt at 5′ end and ag at 3′ end) at exon/intron boundaries required for intron splicing are highlighted in gray. Exon sequences are indicated in bold uppercase letters and start and stop codons are underlined.

is sister to this clade, also with weak support (68% bootstrap). These differences have no effect on the interpretation of the phylogenetic distribution of *rpl22* gene loss/transfer; both topologies indicate two independent transfers of *rpl22* to the nucleus, one in Fabaceae and the second in *Castanea* (arrow heads in fig. 7). There is another other putative loss in *Passiflora* (closed circles in fig. 7); in this case, there is a truncated portion of *rpl22* in the plastid genomes of three species but as yet, no studies have been performed to demonstrate if this partial copy is nonfunctional or if there is a functional copy in the nucleus.

Phylogenetic Analysis of *rpl22*

Phylogenetic relationships among the nuclear-encoded *rpl22* gene from the five species of Fabaceae and Fagaceae and the plastid-encoded copies from 89 other seed plants (four gymnosperm outgroups and 85 angiosperms) were examined by performing MP and ML analyses of DNA sequences for 94 taxa. The aligned data set contained 779 nucleotides. MP analyses identified 75 shortest trees with a length of 3,078 steps, a CI of 0.34 (excluding uninformative characters) and a RI of 0.66 (tree not shown). ML analyses resulted in a tree with a $-\ln L = 13996.74$ (fig. 8). The MP and ML trees were highly congruent, and in cases where there was incongruence bootstrap support was weak (<50%) because the data set comprised only a single gene for 94 taxa. Despite the limited number of characters, the overall topology was similar to the 81 gene phylogeny (cf. figs. 7–8). The most important result is that the nuclear-encoded copies of *rpl22* were nested within the eudicot

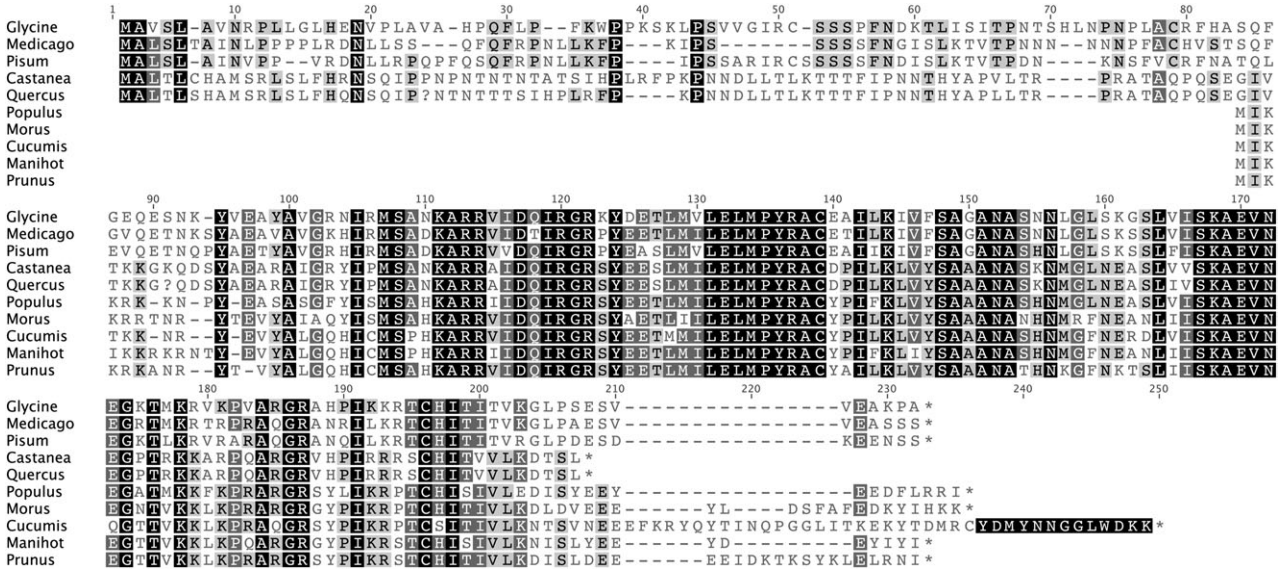

Fig. 6. Amino acid sequence alignment of the nuclear copies of *rpl22* of three Fabaceae (*Glycine max*—AK286885, *Medicago truncatula*—L00667, *Pisum sativum*—M60952) and two Fagaceae (*Castanea mollissima*—HQ336408 and *Quercus rubra*—HQ336408) with the plastid copies of *rpl22* from five eurosid I species (see supplementary table 1 for accession numbers). The first 83 amino acids in Fabaceae and Fagaceae are exon 1 and represent the transit peptide.

clade with 64% bootstrap support in the ML tree (fig. 8). More specifically, Fabaceae and Fagaceae were sister taxa (73% bootstrap support) and were nested in a clade that included other rosids.

