3-2007

The Complete Nucleotide Sequence of the Coffee (Coffea Arabica L.) Chloroplast Genome: Organization and Implications for Biotechnology and Phylogenetic Relationships Amongst Angiosperms

Nalapalli Samson

Michael G. Bausher

Seung-Bum Lee

Robert K. Jansen

Henry Daniell
University of Pennsylvania

Recommended Citation

Samson, N., Bausher, M. G., Lee, S., Jansen, R. K., & Daniell, H. (2007). The Complete Nucleotide Sequence of the Coffee (Coffea Arabica L.) Chloroplast Genome: Organization and Implications for Biotechnology and Phylogenetic Relationships Amongst Angiosperms. *5* (2), 339-353. http://dx.doi.org/10.1111/j.1467-7652.2007.00245.x

At the time of publication, author Henry Daniell was affiliated with the University of Central Florida. Currently, he is a faculty member at the School of Dental Medicine at the University of Pennsylvania

The Complete Nucleotide Sequence of the Coffee (Coffea Arabica L.) Chloroplast Genome: Organization and Implications for Biotechnology and Phylogenetic Relationships Amongst Angiosperms

Abstract

The chloroplast genome sequence of *Coffea arabica* L., the first sequenced member of the fourth largest family of angiosperms, Rubiaceae, is reported. The genome is 155 189 bp in length, including a pair of inverted repeats of 25 943 bp. Of the 130 genes present, 112 are distinct and 18 are duplicated in the inverted repeat. The coding region comprises 79 protein genes, 29 transfer RNA genes, four ribosomal RNA genes and 18 genes containing introns (three with three exons). Repeat analysis revealed five direct and three inverted repeats of 30 bp or longer with a sequence identity of 90% or more. Comparisons of the coffee chloroplast genome with sequenced genomes of the closely related family Solanaceae indicated that coffee has a portion of *rps19* duplicated in the inverted repeat and an intact copy of *infA*. Furthermore, whole-genome comparisons identified large indels (> 500 bp) in several intergenic spacer regions and introns in the Solanaceae, including *trnE* (UUC)–*trnT* (GGU) spacer, *ycf4*–*cemA* spacer, *trnI* (GAU) intron and *rrn5*–*trnR* (ACG) spacer. Phylogenetic analyses based on the DNA sequences of 61 protein-coding genes for 35 taxa, performed using both maximum parsimony and maximum likelihood methods, strongly supported the monophyly of several major clades of angiosperms, including monocots, eudicots, rosids, asterids, eurosids II, and euasterids I and II. *Coffea* (Rubiaceae, Gentianales) is only the second order sampled from the euasterid I clade. The availability of the complete chloroplast genome of coffee provides regulatory and intergenic spacer sequences for utilization in chloroplast genetic engineering to improve this important crop.

Keywords

chloroplast genetic engineering, chloroplast genome, coffee, phylogeny, Rubiaceae

Disciplines

Dentistry

Comments

At the time of publication, author Henry Daniell was affiliated with the University of Central Florida. Currently, he is a faculty member at the School of Dental Medicine at the University of Pennsylvania

Published in final edited form as:
Plant Biotechnol J. 2007 March ; 5(2): 339–353. doi:10.1111/j.1467-7652.2007.00245.x.

The complete nucleotide sequence of the coffee (*Coffea arabica* L.) chloroplast genome: organization and implications for biotechnology and phylogenetic relationships amongst angiosperms

Nalapalli Samson[1], **Michael G. Bausher**[2], **Seung-Bum Lee**[1], **Robert K. Jansen**[3], and **Henry Daniell**[1,*]

[1]University of Central Florida, Department of Molecular Biology and Microbiology, Biomolecular Science, Building #20, Orlando, FL 32816-2364, USA

[2]USDA-ARS, Horticultural Research Laboratory, Fort Pierce, FL 34945-3030, USA

[3]Section of Integrative Biology and Institute of Cellular and Molecular Biology, Patterson Laboratories 141, University of Texas, Austin, TX 78712, USA

Summary

The chloroplast genome sequence of *Coffea arabica* L., the first sequenced member of the fourth largest family of angiosperms, Rubiaceae, is reported. The genome is 155 189 bp in length, including a pair of inverted repeats of 25 943 bp. Of the 130 genes present, 112 are distinct and 18 are duplicated in the inverted repeat. The coding region comprises 79 protein genes, 29 transfer RNA genes, four ribosomal RNA genes and 18 genes containing introns (three with three exons). Repeat analysis revealed five direct and three inverted repeats of 30 bp or longer with a sequence identity of 90% or more. Comparisons of the coffee chloroplast genome with sequenced genomes of the closely related family Solanaceae indicated that coffee has a portion of *rps19* duplicated in the inverted repeat and an intact copy of *infA*. Furthermore, whole-genome comparisons identified large indels (> 500 bp) in several intergenic spacer regions and introns in the Solanaceae, including *trnE* (UUC)–*trnT* (GGU) spacer, *ycf4–cemA* spacer, *trnI* (GAU) intron and *rrn5–trnR* (ACG) spacer. Phylogenetic analyses based on the DNA sequences of 61 protein-coding genes for 35 taxa, performed using both maximum parsimony and maximum likelihood methods, strongly supported the monophyly of several major clades of angiosperms, including monocots, eudicots, rosids, asterids, eurosids II, and euasterids I and II. *Coffea* (Rubiaceae, Gentianales) is only the second order sampled from the euasterid I clade. The availability of the complete chloroplast genome of coffee provides regulatory and intergenic spacer sequences for utilization in chloroplast genetic engineering to improve this important crop.

Keywords

chloroplast genetic engineering; chloroplast genome; coffee; phylogeny; Rubiaceae

[*]*Correspondence* (fax 407-823-0956; daniell@mail.ucf.edu).

Introduction

Coffee is one of the most economically important crops in the world. Approximately 2.25 billion cups of coffee are consumed in the world on a daily basis. In many years, this crop is second in value only to oil as a source of foreign exchange to several developing countries. Worldwide, an estimated 15 billion coffee trees are grown on 100 000 km^2 of land. Coffee is a member of Rubiaceae, the fourth largest family of angiosperms (Stevens, 2006). Most coffee beans come from two species of *Coffea*. *Coffea arabica* is considered to have the best quality and aroma, but this species is highly susceptible to several pathogens, including the fungus *Fusarium oxysporum*, nematodes (mainly *Meloidogyne* sp. and *Pratylenchus* spp.), coffee rust (*Hemileia vastatrix*), coffee stem borers (beetles in the family Cerambycidae) and coffee berry borer (*Hypothenemus hampei*). *Coffea canephora* is resistant to most of these pests, but the quality of the beans is poor. There is a need to improve the resistance of *C. arabica* to pests, which cause severe damage to coffee trees and substantial tree mortality in Africa, Asia and Latin America (USDA, 2005).

During the past 15 years, different research groups have successfully performed the genetic transformation of coffee. The earlier reports on coffee genetic transformation in the 1990s showed co-cultivation of *C. arabica* protoplasts with different strains of *Agrobacterium tumefaciens* using neomycin phosphotransferase type II gene (*NPT II*) selection and β-glucuronidase (GUS) marker genes (Spiral and Pétiard, 1991). Barton *et al*. (1991) obtained transformed somatic embryos of *C. arabica* by the electroporation method to integrate foreign genes. Van Boxtel *et al*. (1995) showed transient GUS expression in *C. arabica* using the biolistic method. Ocampo and Manzanera (1991) demonstrated that *C. arabica* tissue could be infected by wild strains of *A. tumefaciens*. Van Boxtel *et al*. (1997) evaluated the effectiveness of five selective agents (chlorsulphuron, glufosinate, glyphosate, hygromycin and kanamycin) for the selection of transformed embryogenic cell lines. Because of public and consumer concerns about the use of antibiotic-resistant marker genes in transformed plants, Penna *et al*. (2002) proposed several positive selection systems as alternative methods.

Despite considerable advances in coffee genetic transformation, the ability to introduce useful traits, such as insect resistance, did not occur until the turn of the century, when Leroy *et al*. (2000) expressed synthetic *Cry1Ac* to enhance leaf miner resistance. Ogita *et al*. (2003) obtained transgenic coffee plants with suppressed caffeine synthesis using RNA interference (RNAi) technology. They successfully produced low-caffeine plantlets of *C. canephora* through the down-regulation of *CaMXMT1* (theobromine synthase) and *CaDXMT1* (caffeine synthase), achieving 70% reduction of both theobromine and caffeine in the leaves compared with control plants. Another desirable trait of coffee quality is uniformity in fruit ripening, which has a major impact on the quality of coffee (Ribas *et al*., 2006). ACC oxidase (*Ca ACO*) genes promote uniform fruit ripening (Pereira *et al*., 2005). During the maturation of fruits, there is a dramatic increase in ethylene biosynthesis that promotes subsequent steps of fruit ripening with biochemical and physiological changes. Recently, the ACC oxidase gene involved in ethylene production has been cloned and characterized (Pereira *et al*., 2005). These genes are potential candidates for coffee transformation.

Crop improvement via interspecific and/or intraspecific breeding programmes is very time consuming because coffee is a perennial plant, and these programmes are impeded by the slow flowering time of 3–6 years. Because several backcrosses are needed to select for desirable traits, it takes a considerable time to develop disease-resistant trees. Despite the numerous backcrosses performed, unwanted chromosomal regions often remain in the offspring of interspecific crosses, reducing their survival. An alternative approach for

developing disease-resistant coffee plants may be chloroplast genetic engineering (Daniell *et al* ., 2004a,b, 2005; Grevich and Daniell, 2005). The presence of numerous copies of the plastid genome within chloroplasts favours a high level of expression of introduced foreign genes. Mendez-Lopez *et al.* (2003) recently demonstrated that a *Bt* toxin from *Bacillus thuringiensis* serovar *israelensis* is highly toxic to the first year instar of coffee berry borer larvae. They found that a paraporal crystal of *B. thuringiensis* serovar *israelensis* is lethal to the larvae at 219.5 ng/cm^2. Expression of this protein in the chloroplast should provide a very high level of toxicity to insects (Kota *et al.*, 1999; DeCosa *et al.*, 2001). The expression of antimicrobial peptides, MSI-99, via the chloroplast genome produced disease resistance of up to 88%–96% against bacterial (*Pseudomonas syringae*) and fungal (*Aspergillus flavus, Fusarium moniliforme and Verticillium*) pathogens (DeGray *et al.*, 2001). Multigene engineering in a single transformation event (Ruiz *et al.*, 2003; Quesada-Vargas *et al.*, 2005), transgene containment via maternal inheritance (Daniell *et al.*, 1998; Scott and Wilkinson, 1999; Daniell, 2002) or cytoplasmic male sterility (Ruiz and Daniell, 2005), lack of gene silencing (DeCosa *et al.*, 2001), elimination of position effect by site-specific transgene integration (Daniell *et al.*, 2002) and pleiotropic effects by subcellular compartmentalization of transgene products (Daniell *et al.*, 2001; Lee *et al.*, 2003; Leelavathi *et al.*, 2003) are other advantages offered by chloroplast genetic engineering.

