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

Robert K. Jansen1 Christopher Saski,2 Seung-Bum Lee,3 Anne K. Hansen,1 and Henry Daniell*3

1Section of Integrative Biology and Institute of Cellular and Molecular Biology, The University of Texas at Austin
2Clemson University, Genomics Institute
3Department of Molecular Biology and Microbiology, University of Central Florida

*Corresponding author: E-mail: daniell@mail.ucf.edu.
Associate editor: Charles Delwiche

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...
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 (Sugita et al. 2003). The loss of rps16 in Medicago truncatula and Populus albida 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 HindIII fingerprinting and NotI 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 hydroshaving, 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 pBluescript 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 containing 48,335 contigs, 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 HindIII fingerprinting and NotI 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.
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, Biosciences).

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.ccb.baylor.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, fig. 2). 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.

**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.ccb.baylor.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, fig. 2). 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.
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

![Circularized gene map of the plastid genomes of three rosids.](image-url)

**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.
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.

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 = 0.03/0.0, 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.

---

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

<table>
<thead>
<tr>
<th></th>
<th>Castanea mollissima (HQ336406)</th>
<th>Prunus persica (HQ336405)</th>
<th>Theobroma cacao (HQ336404)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (bp)</td>
<td>160,799</td>
<td>157,790</td>
<td>160,604</td>
</tr>
<tr>
<td>LSC length (bp)</td>
<td>90,432</td>
<td>85,968</td>
<td>89,395</td>
</tr>
<tr>
<td>SSC length (bp)</td>
<td>18,995</td>
<td>19,060</td>
<td>20,187</td>
</tr>
<tr>
<td>IR length (bp)</td>
<td>25,686</td>
<td>26,381</td>
<td>25,511</td>
</tr>
<tr>
<td>Number of genes</td>
<td>127</td>
<td>128</td>
<td>128</td>
</tr>
<tr>
<td>Number of gene with introns (with 2 introns)</td>
<td>12 (3)</td>
<td>18 (3)</td>
<td>18 (3)</td>
</tr>
<tr>
<td>GC content</td>
<td>36.8%</td>
<td>36.8%</td>
<td>36.9%</td>
</tr>
</tbody>
</table>

**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.
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 \geq 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 $-\ln L = 1108190.01$ (fig. 7). ML bootstrap values were also high, with values of \geq 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* 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*...
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 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 lnL = 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
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=72759&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, Fabaceae (*Castanea, Quercus*), Passifloraceae (*Passiflora*), Rosaceae (*Prunus*), and Malvaceae (*Theobroma*), have demonstrated that *rpl22* has also been transferred to the nucleus in Fabaceae 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...
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.
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

![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.](image-url)
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; Magallon 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 eudicots clade, then it would have occurred much earlier, approximately 94–105 Ma (Magallon 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; Sugiyama 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 Science 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


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