Discussion

The availability of plastid genome sequences has increased rapidly during the past decade resulting in over 100 publicly available sequences for most major lineages of angiosperms (http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=2759&opt=plastid). These data have provided many new insights into phylogenetic relationships among flowering plants (Jansen et al. 2007; Moore et al. 2007, 2010), genome-wide patterns and rates of nucleotide substitutions (Chang et al. 2006; Guisinger et al. 2008, 2010; Zhong et al. 2009), and genomic rearrangements (Chang et al. 2006; Chumley et al. 2006; Funk et al. 2007; Jansen et al. 2007; Lee et al. 2007; McNeal et al. 2007; Cai et al. 2008; Haberle et al. 2008). There has also been considerable effort to sequence plastid genomes from crop plants because of the increased interest in plastid genetic engineering (Verma and Daniell 2007). These sequences provide valuable information on endogenous regulatory regions for optimal transgene expression, especially in view of the high level of sequence divergence of intergenic spacer regions (Ruhlman et al. 2010). Our plastid genome sequences for three economically important tree species (cacao, chestnut, and peach) add to this important resource. The three rosid genomes reported here have the ancestral angiosperm genome organization and gene content (Raubeson et al. 2007) except for the presence of a pseudogene of *rpl22* in *Castanea*. The rest of our discussion focuses on the evolutionary implications of the transfer of *rpl22* to the nucleus.

Previous studies indicated that the *rpl22* gene is present in the plastid genome of all land plants except legumes (Doyle et al. 1995), and that this gene has been transferred to the nucleus in *Pisum* (Gantt et al. 1991). The availability of plastid genome sequences of more than 125 land plants on GenBank has confirmed that this gene is plastid-encoded in most land plants. *rpl22* was reported as missing in *Gossypium hirsutum* (Lee et al. 2006), and it was suggested to be a pseudogene in *Citrus sinensis* (Bausher et al. 2006). The report for *Gossypium* was later determined to be an annotation error (see Bausher et al. 2006 for correction). Although Bausher et al. (2006) reported that *rpl22* may be a pseudogene in *Citrus*, they suggested that experimental studies are needed to determine if the truncated copy in the plastid genome is functional. Until these studies are performed, we do not consider the evidence strong enough to report this gene to be absent from *Citrus*. Our plastid genome sequences from four families of rosids, Fagaceae (*Castanea*, *Quercus*), Passifloraceae (*Passiflora*), Rosaceae (*Prunus*), and Malvaceae (*Theobroma*), have demonstrated that *rpl22* has also been transferred to the nucleus in Fagaceae and has likely been lost in the plastid genomes of *Passiflora*. In the case of legumes, we have also identified copies of *rpl22* in the nucleus of *Glycine* and *Medicago* by Blast searches of genome sequences for these crop species. The overall organization of the nuclear-encoded *rpl22* in Fabaceae and Fagaceae is quite similar with each gene containing two exons with exon 1 serving as the transit peptide that facilitates targeting of the gene back to the plastid and exon 2 encoding the L22 ribosomal protein (fig. 6). Despite this similarity, the most likely interpretation of our results is that there have been at least two independent transfers of *rpl22* to the nucleus, one in Fabaceae, a second in Fagaceae, and possibly a third transfer in *Passiflora* (fig. 7). An alternative explanation is that there has