Plastid genetic engineering should open up the possibility of enhancing the quality of coffee by the over-expression of enzymes, such as methionine synthase/cysteine synthase involved in methionine/cysteine production; increased levels of these amino acids in coffee are known to improve flavour (Carneiro, 1997). Naturally decaffeinated coffee will be of great interest to consumers who are concerned about the adverse health effects of caffeine intake. *N*-Methyltransferase (caffeine synthase) is a chloroplast enzyme that catalyses the last two steps in caffeine production through the conversion of 7-methylxanthine to theobromine and theobromine to caffeine (Ashihara and Crozier, 2001). Over-expression of the *N-7-demethylase* gene from *Coffea eugenioides* within plastids of transformed coffee plants is likely to produce plants with low caffeine content, as the *C. eugenioides* gene product will catalyse the metabolism of caffeine to theophylline, which is catabolized to CO_2 and NH_3 through the purine catabolic pathway.

To date, engineering of the chloroplast genome has been extended to only a few important crops, including carrot (Kumar *et al.*, 2004a), cotton (Kumar *et al.*, 2004b), lettuce Coffee chloroplast genome (Lelivelt *et al.*, 2005; Kanamoto *et al.*, 2006), potato (Sidorov *et al.*, 1999), poplar (Okumura *et al.*, 2006), rice (Lee S.M. *et al.*, 2006), soybean (Dufourmantel *et al.*, 2004, 2005) and tomato (Ruf *et al.*, 2001). In order to achieve reproducible plastid transformation in coffee, a knowledge of the complete plastid genome sequence is essential. This facilitates the identification of appropriate spacer regions for the integration of transgenes at optimal sites via homologous recombination, as well as endogenous regulatory sequences for the optimal expression of transgenes (Maier and Schmitz-Linneweber, 2004; Daniell *et al.*, 2005). These data also provide a rich source of nucleotide sequences for phylogenetic and molecular evolutionary questions (Goremykin *et al.*, 2003, 2004, 2005; Leebens-Mack *et al.*, 2005; Bausher *et al.*, 2006; Chang *et al.*, 2006; Jansen *et al.*, 2006; Lee S.B. *et al.*, 2006; Ruhlman *et al.*, 2006).

In this article, we report the complete sequence of the coffee chloroplast genome. In addition to discussing the biotechnological applications of this new genome sequence, we use these data to compare the genome organization and phylogenetic relationships of coffee to other angiosperms, especially the closely related family Solanaceae.

Results

Genome organization

The chloroplast genome of coffee is a circular DNA molecule of 155 189 bp with a quadripartite structure typical of the majority of land plant chloroplast chromosomes. It includes two inverted repeat regions (IRa and IRb) of 25 943 bp separated by large (LSC) and small (SSC) single-copy regions of 85 166 and 18 137 bp, respectively (Figure 1). The proportions of protein, transfer RNA (tRNA), ribosomal RNA (rRNA), intron and intergenic sequences are 51%, 2%, 6%, 9% and 32%, respectively. Of the 130 genes present in the genome (Table 1), 112 are present as a single copy and 18 are duplicated in the IR. The coding region includes 79 protein genes, 29 tRNAs and four rRNAs. The coffee chloroplast genome has 59.35% coding sequence, 51.76% of which codes for proteins. Eighteen genes contain introns, 15 with two exons and three with three exons. Twelve protein-coding genes and six tRNAs have introns (Table 2). A portion of the *rps19* gene is duplicated at the IRa–LSC boundary as a result of expansion of the IR. A similar duplication of portions of *rps19* occurs in all members of the related family Solanaceae except tobacco (Chung *et al.*, 2006). In addition, in the case of coffee, we observed that the *infA* gene is intact, whereas it is a pseudogene in tobacco and in most other members of the Solanaceae. The AT and GC contents of the coffee chloroplast genome are 63% and 37%, respectively, very similar to those of rice, maize, citrus, cotton and tobacco.

Repeat analysis revealed five direct repeats and three IRs of 30 bp or longer with sequence identity of 90% or more (Table 3, Figure 1). Three direct repeats and two IRs were located in intergenic spacer (IGS) regions. The remaining three repeats (one inverted and two direct) were located in the protein-coding genes *psaA*, *psaB*, *ycf1* and *ycf2*. The longest repeat of 64 bp occurred in the intergenic region between *rrn5* and *trnR* (ACG) within the IR (Figure 1). Comparative repeat analysis (CRA) of other representative eudicot crop plants identified eight (*C. arabica*), 12 (*Solanum lycopersicum* and *Vitis vinifera*), 14 (*Solanum bulbocastanum*), 15 (*Daucus carota*), 19 (*Citrus sinensis*), 25 (*Gossypium hirsutum*) and 39 (*Glycine max*) repeats.

There are 130 IGS regions ranging from 1 to 1620 bp in the coffee chloroplast genome, representing 31% of the genome. Eight IGS regions, *rps16–trnQ* (UUG), atpH–*atp I, rpoB–trnC* (GCA), *petN–petM, trnT* (GGU)–*psbD, ndhC–trnV* (UAC), *psbE*–petL and *rps12-3end–trnV* (GAC), are longer than 1000 bp, and all are located in the LSC, except for the *rps12* and *trnv* (GAC) region within the IR. Whole-genome sequence comparisons between coffee and four Solanaceae members using the MultiPipMaker program revealed several large (> 500 bp) deletions (red boxes in Figure 2). Three of these deletions within one IGS region [*rps16–trnQ* (UUG) and two introns (*trnI* (GAU) of IRa and IRb)] are present in all members of Solanaceae. A fourth deletion in the IGS region of *trnE* (UUC)–*trnT* (GGU) is only found in the two *Solanum* species, and a fifth located in the IGS region of *ycf4–cemA* is restricted to *Nicotiana tabacum*.

Phylogenetic relationships

Our data included 61 protein-coding genes for 35 taxa (Table 4), including 33 angiosperms and two gymnosperm outgroups (*Pinus* and *Ginkgo*). The data set comprised 46 437 aligned nucleotide positions but, when the gaps were excluded, there were 39 936 characters. Gaps were excluded to avoid alignment ambiguities, which were more pronounced for several divergent genes, including *ccsA*, *matK*, *rpoC2*, *rps15* and *rps18*.

Maximum parsimony (MP) analyses resulted in a single, fully resolved tree with a length of 61 797, a consistency index of 0.41 (excluding uninformative characters) and a retention index of 0.58 (Figure 3). Bootstrap analyses indicated that 24 of the 32 nodes were

supported by values of 95% or greater, and 23 of these had a bootstrap value of 100%. Of the remaining eight nodes, five had bootstrap values between 70% and 95%. Maximum likelihood (ML) analysis resulted in a single tree with −ln L = 348 679.23765 (Figure 4). ML bootstrap values were also high, with values of 95% or greater for 28 of the 32 nodes and 21 nodes with 100% bootstrap support. The remaining four nodes had bootstrap values of less than 75%. The ML and MP trees had very similar topologies (compare Figures 3 and 4). Agreement between the two trees included the position of *Amborella* alone as sister to the remaining angiosperms, the placement of the magnoliid genus *Calycanthus* sister to eudicots, the sister relationship between Caryophyllales and asterids, and strong support for the monophyly of several major clades, including monocots, eudicots, rosids, eurosids II, asterids, and euasterids I and II. The only incongruence between the MP and ML trees concerned relationships amongst the rosids. Specifically, the position of the clade that included *Cucumis* and the Myrtales (*Eucalyptus* and *Oenothera*) varied in the trees generated by the different methods, but support for alternative placements was weak. The MP tree (Figure 3) placed the *Cucumis*–Myrtales clade sister to a clade that includes the eurosid I taxa from the Fabales and Malphigiales. In contrast, the ML tree (Figure 4) placed the *Cucumis*–Myrtales clade sister to the eurosid II clade. Support for the different relationships of the *Cucumis*–Myrtales clade was < 50% in both MP and ML trees. In both analyses, members of eurosids I were not monophyletic. Both MP and ML trees provided very strong support (100% bootstrap) for the sister relationship of *Coffea* (Gentianales) and the Solanales in the euasterid I clade.

Discussion

Genome organization

The gene order of the coffee plastid genome is identical to the inferred ancestral angiosperm plastome organization, emphasizing the highly conserved nature of these genomes amongst land plants (Raubeson and Jansen, 2005). Our examination of the occurrence of repeated sequences in the coffee chloroplast genome confirms reports from several recent studies indicating that these genomes contain a number of direct repeats and IRs, even when genomes have not undergone rearrangements, including soybean (Saski *et al.*, 2005), cotton (Lee S.B. *et al.*, 2006), potato and tomato (Daniell *et al.*, 2006), grape (Jansen *et al.*, 2006), citrus (Bausher *et al.*, 2006) and carrot (Ruhlman *et al.*, 2006). In all of the plastid genomes examined so far, repeats are more prevalent in IGSs and introns, with fewer repeats located in the genes *ycf2*, *psaA* and *psaB*. Studies of repeats in highly rearranged genomes (Pombert *et al.*, 2005, 2006; Chumley *et al.*, 2006) have demonstrated a correlation between the number of repeats and the degree of gene order change, and many rearrangement endpoints have associated repeat elements. In these genomes, it is evident that repeated sequences have played a role in changes in gene order and content. The role of repeat elements in genomes that have not experienced rearrangements is unknown.

