PLASTID GENOME EVOLUTION IN PARTIALLY AND FULLY MYCOHETEROTROPHIC EUDICOTS

by

Hayley Darby

B.A., Portland State University, 2010

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Botany)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

December 2015

Abstract

Plastid-genome evolution following photosynthesis loss is characterized by substantial change, contrasting with strong conservation in most photosynthetic land plants. Common features of reduced plastid genomes across diverse heterotrophic lineages point to a predictable trajectory of genome degradation, but this has been only partly tested. Here I document the molecular evolution of plastid genomes belonging to several mycoheterotroph lineages in Ericaceae, Gentianaceae and Polygalaceae, which include several independent origins of mycoheterotrophy in eudicot angiosperms that span different time scales since photosynthesis loss. I used next-generation and Sanger sequencing techniques to assemble complete plastomes or gene sets for comparative analyses of gene content and genome structure, and phylogenomic inference. I also sequenced several partially mycoheterotrophic and fully autotrophic relatives. Patterns of gene loss in mycoheterotroph plastomes are generally consistent with a previously hypothesized trajectory of change, starting with the loss of plastid NAD(P)H dehydrogenase before full loss of photosynthesis, and ending (here) with substantial reduction in genes involved in the translation apparatus and other nonphotosynthetic functions. Several retentions (delayed losses) of subunit genes for plastid-encoded polymerase, plastid ATP synthase and Rubisco are also consistent with hypothesized secondary (nonphotosynthetic) functions for these complexes. Two within-genus comparisons (for *Epirixanthes* in Polygalaceae and *Voyria* in Gentianaceae) demonstrate substantially different levels of genome degradation, consistent with heterogeneity in rates of genome change after a given origin of full mycoheterotrophy. Mycoheterotrophs in two families (Ericaceae, Polygalaceae) have extensive genome rearrangement compared to most land plants, contrasting with near colinearity in mycoheterotrophic members of Gentianaceae (despite sometimes extensive genome reduction in the latter). However, these contrasting patterns are

apparently not associated with transitions to mycoheterotrophy, as photosynthetic relatives in Ericaceae and Polygalaceae are also substantially rearranged—or with inverted repeat loss (evident in *Epirixanthes pallida*, Polygalaceae), as autotrophic *Polygala* retains its inverted repeats. Phylogenomic inferences of core eudicot phylogeny made using the retained genes are generally well supported and robust to a variety of phylogenetic approaches, and are also congruent with recent phylogenetic studies in each mycoheterotrophic family.

Preface

All steps of this work were conducted predominantly by me, but with the assistance of others as follows: Vivienne Lam (University of British Columbia) was responsible for DNA extraction for four samples: *Exochaenium oliganthum* (Gentianaceae)*, Voyria clavata* (Gentianaceae), and *Voyria caerulea* (Gentianaceae); Marybel Soto Gomez (University of British Columbia) prepared DNA libraries for two samples: *Exochaenium oliganthum* and *Voyria clavata*. Additional thanks are due to Vincent S.F.T. Merckx (Naturalis Biodiversity Center), G. Beatty and J. Provan (Queen's University Belfast), J.R. Abbott (University of Florida), K. Neubig (University of Florida), S. Stefanović (University of Toronto, Mississagua), and R. Bertin (College of the Holy Cross) who kindly provided multiple plant samples as DNA or silica dried specimens. David Tack (University of British Columbia) and Daisie Huang (University of British Columbia) provided bioinformatics scripting assistance.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to start by offering my thanks to my supervisor Dr. Sean Graham who gave me a challenging and stimulating research topic. With his guidance I have gained a deeper understanding of many aspects of plant evolution and phylogenetics. He has taught me to think critically and to value clearly communicated research.

I would also like to extend my appreciation to my committee members Dr. Mary Berbee and Dr. Jeanette Whitton. Their thought-provoking questions have directed me to explore ideas that I may not have otherwise considered.

I thank my current and former lab-mates, Marybel Soto Gomez, Qianshi Lin, Isabel Marques, David Bell, Wesley Gerelle, Vivienne Lam and Greg Ross who taught me laboratory protocols and analytical methods, and engaged me with exciting botanical discussions. Special thanks to Greg Ross who made preparing DNA libraries fun, to Marybel Soto Gomez for providing invaluable feedback in preparing for talks, and to Vivienne Lam for teaching me everything she knows about plastome evolution in peculiar plants.

Special thanks are owed to my parents, who encourage me to do pretty much whatever I want, academically and otherwise.

And thanks to my darling J.P. for making all of my breaks from school fun and for being my emissary while I'm away from home. Most of all, I thank him for his enduring support.

To my parents, who gave me a spot in the garden.