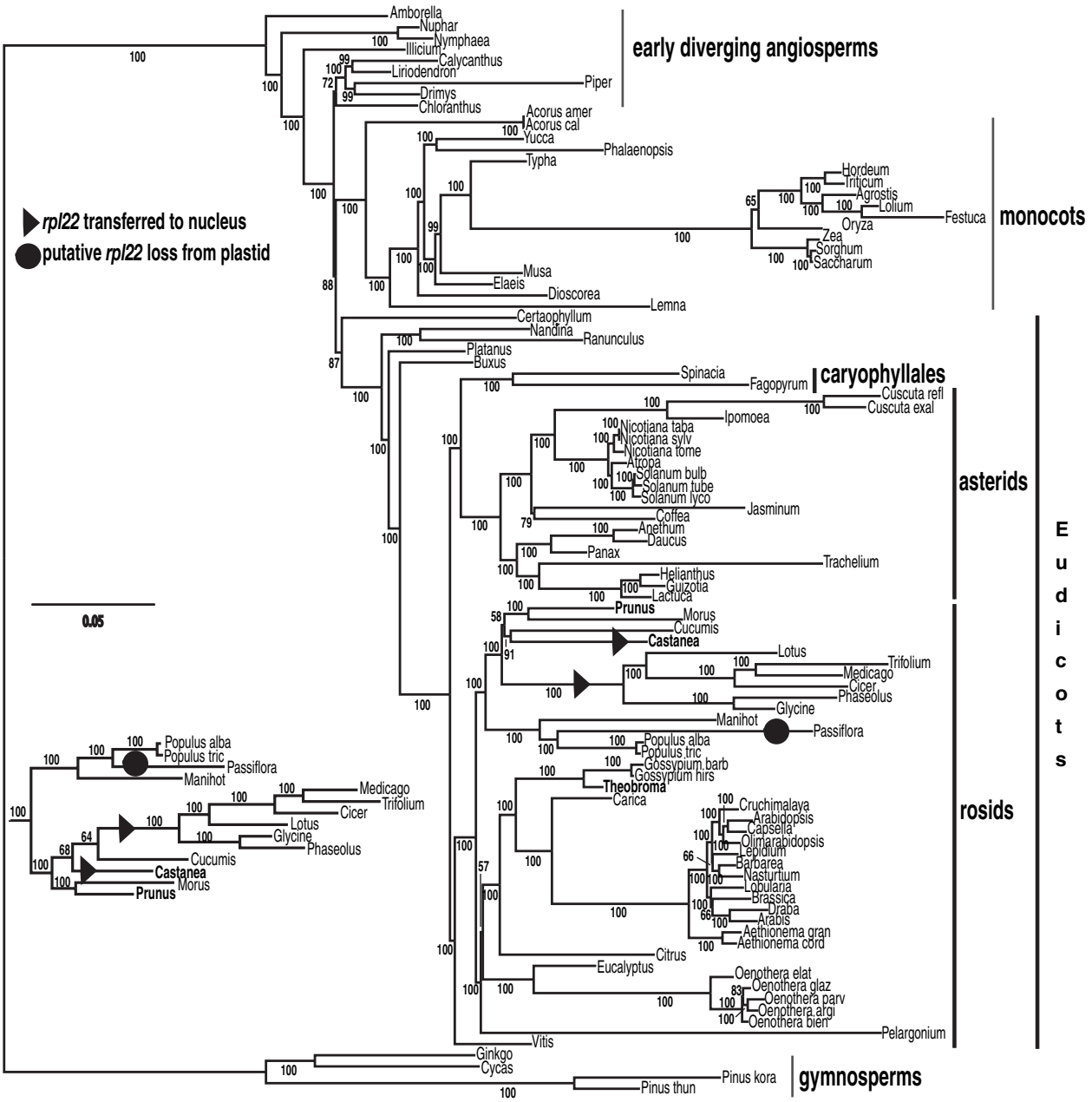

Fig. 7. ML phylogram of 97 taxa based on 81 plastid gene sequences. The tree has a −lnL of 1108190.01. Bootstrap support values >50% are provided at nodes, and major groups are angiosperms are labeled and following APG III (Angiosperm Phylogeny Group [APG] 2009). The three new rosid genomes reported here are highlighted in bold. Triangles on nodes indicate nuclear transfers and circles indicate putative losses of *rpl22*. Scale bar indicates the increment of 0.05 substitutions/site. Inset tree shows the MP topology for eurosid I taxa.

been only a single transfer in the ancestor of the eurosid I clade and that intact copies remain in most plastid genomes. Our data argue against this alternative because the Fabaceae and Fagaceae are not sister taxa and the transit peptides in exon 1 of the nuclear-encoded *rpl22* for Fabaceae and Fagaceae are highly divergent (29.2% amino acid identity) compared with the portion of the gene that came from the plastid in exon 2, which has a much higher sequence identity in comparison with other rosids (57.1%). More importantly, sequence identity of the transit peptides within each of the families (Fabaceae, 46.1% and Fagaceae, 78.5%) is much higher than between the families (29.2%). Although our Blast searches were not able to identify the origin of the transit peptide in either Fabaceae or Fagaceae, high sequence divergence between these two groups clearly supports their independent origin.