In the case of *C. arabica*, we identified fewer repeats (eight) than have been reported in a number of other crop plants, including *Daucus carota* (14; Ruhlman *et al.*, 2006), *Gossypium hirsutum* (54; Lee S.B. *et al.*, 2006), *Citrus sinensis* (29; Bausher *et al.*, 2006), *Vitis vinifera* (36; Jansen *et al.*, 2006), *Glycine max* (287; Saski *et al.*, 2005), *Solanum bulbocastanum* (31; Daniell *et al.*, 2006) and *Solanum lycopersicum* (40; Daniell *et al.*, 2006). The reason for this difference is that we used the CRA program instead of REPuter. REPuter uses pair-wise comparisons to identify repeats, and calculates the number of unique pairs, not the actual number of repeats. Thus, a repeat with multiple copies will be recorded multiple times. REPuter also over-estimates the number of repeats by recognizing several nested series of repeats within a given region containing multiple repeats. The use of an improved algorithm in the CRA program filters out these repeats and therefore more accurately identifies the number of repeated sequences. The number of repeats revealed by CRA was in

the range 8–39, rather than the range 8–287 reported in the published literature. The maximum number of repeats was identified in *Glycine max* (39), the fewest in *C. arabica*, and similar numbers of repeats were found in other crop plant chloroplast genomes (Table 5). The most probable explanation for the larger numbers of repeats in the *Glycine max* genome is that it contains some rearrangements, a common feature of legume genomes. Another possible explanation for the lower number of repeats in *C. arabica* could be related to A-T richness. It is widely accepted that repeats in chloroplast genomes tend to be A-T rich. Although we did not compare A-T richness in the crop plant genomes, this would be worthwhile in a more comprehensive, comparative study of chloroplast genomes.

Millen *et al*. (2001) observed that the *infA* gene, which codes for translation initiation factor 1, stands out as an unusually unstable angiosperm chloroplast gene, having been lost from the chloroplast genome on many separate occasions and transferred to the nucleus multiple times. In coffee, *infA* is intact, in contrast with the closely related family Solanaceae, where it is a pseudogene in tobacco and 17 other species examined. Millen *et al*. (2001) surveyed four genera of Rubiaceae, *Coffea*, *Galium*, *Ixora* and *Pentas*, using a combination of Southern hybridization with an *infA* probe and DNA sequencing. Two genera, *Coffea* and *Ixora*, had *infA*, whereas, in the other two genera, *Pentas* and *Galium*, the gene was absent in the chloroplast genome. Thus, there is variation within the Rubiaceae with regard to the presence or absence of *infA*.

For the expression of foreign proteins, genes are targeted into the spacer regions of the chloroplast genome for stable integration. The use of a 100% identity of the flanking sequence is optimal for stable integration of foreign genes by homologous recombination into the plastid genome. However, spacer regions are not 100% identical even in members of the same family. Recently, Daniell *et al*. (2006) examined the similarities between the sequenced chloroplast genomes of Solanaceae, and found that only four spacer regions had 100% identity amongst the four genomes studied. They also found that, between *Solanum lycopersicum* and *Solanum bulbocastanum*, 21 IGS regions had 100% identity, whereas only eight IGS regions had 100% sequence identity between *Solanum lycopersicum* and *Atropa belladonna*, and *Nicotiana tabacum* and *Solanum bulbocastanum*. In addition, several deletions and insertions were found in the IGS regions of *trnQ* (UUG)–*rps16, trnE* (UUC)–*trnT* (GGU), *trnK* (UUU)–*rps16, trnS* (GCU)–*trnG* (GCC), *ycf2*–*trnI* (CAU), *ycf 4*–*cemA* and *ycf15*–*trnL* (CAA) amongst the Solanaceae genomes examined. Analysis of the coffee chloroplast genome also identified deletions in the IGS regions of *rps16*–*trnQ* (UUG), *trnE* (UUC)–*trnT* (GGU), *ycf4*–*cemA* and in the *trnI* (GAU) intron relative to *Atropa belladonna, Solanum bulbocastanum*, *Nicotiana tabacum* and *Solanum lycopersicum* (Figure 2). A correlation between a low frequency of plastid transformation and the use of flanking sequences of lower sequence identity has been noted (Daniell *et al*., 2006). This study further highlights the importance of choosing IGS regions that have high sequence identity. Alternatively, species-specific vectors from the appropriate IGS region obtained from the coffee chloroplast genome could be used for successful integration of foreign genes into the chloroplast genome.

Phylogenetic relationships

The phylogenies based on 61 protein-coding chloroplast genes for 33 angiosperms (Figures 3 and 4) are congruent with most relationships suggested in recent phylogenies based on complete chloroplast genome sequences (Goremykin *et al*., 2003, 2004, 2005; Leebens-Mack *et al*., 2005; Chang *et al*., 2006; Jansen *et al*., 2006; Lee S.M. *et al*., 2006; Ruhlman *et al*., 2006). There is strong support for the monophyly of many major clades of angiosperms, including monocots, eudicots, rosids, asterids, eurosids II, asterids I and asterids II. Furthermore, our phylogenies, which include two additional asterids (*Coffea* and *Lactuca*), continue to strongly support a sister relationship between the asterids and Caryophyllales.

In the previous phylogenies based on complete genome sequences, several areas of incongruence were identified, including the identification of the basal angiosperm lineage, position of the magnoliids (represented by *Calycanthus*) and monophyly of the eurosid I clade. These incongruences have been attributed to two phenomena: limited taxon sampling and methods of phylogenetic reconstruction (Soltis and Soltis, 2004; Soltis *et al.*, 2004; Stefanovic *et al.*, 2004; Goremykin *et al.*, 2005; Leebens-Mack *et al.*, 2005; Lockhart and Penny, 2005; Martin *et al.*, 2005; Chang *et al.*, 2006; Jansen *et al.*, 2006). Our phylogenetic analyses, which include several additional genomes, are congruent with regard to the relationships in two of these areas. First, both MP and ML trees (Figures 3 and 4) indicate that *Amborella* alone is sister to the remaining angiosperms, whereas, in previous analyses (Leebens-Mack *et al.*, 2005; Jansen *et al.*, 2006), MP trees supported *Amborella* as the earliest diverging lineage and ML trees indicated that *Amborella* + Nympheales together was the most basal group. Second, in our analyses, the magnoliid *Calycanthus* is sister to the eudicots in both MP and ML trees (Figures 3 and 4). In previous whole-genome comparisons, *Calycanthus* was sister to eudicots in MP trees, but was sister to a clade that included monocots + eudicots in ML trees. Although both MP and ML trees agree with regard to the relationships between basal angiosperms and magnoliids, support for the relationships is not strong, especially in the ML tree (Figure 4). Additional taxon sampling of magnoliids is needed to resolve their phylogenetic position. The monophyly of the eurosid I clade continues to be controversial in phylogenies based on complete chloroplast genome sequences. MP analyses (see figure 4 in Jansen *et al.*, 2006) support the monophyly of this group, whereas ML trees do not. Our phylogenies using both methods indicate that the eurosid I clade is not monophyletic (Figures 3 and 4). Sampling of additional rosid taxa and genes from complete chloroplast genomes will be needed to resolve the relationships between these groups.

Finally, our phylogenetic analyses support the placement of *Coffea* (Rubiaceae, Gentianales) sister to the Solanales in the euasterid I clade. The position of the Rubiaceae and Gentianales is not controversial as recent single and multigene phylogenies already strongly support their placement in the same clade as the Solanales: the euasterid I clade (reviewed in Soltis *et al.*, 2005). However, given that euasterids I include 38 families and 35 000 species, much more taxon sampling is required to assess the relationships between the major clades.

In summary, this is the first report on the complete chloroplast genome sequence from a member of the Rubiaceae. The results provide essential information for the design of plastid transformation vectors for this economically important crop. However, several more members of this family must be sequenced before optimal (100% homologous) IGS regions are identified for use in transformation studies. Currently, any IGS region could be used to transform *C. arabica*, but not any other member of this family. One of the advantages of plastid transformation is the high level of transgene expression and foreign protein accumulation. This method has the potential to produce chloroplast-transformed coffee plants highly resistant to insect pests, such as leaf miners, nematodes and coffee berry borers. Over-expression of the caffeine-degrading enzyme *N*-7-demethylase compartmentalized in plastid-transformed plants of coffee would also enable the production of naturally decaffeinated coffee.

The MP and ML tree topologies strongly support the monophyly of several major clades of angiosperms, including monocots, eudicots, rosids, asterids, eurosids II, and euasterids I and II. The trees also provide support for relationships between several major clades, including the position of *Amborella* as the earliest diverging angiosperm lineage, a sister relationship between monocots and a clade including magnoliids and eudicots, the position of magnoliids sister to eudicots, and a sister relationship between Caryophyllales and asterids. The only

incongruence between MP and ML trees concerns the relationships amongst rosid clades, but the support of the different tree topologies is weak.

Experimental procedures

Plant material of *C. arabica* L. was obtained from Banana Tree Co. (Easton, PA, USA). Prior to chloroplast isolation, plants were kept in the dark for 2 days to reduce the levels of starch. Chloroplasts from leaves were isolated using the sucrose step gradient method of Palmer (1986) as modified by Jansen *et al.* (2005). About 10 g of leaf tissue was homogenized in Sandbrink isolation buffer (Sandbrink *et al.*, 1989) using prechilled tissue blender bursts at high speed for 5 s to obtain sufficient quantities of chloroplast. The homogenate was filtered using four layers of cheesecloth followed by one layer of miracloth (Calbiochem, La Jolla, CA, catalogue no. 475855) without squeezing. The filtrate was transferred to prechilled centrifuge tubes and centrifuged at 1000 ***g*** for 15 min at 4 °C. The pellets were resuspended in 7 mL of ice-cold wash buffer and gently loaded over the step gradient consisting of 18 mL of 52% sucrose overlayered with 7 mL of 30% sucrose. The sucrose step gradient was centrifuged at 76 800 ***g*** for 30–60 min at 4 °C in an SW-27 centrifuge (Beckman Coulter, CA). The chloroplast band from the 30%–52% interface was removed using a wide-bore pipette, diluted with 10 volumes of wash buffer and centrifuged at 1500 ***g*** for 15 min at 4 °C. Purified chloroplast pellets were resuspended in a final volume of 2 mL. The entire chloroplast genome was amplified by rolling circular amplification (RCA) using the Repli-g RCA Kit (Qiagen GmbH, Hilden, Germany), following the methods described by Jansen *et al.* (2005). RCA was performed at 30 °C for 16 h; the reaction was terminated with a final incubation at 65 °C for 10 min. Digestion of the RCA product with the restriction enzymes *Bst*XI, *EcoRI* and *Hin*dIII verified successful genome amplification, as well as the DNA quality for sequencing.