Chapter 1: Introduction

The plastid genome (plastome) of photosynthetic land plants is generally highly conserved in gene content and order, length and overall architecture (reviewed in Palmer, 1985; Wicke et al., 2011). It typically codes for ~110-120 unique genes, and its ~120-160 kb length is quadripartite in structure: a subset of duplicated genes are located in inverted-repeat (IR) regions of variable length across taxa that separate two asymmetrical single-copy regions. The latter regions are referred to as the large and small single copy (LSC and SSC) regions, respectively. Published plastome sequences of heterotrophic plants depart in some or all of these characteristics, in a lineage-dependent manner. For example, the plastomes of the mycoheterotroph *Petrosavia stellaris* (Petrosaviaceae) and the obligate holoparasite *Conopholis americana* (Orobanchaceae) have reduced length, gene content and an atypical gene order due to rearrangements (Wicke et al., 2013; Logacheva et al., 2014), while that of *Sciaphila densiflora* (Triuridaceae) is highly reduced in gene content while retaining nearly complete colinearity with its close photosynthetic relatives (Lam et al., 2015). Although they have heterogeneous patterns of gene loss and genome rearrangement, comparative analysis of genome evolution in different mycoheterotrophic lineages may allow us to make broad generalizations on the effect of photosynthesis loss on plastome molecular evolution (e.g., Barrett and Davis, 2012; Barrett et al. 2014).

Mycoheterotrophy is a plant nutritional strategy that is distinct from direct plant parasitism, and is referred to as "full" mycoheterotrophy when photosynthesis has been lost. Fully mycoheterotrophic plants are completely dependent on fungal partners for their nutritional needs. Parasitic plants use haustoria to penetrate and parasitize the tissues of green plants, but mycoheterotrophs attract and consume fungal hyphae in modified root systems (Leake and Cameron, 2010; Merckx, Freudenstein, et al., 2013) The hyphae may belong to fungi involved in

mycorrhizal networks (these mycoheterotrophic plants thus indirectly parasitize the green-plant partners of mycorrhizal fungi), or in a few cases belong to saprophytic fungi (Bidartondo, 2005). Although relatively rare in terms of species number (less than 1% of land-plant species are full-blown heterotrophs), plant parasites and full mycoheterotrophs have evolved repeatedly across land-plant phylogeny (Merckx, 2013). There are 514 known species of fully mycoheterotroph plants, representing an estimated 46 or 47 independent losses of photosynthesis. Of these, a minimum of seven origins of full mycoheterotrophy (representing 47 species) are known in the core eudicots (Merckx et al., 2013a), where full mycoheterotrophy has evolved independently in three families (Ericaceae, Gentianaceae and Polygalaceae). In addition, partial mycoheterotrophs (plants that both photosynthesize and derive some nutrition from fungal partners) are known in Ericaceae, Gentianaceae and possibly also Polygalaceae (Tedersoo et al., 2007; Zimmer et al., 2007; Hynson et al., 2009; Cameron and Bolin, 2010; Merckx et al., 2013a). Mycoheterotrophic eudicots associate with arbuscular mycorrhiza-forming glomeromycete fungi in Polygalaceae and Gentianaceae, and ectomycorrhizal basidiomycete and ascomycete fungi in Ericaceae (Hynson and Bruns, 2009; Merckx, Freudenstein, et al., 2013).

With the exception of Ericaceae (see Braukmann and Stefanović, 2012), plastid genome evolution in eudicot mycoheterotrophs has not been explored. Examining independent losses of photosynthesis in these lineages of plants would be useful to more fully understand the breadth of plastome evolution in plants, and would provide counterpoints for recently published plastid genomes produced for monocot and liverwort mycoheterotrophs (Wickett et al., 2008; Delannoy et al., 2011; Logacheva et al., 2011, 2014; Barrett and Davis, 2012; Barrett et al., 2014; Lam et al., 2015; Schelkunov et al., 2015). Using evidence from published plastome sequences of heterotrophic plants and known functions of plastid genes, Barrett and Davis (2012) and Barrett

et al. (2014) proposed models for plastid genome degradation during or following the transition to a heterotrophic lifestyle. Their closely related ratchet-like models begin with the loss of plastid NAD(P)H genes, likely before the loss of photosynthesis in partial mycoheterotrophs, followed by concerted degradation of photosynthesis genes and the plastid-encoded RNA polymerase ('PEP,' which transcribes most photosynthesis genes, Hajdukiewicz et al., 1997; reviewed in Yagi and Shiina, 2014). Later-stage plastid genome gene loss or degradation apparently involves plastid ATP synthase loci (which appear to be retained after the initial loss of photosynthesis; Knauf and Hachtel, 2002; Wickett et al., 2008; Barrett et al., 2014; Logacheva et al., 2014), followed by genes involved in the plastid genetic apparatus and other non-photosynthetic functions. The degradation is ratchet-like because genes are assumed to not re-evolve once lost. Thus, the extent of degradation in mycoheterotrophs may correlate with the degree and recency of dependence on heterotrophic nutrition.