Gantt et al. (1991) argued that the nuclear transfer of *rpl22* likely occurred at least 100 Ma based on a phylogenetic analysis of this gene that placed the nuclear copy of *Pisum* outside all other angiosperms. They further suggested that the transfer might even be older, up to 200 Ma, because trees only one step longer place *Pisum* outside of all land plants. Analyses of the Gantt et al. (1991) were very limited in terms of taxon sampling due to the paucity

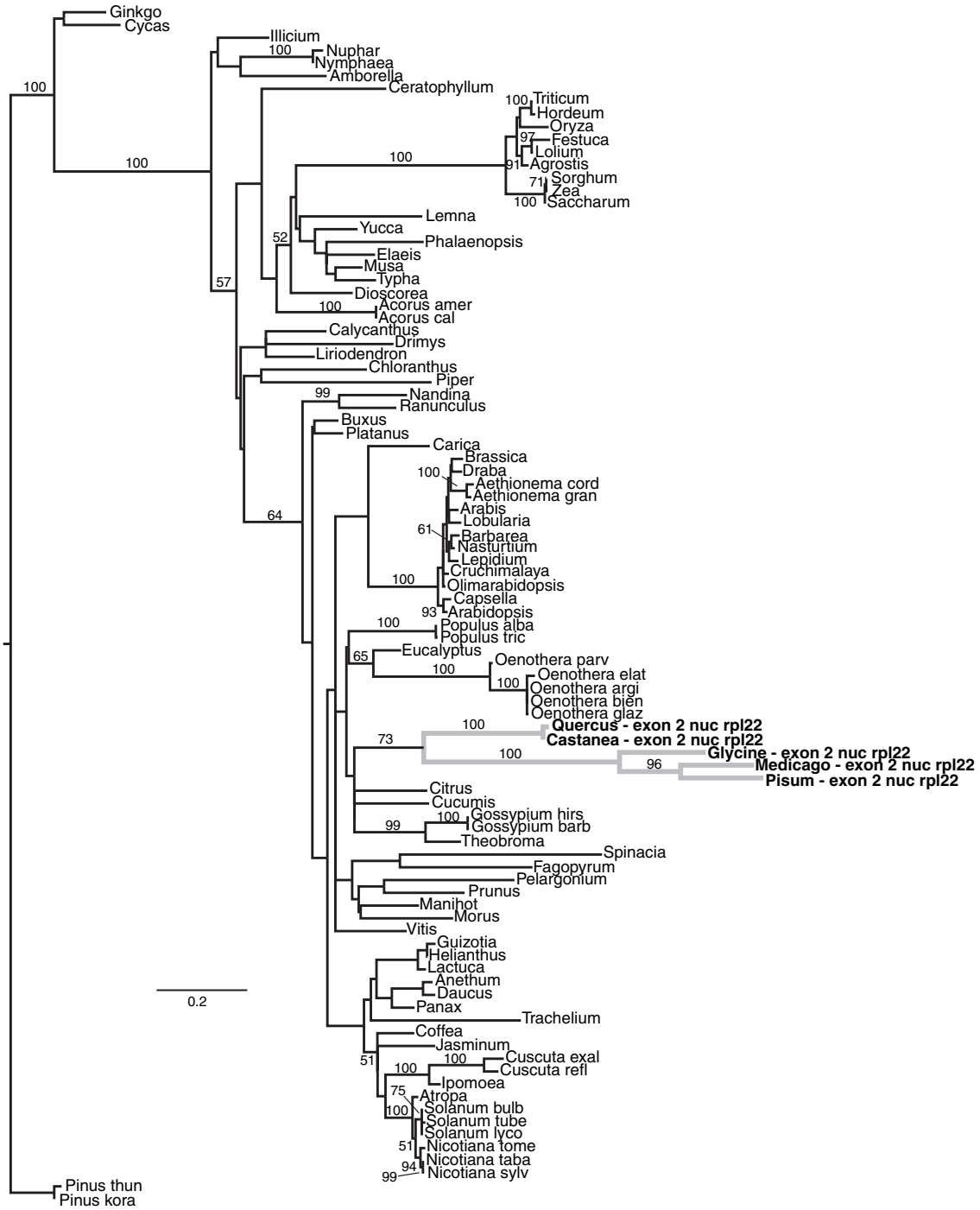

Fig. 8. ML phylogram of 91 taxa based on *rpl22* gene sequences. The tree has a −lnL of 13996.74. Bootstrap support values >50% are provided at nodes. Scale bar indicates the increment of 0.02 substitutions/site. Nuclear copies of *rpl22* in Fabaceae and Fagaceae are indicated in bold font.

of *rpl22* sequences available at that time. Their study only included seven land plant sequences, six of which were angiosperms. They also utilized four distant outgroups from algae and eubacteria. Thus, the placement of the nuclear copy of *Pisum* outside of angiosperms in Gantt et al. (1991) likely represents an artifact of limited taxon sampling, a well-known phenomenon that can lead to erroneous conclusions in phylogenetics (Pollock et al. 2002; Zwickl and Hillis 2002; Stefanovic et al. 2004; Leebens-Mack et al. 2005). Our phylogenetic analyses of *rpl22* for 94 taxa (fig. 8) clearly indicate that both the *Castanea* and *Pisum* nuclear copies of *rpl22* are nested within eudicots with some members of the rosid clade. Bootstrap support is not very strong in our analysis because the tree is based on only a single gene sequence and there are a large number of taxa. However, some internal nodes are moderately supported, indicating that the transfer of *rpl22* occurred much more recently than suggested by Gantt et al. (1991). We did

not estimate divergence times with the data presented in this paper because there have been several such studies during the past decade (Wikström et al. 2001; Davies et al. 2004; Magallón and Castillo 2009; Wang et al. 2009; Smith et al. 2010), and most of these are in general agreement concerning the time of origin of the major clades of angiosperms, especially those relevant to the timing of the *rpl22* transfer. Assuming that there were three independent transfers to the nucleus in the Fabaceae, Fagaceae, and Passifloraceae, these