DNA sequencing and genome assembly

Purified RCA products were subjected to nebulization, followed by end repair, and size fractionated by agarose gel electrophoresis to obtain fragment lengths ranging from 2.0 to 3.5 kb. Repaired products were blunt-end cloned into pCR®-4Blunt-TOPO, followed by transformation into ElectroMax™ DH5 alpha cells by electroporation (TOPO® Shotgun Cloning Kit; Invitrogen, Carlsbad, CA, USA). Transformed cells were selected on Luria–Bertani (LB) agar containing 100 μg/μL ampicillin and arrayed into 30 × 96-well microtitre plates. Sequencing reactions were carried out in both the forward and reverse directions using the BigDye® Terminator v3.1 Cycle Sequencing Kit and separated by a 3730xL DNA Sequence Analyser (Applied Biosystems, Foster City, CA, USA). Sequence data were assembled using Sequencher version 4.5 (GeneCodes, Ann Arbor, MI, USA) following quality and vector trimming. Gap regions were filled by sequencing polymerase chain reaction (PCR) fragments generated from primers designed to flank the gaps. The assembly was considered to be complete when a sequence with a confidence score of 20, as judged by KB Basecaller software (Applied Biosystems), was accumulated at every base position with at least 4× coverage.

Gene annotation

Dual Organellar GenoMe Annotator (DOGMA) (Wyman *et al.*, 2004) was used to annotate the complete coffee chloroplast genome, after uploading a FASTA-formatted file of the complete nucleotide sequence to the program's server. BlastX and BlastN searches, against a custom database of previously published chloroplast genomes, identified putative protein-coding genes, and tRNAs or rRNAs. For genes with low sequence identity, manual annotation was performed, after identifying the position of the start and stop codons, as well as the translated amino acid sequence, using the chloroplast/bacterial genetic code.

Examination of repeat structure

Repeats were identified for the coffee genome using CRA (N. Holtshulte and S. Wyman (Williamstown, MA), unpubl. data; http://bugmaster.jgi-psf.org/repeats/). This program filters the redundant output of REPuter (Kurtz *et al.*, 2001) and identifies shared repeats amongst the input genomes. For repeat identification, the following constraints were set in CRA: a minimum repeat size of 30 bp and a Hamming distance of 3 (i.e. a sequence identity of 90% or more). Manual verification of the identified repeats was obtained using the program EDITSEQ of DNA star, whilst performing an intragenomic BLAST search of the identified repeat sequences.

Whole-genome sequence alignment

MultiPipMaker (Schwartz *et al.*, 2003; http://bio.cse.psu.edu) was used for whole-genome alignment of coffee with four published chloroplast genomes from the related family Solanaceae (*Atropa bellodonna*, NC_004561, Schmitz-Linneweber, 2002; *Solanum bulbocastanum*, NC_007943, Daniell *et al.*, 2006; *Nicotiana tabacum*, NC_001879, Shinozaki *et al.*, 1986; *Solanum lycopersicum*, DQ347959, Daniell *et al.*, 2006).

Phylogenetic analyses

The 61 genes included in the analyses of Goremykin *et al.* (2003), Leebens-Mack *et al.* (2005), Jansen *et al.* (2006), Lee S.B. *et al.* (2006) and Ruhlman *et al.* (2006) were extracted from the chloroplast genome sequence of *Coffea* using DOGMA (Wyman *et al.*, 2004). The same set of 61 genes was extracted from the chloroplast genome sequences of 34 other sequenced chloroplast genomes (see Table 4 for a complete list of the genomes examined). All 61 protein-coding genes of the 35 taxa were translated into amino acid sequences and aligned using MUSCLE (Edgar, 2004), followed by manual adjustments; the nucleotide sequences of these genes were aligned by constraining them to the aligned amino acid sequences. A Nexus file with character sets for phylogenetic analyses was generated after nucleotide sequence alignment had been completed. The complete nucleotide alignment is available online at: http://www.biosci.utexas.edu/IB/faculty/jansen/lab/research/data_files/index.htm.

Phylogenetic analyses using MP and ML were performed with PAUP* version 4.10 (Swofford, 2003) and GARLI version 0.942 (Zwickl, 2006). Phylogenetic analyses excluded gap regions to avoid ambiguities in regions with problematic alignment, which was especially needed for the more divergent genes *ccsA*, *matK*, *rpoC2*, *rps15* and *rps18*. All MP searches included 100 random addition replicates and tree bisection–reconnection (TBR) branch swapping with the Multrees option. MODELTEST 3.7 (Posada and Crandall, 1998) was used to determine the most appropriate model of DNA sequence evolution for the combined 61-gene data set. Hierarchical likelihood ratio tests and the Akaike information criterion were used to assess which of the 56 models best fitted the data, which was determined to be GTR + I + Γ by both criteria. ML analyses were performed using GARLI version 0.93 (Zwickl, 2006). Two independent GARLI runs were performed and the ML scores for the best tree were optimized in PAUP. Non-parametric bootstrap analyses (Felsenstein, 1985) were performed in PAUP for MP analyses with 1000 replicates with TBR branch swapping, one random addition replicate and the Multrees option, and for ML analyses with 100 replicates with nearestneighbour interchange (NNI) branch swapping, one random addition replicate and the Multrees option.

Acknowledgments

The investigations reported in this article were supported in part by grants from the United States Department of Agriculture (USDA 3611-21000-017-00D) and National Institutes of Health (NIH R01 GM 63879) to Henry Daniell and the National Science Foundation (NSF DEB 0120709) to Robert K. Jansen. The authors would like to thank Jerry Mozoruk for technical assistance in sample preparation, initial genome assembly and DNA preparation.

References

Asano T, Tsudzuki T, Takahashi S, Shimada H, Kadowaki K. Nucleotide sequence of the sugarcane (*Saccharum officinarum*) chloroplast genome: a comparative analysis of four monocot chloroplast genomes. DNA Res. 2004; 11:93–99. [PubMed: 15449542]

Ashihara H, Crozier A. Caffeine: a well known but little mentioned compound in plant science. Trends Plant Sci. 2001; 6:407–413. [PubMed: 11544129]

Barton, C.; Adam, TL.; Zaarowitz, MA. Stable transformation of foreign DNA into *Coffea arabica* plants. 14th International Conference on Coffee Science; San Francisco, CA, USA. Paris: ASIC (Association Scientifique Internationale du Café); 1991. p. 460-464.

Bausher MG, Singh ND, Lee S-B, Jansen RK, Daniell H. The complete chloroplast genome sequence of *Citrus sinensis* (L.) Osbeck var 'Ridge Pineapple': organization and phylogenetic relationships to other angiosperms. BMC Plant Biol. 2006; 6:21. [PubMed: 17010212]

Carneiro M. Coffee biotechnology and its application in genetic transformation. Euphytica. 1997; 96:167–172.

Chang C-C, Lin H-C, Lin I-P, Chow T-Y, Chen H-H, Chen W-H, Cheng C-H, Lin C-Y, Liu S-M, Chang C-C, Chaw S-M. The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. Mol. Biol. Evol. 2006; 23:279–291. [PubMed: 16207935]

Chumley TW, Palmer JD, Mower JP, Fourcade HM, Caile PJ, Boore JL, Jansen RK. The complete chloroplast genome sequence of *Pelargonium* × *hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Mol. Biol. Evol. 2006; 23:1–16. [PubMed: 16151190]

Chung H-J, Jung JD, Park H-W, Kim J-H, Cha HW, Min SR, Jeong W-J, Liu JR. The complete chloroplast genome sequence of *Solanum tuberosum* and comparative analysis with Solanaceae species identified the presence of a 241-bp deletion in cultivated potato chloroplast DNA sequence. Plant Cell Rep. 2006; 25:1369–1379. [PubMed: 16835751]

Daniell H. Molecular strategies for gene containment in transgenic crops. Nat. Biotechnol. 2002; 20:581–586. [PubMed: 12042861]

Daniell H, Datta R, Varma S, Gray S, Lee SB. Containment of herbicide resistance through genetic engineering of the chloroplast genome. Nat. Biotechnol. 1998; 16:345–348. [PubMed: 9555724]

Daniell H, Lee SB, Panchal T, Wiebe PO. Expression of the native cholera toxin B subunit gene and assembly as functional oligomers in transgenic tobacco chloroplasts. J. Mol. Biol. 2001; 311:1001–1009. [PubMed: 11531335]

Daniell H, Khan MS, Allison L. Milestones in chloroplast genetic engineering: an environmentally friendly era in biotechnology. Trends Plant Sci. 2002; 7:84–91. [PubMed: 11832280]

Daniell, H.; Cohill, P.; Kumar, S.; Dufourmantel, N. Chloroplast genetic engineering. In: Daniell, H.; Chase, C., editors. Molecular Biology and Biotechnology of Plant Organelles. Dordrecht: Kluwer Academic Publishers; 2004a. p. 423-468.