events would have occurred approximately 56–68 Ma, 34–37 Ma, and 26–27 Ma based on the range of divergence times for each family, respectively (Wikström et al. 2001). However, if the alternative less likely scenario is correct and the transfer occurred in the ancestor of the eurosid I clade, then it would have occurred much earlier, approximately 94–105 Ma (Magallón and Castillo 2009).

Despite high rates of transfer of plastid DNA to the nucleus (Timmis et al. 2004; Matsuo et al. 2005;Noutsos et al. 2005), very few successful functional transfers of genes have been documented. The reason for this is that transfer of functional copies of genes requires a series of unlikely events, including the acquisition of the required nuclear machinery to regulate transcription and a transit peptide to target the product back to the plastid. The three well-characterized transfers in land plants, *infA*, *rpoA*, and *rpl32*, used two different strategies. Both *infA* and *rpoA* acquired a transit peptide de novo (Millen et al. 2001; Sugiura et al. 2003), and in the case of *infA*, the transfer happened at least 24 times independently. In contrast, the *rpl32* gene acquired its transit peptide by transferring into a duplicate copy of a nuclear gene (Cu–Zn superoxide dismutase) that was already targeted to the plastid (Cusack and Wolfe 2007; Ueda et al. 2007). Transfer of *rpl22* followed a similar strategy to *infA*. The two independent transfers in Fabaceae (*Glycine*, *Medicago*, and *Pisum*) and Fagaceae (*Castanea* and *Quercus*) appear to have acquired transit peptides de novo based on the very low sequence identity between these two proteins (29.2%).

Supplementary Material

Supplementary table 1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

Support for this work was provided by grants from National Sciecne Foundation (DEB-0717372 to R.K.J.) and National Institutes of Health GM 63879 and USDA 58-3611-7-610 (to H.D.). We thank Chris Blazier and two anonymous reviewers for critical comments on an earlier version of this manuscript.

References

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

Bausher MG, Singh ND, Lee S-B, Jansen RK, Daniell H. 2006. The complete chloroplast genome sequence of *Citrus sinensis* (L.) Osbeck var 'Ridge Pineapple': organization and phylogenetic relationships to other angiosperms. *BMC Plant Biol*. 6:21.