Daniell H, Ruiz O, Dhingra A. Chloroplast genetic engineering to improve agronomic traits. Methods Mol. Biol. 2004b; 286:111–138. [PubMed: 15310917]

Daniell H, Kumar S, Dufourmantel N. Breakthrough in chloroplast genetic engineering of agronomically important crops. Trends Biotechnol. 2005; 23:238–245. [PubMed: 15866001]

Daniell H, Lee SB, Grevich J, Saski C, Guda C, Tomkins J, Jansen RK. Complete chloroplast genome sequences of *Solanum bulbocastanum*, *Solanum lycopersicum* and comparative analyses with other Solanaceae genomes. Theor. Appl. Genet. 2006; 112:1503–1518. [PubMed: 16575560]

DeCosa B, Moar W, Lee SB, Miller M, Daniell H. Overexpression of the *Bt* Cry2Aa2 operon in chloroplasts leads to formation of insecticidal crystals. Nat. Biotechnol. 2001; 19:71–74. [PubMed: 11135556]

DeGray G, Rajasekaran K, Smith F, Sanford J, Daniell H. Expression of an antimicrobial peptide via the chloroplast genome to control phytopathogenic bacteria and fungi. Plant Physiol. 2001; 127:852–862. [PubMed: 11706168]

Dufourmantel N, Pelissier B, Garcon F, Peltier G, Ferullo JM, Tissot G. Generation of fertile transplastomic soybean. Plant Mol. Biol. 2004; 55:479–489. [PubMed: 15604694]

Dufourmantel N, Tissot G, Goutorbe F, Garcon F, Muhr C, Jansens S, Pelissier B, Peltier G, Dubald M. Generation and analysis of soybean plastid transformants expressing *Bacillus thuringiensis* Cry1Ab protoxin. Plant Mol. Biol. 2005; 58:659–658. [PubMed: 16158241]

Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004; 5:113. [PubMed: 15318951]

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

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH. Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. Mol. Biol. Evol. 2003; 20:1499–1505. [PubMed: 12832641]

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH. The chloroplast genome of *Nymphaea alba*: whole-genome analyses and the problem of identifying the most basal angiosperm. Mol. Biol. Evol. 2004; 21:1445–1454. [PubMed: 15084683]

Goremykin VV, Holland B, Hirsch-Ernst KI, Hellwig FH. Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. Mol. Biol. Evol. 2005; 22:1813–1822. [PubMed: 15930156]

Grevich JJ, Daniell H. Chloroplast genetic engineering: recent advances and future perspectives. Crit. Rev. Plant Sci. 2005; 24:83–107.

Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun CR, Meng BY, Li YQ, Kanno A, Nishizawa Y, Hirai A, Shinozaki K, Sugiura M. The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Mol. Gen. Genet. 1989; 217:185–194. [PubMed: 2770692]

Hupfer H, Swaitek M, Hornung S, Herrmann RG, Maier RM, Chiu WL, Sears B. Complete nucleotide sequence of the *Oenothera elata* plastid chromosome, representing plastome 1 of the five distinguishable *Euoenthera* plastomes. Mol. Gen. Genet. 2000; 263:581–585. [PubMed: 10852478]

Jansen RK, Boore JL, dePamphilis CW, Chumley TW, Haberle RC, Wyman SK, Alverson AJ, Peery R, Herman SJ, Fourcade HM, Kuehl JV, McNeal JR, Leebens-Mack J, Cui L. Methods for obtaining and analyzing whole chloroplast genome sequences. Methods Enzymol. 2005; 395:348–384. [PubMed: 15865976]

Jansen RK, Kaittanis C, Saski C, Lee SB, Tomkins J, Alverson AJ, Daniell H. Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evol. Biol. 2006; 6:32. [PubMed: 16603088]

Kanamoto H, Yamashita A, Asao H, Okumura S, Takase H, Hattori M, Akiho Yokota A, Tomizawa K. Efficient and stable transformation of *Lactuca sativa* L. cv. Cisco (lettuce) plastids. Transgenic Res. 2006; 15:205–217. [PubMed: 16604461]

Kato T, Kaneko T, Sato S, Nakamura Y, Tabata S. Complete structure of the chloroplast genome of a legume, *Lotus japonicus*. DNA Res. 2000; 7:323–330. [PubMed: 11214967]

Kim KJ, Lee HL. Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Res. 2004; 11:247–261. [PubMed: 15500250]

Kota M, Daniell H, Varma S, Garczynski SF, Gould F, Moar WJ. Overexpression of the *Bacillus thuringiensis* (Bt) Cry2Aa2 protein in chloroplasts confers resistance to plants against susceptible and Bt-resistant insects. Proc. Natl Acad. Sci. USA. 1999; 96:1840–1845. [PubMed: 10051556]

Kumar S, Dhingra A, Daniell H. Chloroplast-expressed betaine aldehyde dehydrogenase gene in carrot cultured cells, roots, and leaves confers enhanced salt tolerance. Plant Physiol. 2004a; 136:2843–2854. [PubMed: 15347789]

Kumar S, Dhingra A, Daniell H. Stable transformation of the cotton chloroplast genome and maternal inheritance of transgenes. Plant Mol. Biol. 2004b; 56:203–216. [PubMed: 15604738]

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001; 29:4633–4642. [PubMed: 11713313]

Lee S-B, Kwon H, Kwon S, Park S, Jeong M, Han S, Daniell H. Accumulation of trehalose within transgenic chloroplast confers drought tolerance. Mol. Breed. 2003; 11:1–13.

Lee S-B, Kaittanis C, Jansen RK, Hostetler JB, Tallon LJ, Town CD, Daniell H. The complete chloroplast genome sequence of *Gossypium hirsutum*: organization and phylogenetic relationships to other angiosperms. BMC Genomics. 2006; 7:61. [PubMed: 16553962]

Lee SM, Kang K, Chung H, Yoo SJ, Xu XM, Lee S-B, Cheong JJ, Daniell H, Kim M. Plastid transformation in the monocotyledonous cereal crop, rice (*Oryza sativa*), and transmission of transgenes to their progeny. Mol. Cells. 2006; 21:401–410. [PubMed: 16819304]

Leebens-Mack J, Raubeson LA, Cui L, Kuehl J, Fourcade M, Chumley T, Boore JL, Jansen RK, dePamphilis CW. Identifying the basal angiosperms in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Mol. Biol. Evol. 2005; 22:1948–1963. [PubMed: 15944438]

Leelavathi S, Naveen Gupta N, Maiti S, Ghosh A, Reddy VS. Overproduction of an alkali- and thermostable xylanase in tobacco chloroplasts and efficient recovery of the enzyme. Mol. Breed. 2003; 11:59–67.

Lelivelt C, McCabe M, Newell C, deSnoo C, Dun K, Birch-Machin I, Gray J, Mills K, Nugent J. Stable plastid transformation in lettuce (*Lactuca sativa* L.). Plant Mol. Biol. 2005; 58:763–774. [PubMed: 16240172]

Leroy T, Henry A-M, Royer M, Altosaar I, Frutos R, Duris D, Philippe R. Genetically modified coffee plants expressing the *Bacillus thuringiensis* cry1Ac gene for resistance to leaf miner. Plant Cell Rep. 2000; 19:382–389.

Lockhart PJ, Penny D. The place of *Amborella* within the radiation of angiosperms. Trends Plant Sci. 2005; 10:201–202. [PubMed: 15882650]

Maier, RM.; Schmitz-Linneweber. Chloroplast genomes. In: Daniell, H.; Chase, C., editors. Molecular Biology and Biotechnology of Plant Organelles. The Netherlands: Springer; 2004. p. 115-150.

Maier RM, Neckermann K, Igloi GL, Kossel H. Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. J. Mol. Biol. 1995; 251:614–628. [PubMed: 7666415]

Martin W, Deusch O, Stawski N, Grunheit N, Goremykin V. Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci. 2005; 10:203–209. [PubMed: 15882651]

Mendez-Lopez I, Basurto-Rios R, Ibarra JE. *Bacillus thuringiensis* serovar *israelensis* is highly toxic to the coffee berry borer, *Hypothenemus hampei* Ferr. (Coleoptera: Scolytidae). FEMS Microbiol. Lett. 2003; 226:73–77. [PubMed: 13129610]

Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Hibberd JM, Gray JC, Morden CW, Calie PJ, Jermiin LS, Wolfe KH. Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. Plant Cell. 2001; 13:645–658. [PubMed: 11251102]

Ocampo, C.; Manzanera, LM. Advances in genetic manipulation of coffee plant. 14th International Conference on Coffee Science; San Francisco, CA, USA. Paris: ASIC (Association Scientifique Internationale du Café); 1991. p. 73-81.

Ogihara Y, Isono K, Kojima T, Endo A, Hanaoka M, Shiina T, Terachi T, Utsugi S, Murata M, Mori N, Takumi S, Ikeo K, Gojobori T, Murai R, Murai K, Matsuoka Y, Ohnishi Y, Tajiri H, Tsunewaki K. Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA. Mol. Gen. Genet. 2002; 266:740–746.

Ogita S, Uefuji H, Yamaguchi Y, Koizumi N, Sano H. RNA interference producing decaffeinated coffee plants. Nature. 2003; 423:823. [PubMed: 12815419]

Okumura S, Sawada M, Park YW, Hayashi T, Shimamura M, Takase H, Tomizawa K. Transformation of poplar (*Populus alba*) plastids and expression of foreign proteins in tree chloroplasts. Transgenic Res. 2006; 15:637–646. [PubMed: 16952016]

Palmer JD. Isolation and structural analysis of chloroplast DNA. Meth. Enzymol. 1986; 118:167–186.

Penna S, Saggi L, Swennen R. Positive selectable marker genes for routine plant transformation. In Vitro Cell Dev. Biol. Plant. 2002; 38:125–128.

Pereira LFP, Galvão RM, Kobayash AK, Cação SMB, Vieira LGE. Ethylene production and acc oxidase gene expression during fruit ripening of *Coffea arabica* L. Braz. J. Plant Physiol. 2005; 17:283–289.

Pombert J-F, Otis C, Lemieux C, Turmel M. The chloroplast genome sequence of the green alga *Pseudendoclonium akinetum* Ulvophyceae reveals unusual structural features and new insights into the branching order of chlorophyte lineages. Mol. Biol. Evol. 2005; 22:1903–1918. [PubMed: 15930151]

Pombert J-F, Lemieux C, Turmel M. The complete chloroplast DNA sequence of the green alga *Oltmannsiellopsis viridis* reveals a distinctive quadripartite architecture in the chloroplast genome of early diverging ulvophytes. BMC Biol. 2006; 4:3. [PubMed: 16472375]

Posada D, Crandall KA. Modeltest: testing the model of DNA substitution. Bioinformatics. 1998; 14:817–818. [PubMed: 9918953]

Quesada-Vargas T, Ruiz ON, Daniell H. Characterization of heterologous multigene operons in transgenic chloroplasts: transcription, processing, and translation. Plant Physiol. 2005; 138:1746–1762. [PubMed: 15980187]

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

Ribas AF, Pereira IIFP, Vieira LGE. Genetic transformation of coffee. BrazJ. Plant Physiol. 2006; 18:83–94.