Bock R. 2007. Structure, function, and inheritance of plastid genomes. In: Bock R, editor. Cell and molecular biology of plastids. Berlin (Germany): Springer–Verlag. p. 1610–2096.

Bubunenko MG, Schmidt J, Subramanian AR. 1994. Protein substitution in chloroplast ribosome evolution. A eukaryotic cytosolic protein has replaced its organelle homologue (L23) in spinach. *J Mol Biol*. 240:28–41.

Cai Z, Guisinger M, Kim HG, Ruck E, Blazier JC, McMurtry V, Kuehl JV, Boore J, Jansen RK. 2008. Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. *J Mol Evol*. 67:696–704.

Chang CC, Lin HC, Lin IP, Chow TY, Chen HH, Chen WH, Cheng CH, Lin CY, Liu SM, Chaw SM. 2006. The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. *Mol Biol Evol*. 23:279–291.

Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK. 2006. 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*. 23:2175–2190.

Cusack BP, Wolfe KH. 2007. When gene marriages don't work: divorce by subfunctionalization. *Trends Genet*. 23:270–272.

Daniell H, Lee S-B, Grevich J, Saski C, Guda C, Tompkins J, Jansen RK. 2006. Complete chloroplast genome sequences of *Solanum tuberosum*, *Solanum lycopersicum* and comparative analyses with other Solanaceae genomes. *Theor Appl Genet*. 112:1503–1518.

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

Doyle JJ, Doyle JL, Palmer JP. 1995. Multiple independent losses of two genes and one intron from legume chloroplast genomes. *Syst Bot*. 20:272–294.

Drummond AJ, Ashton B, Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T, Wilson A. 2009. Geneious v4.7. Available from: http://www.geneious.com/.

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

Emanuelsson O, Nielsen H, Brunak S, von Heijne G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol*. 300:1005–1016.

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

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

Funk HT, Berg S, Krupinska K, Maier UG, Krause K. 2007. Complete DNA sequences of the plastid genomes of two parasitic flowering plant species, *Cuscuta reflexa* and *Cuscuta gronovii*. *BMC Plant Biol*. 7:45.

Gantt JS, Baldauf SL, Calie PJ, Weeden NF, Palmer JD. 1991. Transfer of *rpl22* to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. *EMBO J*. 10:3073–3078.

Gornicki P, Faris J, King I, Podkowinski J, Gill B, Haselkorn R. 1997. Plastid localized acetyl-Co-A carboxylase of bread wheat is encoded by a single gene on each of the three ancestral chromosome sets. *Proc Natl Acad Sci U S A*. 94:14179–14184.

Guisinger MM, Chumley TW, Kuehl JV, Boore JL, Jansen RK. 2010. Implications of the plastid genome sequence of *Typha latifolia*

(Typhaceae, Poales) for understanding genome evolution in Poaceae. *J Mol Evol*. 70:149–166.

Guisinger MM, Kuehl JV, Boore JL, Jansen RK. 2008. Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. *Proc Natl Acad Sci U S A*. 105:18424–18429.

Haberle RC, Fourcade HM, Boore JL, Jansen RK. 2008. Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. *J Mol Evol*. 66:350–361.

Huang CY, Ayliffe MA, Timmis JN. 2003. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. *Nature*. 422:72–76.

Jansen RK, Cai Z, Raubeson LA, et al. 15 co-authors. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc Natl Acad Sci U S A*. 104:19369–19374.

Jansen RK, Raubeson LA, Boore JL, et al. (12 co-authors). 2005. Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods Enzymol*. 395:348–384.

Konishi T, Shinohara K, Yamada K, Sasaki Y. 1996. Acetyl-CoA carboxylase in higher plants: most plants other than gramineae have both prokaryotic and eukaryotic forms of this enzyme. *Plant Cell Physiol*. 37:117–122.

Lee HL, Jansen RK, Chumley TW, Kim KJ. 2007. Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. *Mol Biol Evol*. 24:1161–1180.

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

Leebens-Mack J, Raubeson LA, Cui L, Kuehl JV, Fourcade MH, Chumley TW, Boore JL, Jansen RK, dePamphilis CW. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein Zone. *Mol Biol Evol*. 22:1948–1963.