Ruf S, Hermann M, Berger IJ, Carrer H, Bock R. Stable genetic transformation of tomato plastids and expression of a foreign protein in fruit. Nat. Biotechnol. 2001; 19:870–875. [PubMed: 11533648]

Ruhlman T, Lee SB, Jansen RK, Hostetler JB, Tallon LJ, Town CD, Daniell H. Complete plastid genome sequence of *Daucus carota*: implications for biotechnology and phylogeny of angiosperms. BMC Genomics. 2006; 7:224. [PubMed: 16948847]

Ruiz ON, Daniell H. Engineering cytoplasmic male sterility via the chloroplast genome by expression of betaketothiolase. Plant Physiol. 2005; 138:1232–1246. [PubMed: 16009998]

Ruiz ON, Hussein HS, Terry N, Daniell H. Phytoremediation of organomercurial compounds via chloroplast genetic engineering. Plant Physiol. 2003; 132:1344–1352. [PubMed: 12857816]

Sandbrink JM, Vellekoop P, Vanham R, Vanbrederode J. A method for evolutionary studies on RFLP of chloroplast DNA, applicable to a range of plant species. Biochem. Syst. Ecol. 1989; 17:45–49.

Saski C, Lee S-B, Daniell H, Wood T, Tomkins J, Kim H-G, Jansen R. Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. Plant Mol. Biol. 2005; 59:309–322. [PubMed: 16247559]

Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S. Complete structure of the chloroplast genome of *Arabidopsis thaliana*. DNA Res. 1999; 6:283–290. [PubMed: 10574454]

Schmitz-Linneweber C, Maier RM, Alcaraz JP, Cottet A, Herrmann RG, Mache R. The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization. Plant Mol. Biol. 2001; 45:307–315. [PubMed: 11292076]

Schmitz-Linneweber C, Du Regel RTG, Hupfer H, Herrmann RG, Maier RM. The plastid chromosome of *Atropa belladonna* and its comparison with that of *Nicotiana tabacum*: the role of RNA editing in generating divergence in the process of plant speciation. Mol. Biol. Evol. 2002; 19:1602–1612. [PubMed: 12200487]

Schwartz S, Elnitski L, Li M, Weirauch M, Riemer C, Smit A, Green ED, Hardison RC, Webb M. MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. Nucleic Acids Res. 2003; 31:3518–3524. [PubMed: 12824357]

Scott SE, Wilkinson MJ. Low probability of chloroplast movement from oilseed rape (*Brassica napus*) into wild *Brassica rapa*. Nat. Biotechnol. 1999; 17:390–392. [PubMed: 10207890]

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H, Sugiura M. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J. 1986; 5:2043–2049. [PubMed: 16453699]

Sidorov VA, Kasten D, Pang S-Z, Hajdukiewicz PTJM, Staub JM, Nehra NS. Stable chloroplast transformation in potato: use of green fluorescent protein as a plastid marker. Plant J. 1999; 19:209–216. [PubMed: 10476068]

Soltis DE, Soltis PS. *Amborella* not a 'basal angiosperm'? Not so fast. Am. J. Bot. 2004; 91:997–1001. [PubMed: 21653455]

Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu YL, Chase MW, Farris JS, Stefanovic S, Rice DW, Palmer JD, Soltis PS. Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. Trends Plant Sci. 2004; 9:477–483. [PubMed: 15465682]

Soltis, DE.; Soltis, PS.; Endress, PK.; Chase, MW. Phylogeny and Evolution of Angiosperms. Sunderland, MA: Sinauer Associates Inc.; 2005.

Spiral, J.; Pétiard, V. Protoplast culture and regeneration in Coffea species. 14th International Conference on Coffee Science; San Francisco, CA, USA. Paris: ASIC (Association Scientifique Internationale du Café); 1991. p. 383-391.

Steane DA. Complete nucleotide sequence of the chloroplast genome from the Tasmanian blue gum, *Eucalyptus globulus* (Myrtaceae). DNA Res. 2005; 12:215–220. [PubMed: 16303753]

Stefanovic S, Rice DW, Palmer JD. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? BMC Evol. Biol. 2004; 4:35. [PubMed: 15453916]

Stevens, PF. [accessed on 27 December 2006] Angiosperm Phylogeny Website. 2006. Version 7. URL http://www.mobot.org/MOBOT/research/APweb/

Swofford, DL. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4.0. Sunderland MA: Sinauer Associates; 2003.

Timme RE, Kuehl JV, Boore JL, Jansen RK. A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats. Am. J. Bot. 2007 in press.

USDA. Tropical Products: World Markets and Trade. Washington, DC: USDA, Foreign Agriculture Service; 2005.

Van Boxtel J, Berthouly M, Carasco C, Dufour M, Eskes A. Transient expression of β-glucuronidase following biolistic delivery of foreign DNA into coffee tissues. Plant Cell Rep. 1995; 14:748–752.

Van Boxtel J, Eskes A, Berthouly M. Glufosinate as an efficient inhibitor of callus proliferation in coffee tissue. In Vitro Cell. Dev. Biol. Plant. 1997; 33:6–12.

Wakasugi T, Tsudzuki J, Ito S, Nakashima K, Tsudzuki T, Sugiura M. Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. Proc. Natl Acad. Sci. USA. 1994; 91:9794–9798. [PubMed: 7937893]

Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with DOGMA. Bioinformatics. 2004; 20:3252–3255. [PubMed: 15180927]

Zwickl, DJ. [accessed on 4 July 2006] GARLI (Genetic Algorithm for Rapid Likelihood Inference). 2006. Version 0.942. URL http://www.bio.utexas.edu/grad/zwickl/web/garli.html

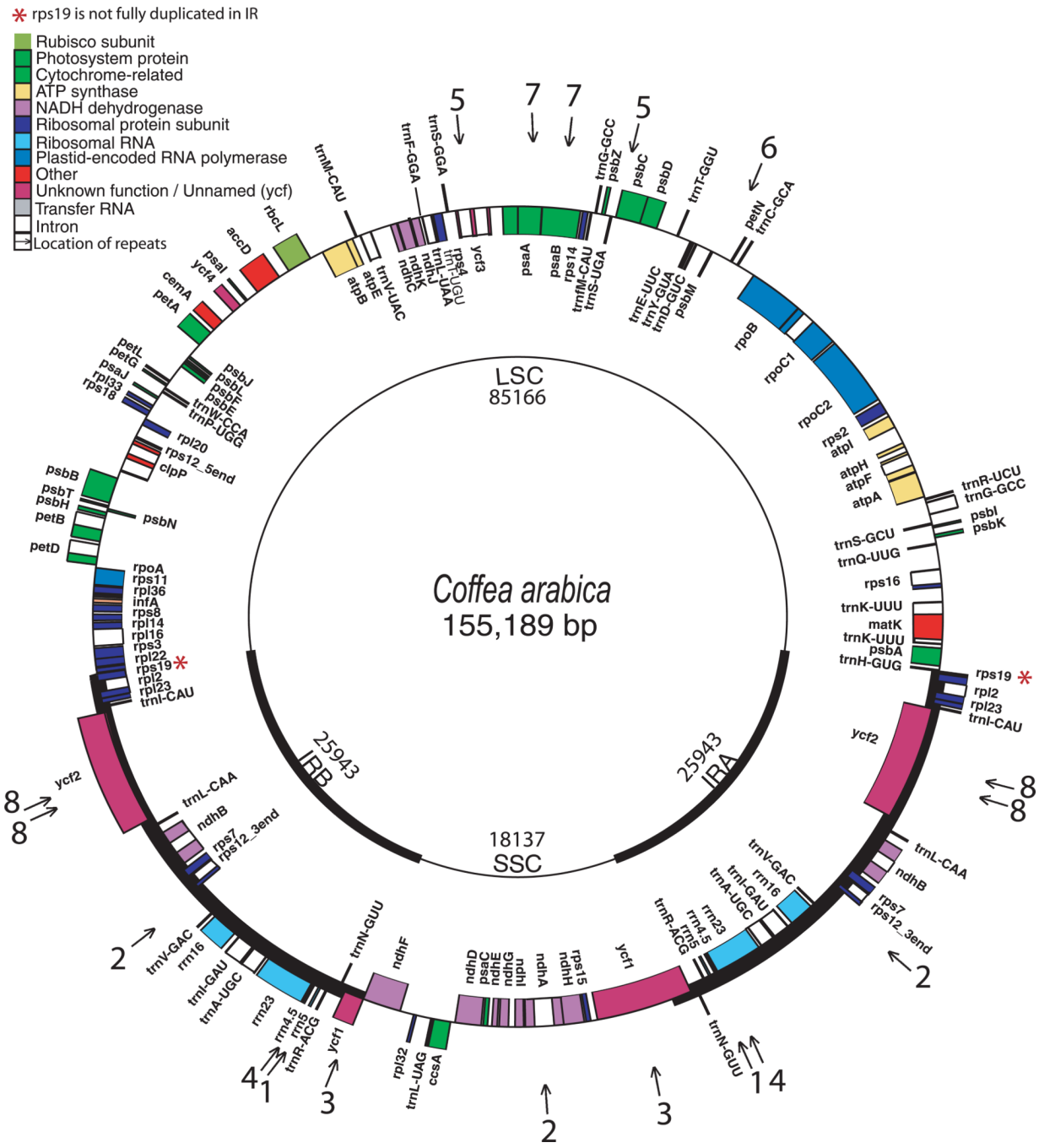

Figure 1.

Circular gene map of the *Coffea arabica* chloroplast genome. The thick lines indicate the extent of the inverted repeats (IRa and IRb, 25 943 bp), which separate the genome into small (SSC, 18 133 bp) and large (LSC, 85 166 bp) single-copy regions. Genes on the outside of the map are transcribed in the clockwise direction and genes on the inside of the map are transcribed in the counterclockwise direction. *The *rps19* gene locates entirely in the IRb region and partly in the IRa region. Arrows show the location of repeats (for more details on repeats, see Table 3).

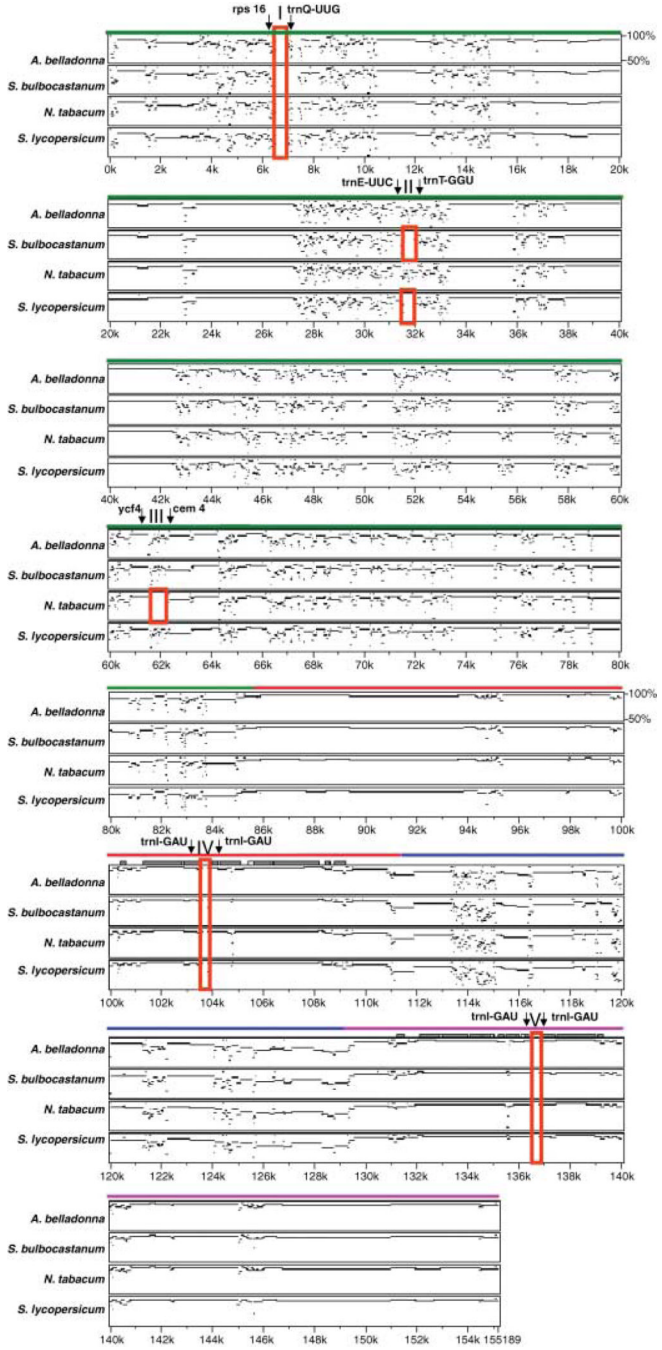

Figure 2.

The chloroplast genome comparison derived through a percentage identity plot of coffee against four Solanaceae members (*Atropa belladonna, Solanum bulbocastanum*, *Nicotiana tabacum* and *Solanum lycopersicum*) using the MultiPipMaker alignment tool. DNA losses are marked with roman numerals and the red boxes. I, region within intergenic spacer (IGS) [*rps16–trnQ* (UUG)]; II, region within IGS [*trnE* (UUC)–*trnI* (GGU)]; III, region within IGS (*ycf4–cemA*); IV, intron [IRb: *trnI* (GAU)]; V, intron [IRa: *trnI* (GAU)].

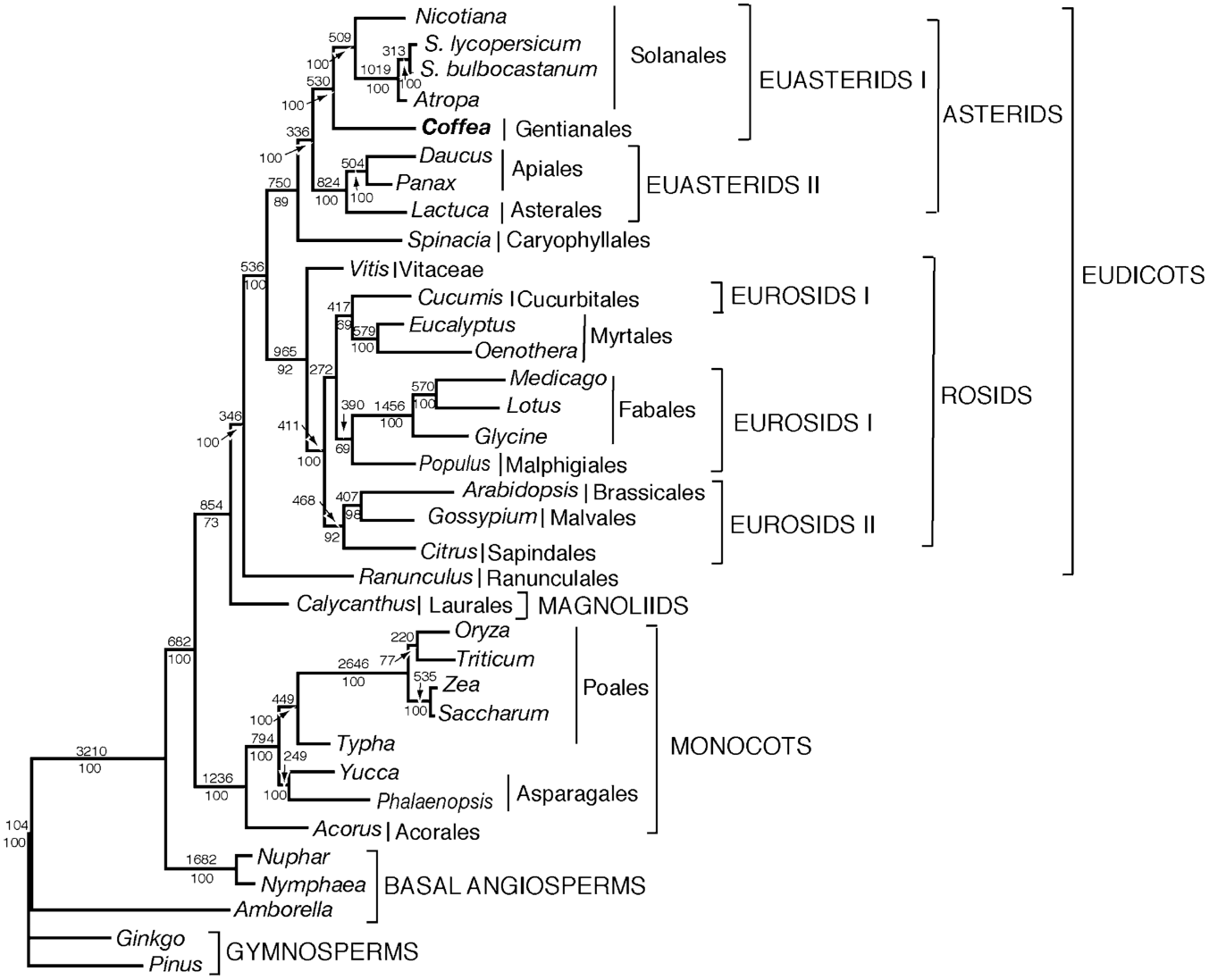

Figure 3.

Maximum parsimony tree based on 61 chloroplast protein-coding genes (data are available at http://www.biosci.utexas.edu/IB/faculty/jansen/lab/research/datafiles/index.htm). The single most parsimonious phylogram has a length of 61 797, a consistency index of 0.41 (excluding uninformative characters) and a retention index of 0.58. Numbers above and below the nodes indicate the number of nucleotide substitutions and bootstrap support values, respectively.

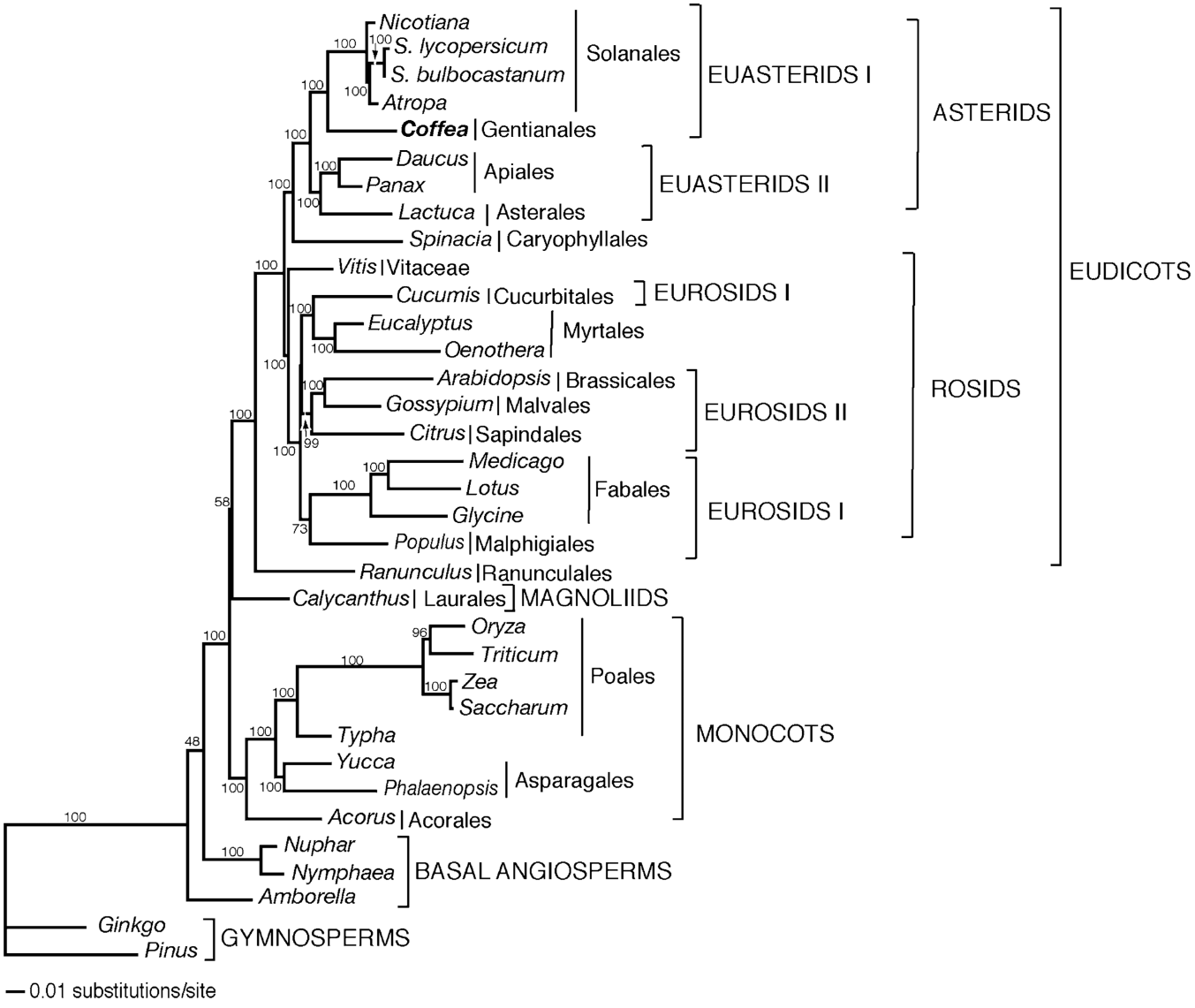

Figure 4.