Magallón S, Castillo A. 2009. Angiosperm diversification through time. *Am J Bot*. 96:349–365.

Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci U S A*. 99:12246–12251.

Matsuo M, Ito Y, Yamauchi R, Obokata J. 2005. The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast–nuclear DNA flux. *Plant Cell*. 17:665–675.

McNeal JR, Kuehl JV, Boore JL, dePamphilis CW. 2007. Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus *Cuscuta*. *BMC Plant Biol*. 5:55.

Millen RS, Olmstead RG, Adams KL, et al. (10 co-authors). 2001. Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. *Plant Cell*. 13:645–658.

Moore MJ, Bell CD, Soltis PS, Soltis DE. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc Natl Acad Sci U S A*. 104:19363–368.

Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Folta KM, Soltis DE. 2006. Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biol*. 6:17.

Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc Natl Acad Sci U S A*. 107:4623–4628.

Noutsos C, Richly E, Leister D. 2005. Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants. *Genome Res*. 15:616–628.

Nylander JAA. 2004. MrModelTest v.2. Program distributed by the author. Sweden: Evolutionary Biology Centre, Uppsala University.

Pollock DD, Zwickl DJ, McGuire JA, Hillis DM. 2002. Increased taxon sampling is advantageous for phylogenetic inference. *Syst Biol*. 51:664–671.

Raubeson LA, Jansen RK. 2005. Chloroplast genomes of plants. In: Henry RJ, editor. Plant diversity and evolution: genotypic and phenotypic variation in higher plants. Wallingford (UK): CAB International.

Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM, Boore JL, Jansen RK. 2007. Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. *BMC Genomics*. 8:174.

Ruhlman T, Samson N, Verma D, Daniell H. 2010. The role of heterologous chloroplast sequence elements in transgene integration and expression. *Plant Physiol*. 152:2088–2104.

Saski C, Lee S-B, Fjellheim S, Guda C, Jansen RK, Luo H, Tomkins J, Rognli OA, Daniell H, Clarke JL. 2007. Complete chloroplast genome sequences of *Hordeum vulgare*, *Sorghum bicolor* and *Agrostis stolonifera* and comparative analyses with other grass genomes. *Theor Appl Genet*. 115:571–590.

Small I, Peeters N, Legeai F, Lurin C. 2004. Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* 4:1581–1590.

Smith SA, Beaulieu JM, Donoghue MJ. 2010. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. *Proc Natl Acad Sci U S A*. 107:5897–5902.

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

Stegemann S, Hartmann S, Ruf S, Bock R. 2003. High-frequency gene transfer from the chloroplast genome to the nucleus. *Proc Natl Acad Sci U S A*. 100:8828–8833.

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

Swofford D. 2003. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.

Timmis JN, Aliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet*. 5:123–135.

Ueda M, Fujimoto M, Arimura S-I, Murata J, Tsutsumi N, Kadowaki K-I. 2007. Loss of the *rpl32* gene from the chloroplast genome and subsequent acquisition of a preexisting transit peptide within the nuclear gene in *Populus*. *Gene* 402:51–56.

Ueda M, Fujimoto M, Takanashi H, Arimura S-I, Tsutsumi N, Kadowaki K-I. 2008. Substitution of the gene for chloroplast *rps16* was assisted by generation of dual targeting signal. *Mol Biol Evol*. 25:1566–1575.

Verma D, Daniell H. 2007. Chloroplast vector systems for biotechnology applications. *Plant Physiol*. 145:1129–1143.

Wang H, Moore MJ, Soltis PS, Bell CD, Brockington SF, Alexandre R, Davis CC, Latvis M, Manchester SR, Soltis DE. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proc Natl Acad Sci U S A*. 106:3853–3858.

Wikström N, Savolainen V, Chase MW. 2001. Evolution of the angiosperms: calibrating the family tree. *Proc R Soc Lond B Biol Sci*. 268:2211–2220.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20:3252–3255.

Zhong B, Yonezawa T, Zhong Y, Hasegawa M. 2009. Episodic evolution and adaptation of chloroplast genomes in ancestral grasses. *PloS ONE*. 4:e529.

Zwickl DJ. 2006. GARLI: Genetic Algorithm for Rapid Likelihood Inference, version 0.951. Available from: http://www.bio.utexas.edu/faculty/antisense/garli/Garli.html.

Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. *Syst Biol*. 51:588–598.