Maximum likelihood tree based on 61 chloroplast protein-coding genes. The single maximum likelihood phylogram has a maximum likelihood value of −ln *L* = 348 679.23765. Numbers at the nodes indicate the bootstrap support values and the branch length scale is shown at the base of the tree.

Table 1

List of genes encoded by the coffee chloroplast genome

RNA genes
Ribosomal RNA genes
*rrn*4.5[*], *rrn*5[*], *rrn*16[*], *rrn*23[*]
Transfer RNA genes
trnA(UGC)[*†], *trnC*(GCA), *trnD*(GUC), *trnE*(UUC), *trnF*(GGA), *trnG*(GCC)[†], *trnH*(GUG), *trnI*(CAU)[*], *trnI*(GAU)[*†], *trnK*(UUU)[†], *trnL*(CAA)[*], *trnL*(UAA)[†], *trnL*(UAG), *trnfM*(CAU), *trnM*(CAU), *trnN*(GUU)[*], *trnP*(UGG), *trnQ*(UUG), *trnR*(ACG)[*], *trnR*(UCU), *trnS*(GCU), *trnS*(GGA), *trnS*(UGA), *trnT*(GGU), *trnT*(UGU), *trnV*(GAC)[*], *trnV*(UAC)[†], *trnW*(CCA), *trnY*(GUA)
Polypeptide genes
Ribosomal protein genes (larger subunit)
rpl2[*†], *rpl*14, *rpl*16[†], *rpl*20, *rpl*22, *rpl*23[*], *rpl*32, *rpl*33, *rpl*36
Ribosomal protein genes (smaller subunit)
*rps*2, *rps*3, *rps*4, *rps*7[*], *rps*8, *rps*11, *rps*12[*‡], *rps*14, *rps*15, *rps*16[†], *rps*18, *rps*19[§]
Transcription/translation apparatus genes: *rpoA, rpoB, rpoC*1[†], *rpoC*2, *infA*
Acetyl-CoA carboxylase: *accD*
ATP-dependent protease: *clpP*[‡]
ATP synthase: *atpA, atpB, atpE, atpF*[†], *atpH, atpI*
Cytochrome *b/f*: *petA, petB*[†], *petD*[†], *petG, petL, petN*
Cytochrome *c* biogenesis: *ccsA*
Membrane protein: *cemA*
NADH dehydrogenase: *ndhA*[†], *ndhB*[*†], *ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
Photosystem I: *psaA, psaB, psaC, psaI, psaJ*
Photosystem II: *psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ*
Rubisco: *rbcL*
Maturase: *matK* [as intron in *trnK*(UUU)]
Conserved reading frames: *ycf*1, *ycf*2[*], *ycf*3[‡], *ycf*4

NADH, reduced nicotinamide adenine dinucleotide; Rubisco, ribulose-1,5-bisphosphate carboxylase/oxygenase.

[*] Gene duplicated in the inverted repeat.

[†] Gene with one intron.

[‡] Gene with two introns.

[§] Gene truncated in IRa.

Table 2

List of genes containing introns in the coffee chloroplast genome

Number	Gene name	Intron size (bp)	Type of gene
1	*trnK*-UUU	2541 *	Transfer RNA
2	*rps16*	834	Protein
3	*trnG*-GCC	693	Transfer RNA
4	*atpF*	693	Protein
5	*rpoC1*	738	Protein
6	*ycf3*‡	749, 699	Protein
7	*trnL*-UAA	494	Transfer RNA
8	*trnV*-UAC	594	Transfer RNA
9	clpP‡	627, 776	Protein
10	*petB*	765	Protein
11	*petD*	719	Protein
12	*rpl16*	1000	Protein
13	*rpl2*†	659	Protein
14	*rps12*†‡	526	Protein
15	*trnI*-GAU†	936	Transfer RNA
16	*trnA*-UGC†	811	Transfer RNA
17	*ndhB*†	602	Protein
18	*ndhA*	1117	Protein

*Intron includes 2528 bp of the protein-coding gene *matK*.

†Gene located in inverted repeat (IR).

‡Gene contains two introns.

Table 3

Location of repeats in the coffee chloroplast genome. The table includes repeats of at least 30 bp in size, with a sequence identity of 90% or more (see Figure 1 for location of repeats on the gene map)

Repeat no.	Repeat length	Repeat	Location
1	64	Direct	IGS (*rrn5–trnR*-ACG) : IGS (*rrn5–trnR*-ACG)
2	41	Direct	IGS (*rps12–trnV*-GAC) : intron (*ndhA*)
3	34	Inverted	IGS (*ycf1–ndhF*) : *ycf1*
4	32	Direct	IGS (*rrn4.5–rrn5*) : IGS (*rrn4.5–rrn5*)
5	33	Inverted	IGS (*psbC–trnS*-UGA) : intron (*ycf3*)
6	30	Inverted	IGS (*petN–psbM*) : IGS (*petN–psbM*)
7	30	Direct	*psaB* : *psaA*
8	30	Direct	*ycf2* : *ycf2*

IGS, intergenic spacer.

Table 4

Taxa included in the phylogenetic analyses with GenBank accession numbers and references

	Taxon	GenBank accession number	Reference
Gymnosperm outgroups			
	Pinus thunbergii	NC_001631	Wakasugi *et al.* (1994)
	Ginkgo biloba	DQ069337-DQ069702	Leebens-Mack *et al.* (2005)
Basal angiosperms			
	Amborella trichopoda	NC_005086	Goremykin *et al.* (2003)
	Nuphar advena	DQ069337-DQ069702	Leebens-Mack *et al.* (2005)
	Nymphaea alba	NC_006050	Goremykin *et al.* (2004)
Magnoliids			
	Calycanthus floridus	NC_004993	Goremykin *et al.* (2003)
Monocots			
	Acorus americanus	DQ069337-DQ069702	Leebens-Mack *et al.* (2005)
	Oryza sativa	NC_001320	Hiratsuka *et al.* (1989)
	Phalaenopsis aphrodite	NC_007499	Chang *et al.* (2006)
	Saccharum officinarum	NC_006084	Asano *et al.* (2004)
	Triticum aestivum	NC_002762	Ogihara *et al.* (2002)
	Typha latifolia	DQ069337-DQ069702	Leebens-Mack *et al.* (2005)
	Yucca schidigera	DQ069337-DQ069702	Leebens-Mack *et al.* 2005
	Zea mays	NC_001666	Maier *et al.* (1995)
Eudicots			
	Arabidopsis thaliana	NC_000932	Sato *et al.* (1999)
	Atropa belladonna	NC_004561	Schmitz-Linneweber *et al.* (2001)
	Citrus sinensis	NC_008334	Bausher *et al.* (2006)
	Coffea arabica	NC_008535	Current study
	Cucumis sativus	NC_007144	Plader, W.W., *et al.*, Warsaw Agricultural University, unpubl. data
	NC_008325	DQ898156	Ruhlman *et al.* (2006)
	Eucalyptus globulus	NC_008115	Steane (2005)
	Glycine max	NC_007942	Saski *et al.* (2005)
	Gossypium hirsutum	NC_007944	Lee S.M. *et al.* (2006)
	Lactuca sativa	DQ383816	Timme *et al.* (2007)
	Lotus corniculatus	NC_002694	Kato *et al.* (2000)
	Medicago truncatula	NC_003119	Lin, S. *et al.*, The University of Oklahoma, unpubl. data
	Nicotiana tabacum	NC_001879	Shinozaki *et al.* (1986)
	Oenothera elata	NC_002693	Hupfer *et al.* (2000)
	Panax schinseng	NC_006290	Kim & Lee (2004)
	Populus trichocarpa	NC_008235	Okumura, S. *et al.,* Research Institute of Innovative Technology for the Earth, Kyoto, unpubl. data
	Ranunculus macranthus	DQ069337–DQ069702	Leebens-Mack *et al.* (2005)
	Solanum lycopersicum	DQ347959	Daniell *et al.* (2006)
	Solanum bulbocastanum	NC_007943	Daniell *et al.* (2006)

Taxon	GenBank accession number	Reference
Spinacia oleracea	NC_002202	Schmitz-Linneweber *et al.* (2001)
Vitis vinifera	NC_007957	Jansen *et al.* (2006)

Table 5

Repeat analysis of the chloroplast genome of selected crop plants using comparative repeat analysis (CRA). Data include direct, inverted, complement and reverse repeats with a minimum size of 30 bp and a sequence identity of 90% or more

Crop plant	No. of direct repeats	No. of inverted repeats	No. of complement repeats	No. of reverse repeats	Total number of repeats
Coffea arabica	5	3	–	–	8
Citrus sinensis	9	6	1	3	19
Daucus carota	13	2	–	–	15
Gossypium hirsutum	16	6	–	3	25
Glycine max	13	17	3	6	39
Solanum lycopersicum	9	1	–	2	12
Solanum bulbocastanum	10	1	–	3	14
Vitis vinifera	8	1	–	3	12