The primary objective of my study is to survey plastid genome evolution in eudicot mycoheterotroph lineages, to use these new data in comparative analyses of gene content and genome structure, and for use in phylogenetic inference to place taxa in the context of core eudicot relationships. I used next-generation (NGS) and Sanger sequencing techniques to assemble complete circle plastomes for mycoheterotroph plants that represent three of the estimated seven origins of full mycoheterotrophy that have occurred in eudicots. Within Gentianaceae, I included *Exochaenium oliganthum* as an example of a recent loss of photosynthesis (estimated to have occurred within the last three million years, Merckx et al., 2013b). Chlorophyllous populations have also been reported for it (Kissling, 2012); chlorophyll retention has also been noted in full mycoheterotrophs such as *Cymbidium macrorhyzon* (Merckx et al., 2013a) and *Corallorhiza* spp.(Cummings and Welschmeyer, 1998; Barrett et al., 2014),

and does not necessarily reflect retention of photosynthesis. In contrast, loss of photosynthesis in *Voyria* dates to at least 31 million years ago, based on a crown-age dating for this fully mycoheterotrophic lineage (Merckx et al., 2013b). I included representatives of the single origin of full mycoheterotrophy in Polygalaceae, the exclusively non-photosynthetic genus *Epirixanthes*, which has an estimated crown age of 14 million years (Mennes et al., 2015b). I also included several partial mycoheterotrophs (species from two genera each in Gentianaceae and Ericaceae), and green relatives for all three families (published sequences for Ericaceae and new sequences in Gentianaceae and Polygalaceae) to provide close points of genomic comparison.

This sampling allowed me to explore plastome evolution over a range of different time scales, across taxa of different evolutionary histories and degrees of heterotrophy, and involving homologous (within-genus) and non-homologous losses of photosynthesis (losses between genera here). I used these new data to address the following specific questions: (1) Do plastid genomes evolve in a predictable manner after the transition to heterotrophy, as proposed by Barrett and Davis (2012) and Barrett et al. (2014)? (2) Do we see any unexpected retention of photosynthetic genes in full mycoheterotrophs, which I think point to secondary functions for them in the plastid? (3) Are plastid genes retrieved from heterotrophs useful in plastid-genome scale phylogenetic inference? (4) What (if any) structural rearrangements are associated with the origins of mycoheterotrophy in these taxa?

Chapter 2: Materials and Methods

2.1 Taxonomic sampling

I sampled two species representing the single origin of full mycoheterotrophy in Polygalaceae (*Epirixanthes*), three species representing two of the estimated four origins in Gentianaceae (*Exochaenium* and *Voyria*), and four partially mycoheterotrophic species, two each from Ericaceae and Gentianaceae. I also sampled at least two putatively fully autotrophic taxa in Polygalaceae (*Polygala* and *Salomonia*), an autotrophic Gentianaceae (*Exacum*), and included several publicly available plastid genomes of autotrophic members of Ericaceae, allowing comparisons between heterotrophic and autotrophic relatives in each case (see Table 1). The full taxon sampling includes sequences from 69 taxa retrieved from GenBank and 91 from the larger matrix presented in Ruhfel et al. (2014), and represents multiple lineages of monocots, magnoliids and other angiosperms (Amborellales, Nymphaeales, Austrobaileyales). My taxon sampling within eudicots includes a single representative for each available family across the core eudicots, with denser sampling in lineages that are more closely related to Polygalaceae, Gentianaceae and Ericaceae (Table S1). It also includes all available eudicot plastid genomes from heterotrophs (parasitic plants belonging to Orobanchaceae, Convolvulaceae and Santalales) and carnivorous plants (members of Lentibulariaceae).

2.2 DNA isolation and library preparation

I prepared sampled species for whole-genome shotgun-sequencing on the Illumina HiSeq 2000 platform (Illumina, Inc., San Diego, USA) to retrieve complete plastid genome sequences. I first extracted DNA from silica-dried tissue samples using the method of Doyle and Doyle (1987).

Several samples (*Orthilia secunda, Pyrola minor, Epirixanthes elongata* and *Salomonia cantoniensis*) were provided by collaborators as DNA extractions. I prepared genomic DNA libraries using three kits (KAPA Library Preparation Kit, KAPA Biosystems, Wilmington, USA; Nugen Ovation Ultralow Library systems, NuGEN, San Carlos, USA; Bioo NextFlex Rapid sequencing kit, Bioo Scientific, Austin, USA), following manufacturer protocols for each kit, using genomic DNA sheared to 400 bp fragments with a Covaris sonicator (model: S220, Woburn, Massachusetts, USA) as a starting point. I confirmed that the libraries met a minimum concentration of 0.5 ng/ul and were in a 500-600 bp size range, by using a Qubit fluorometer (Qubit 2.0 Fluorometer, Life Technologies, Thermo Fisher Scientific, Waltham, USA) and Bioanalyzer (2100 Bioanalyzer, Agilent Technologies, Santa Clara, United States), respectively. Sample concentrations were then quantified on an iQ5 real-time qPCR system (Illumina DNA standard kit, KAPA Biosystems, Boston, USA; Bio-Rad Laboratories, Inc., Hercules, USA) and sequenced as 100 bp paired-end reads, on multiplexed Illumina runs (Cronn et al., 2008) that included 10 to 39 samples per lane.

2.3 De novo contig assembly, plastid gene annotation and plastome reconstruction

The multiplexed Illumina sequence reads were sorted by taxon using CASAVA 1.8.2 (Illumina Inc., San Diego, California, USA). I performed *de novo* assemblies on each sample using CLC Genomics Workbench v. 6.5.1 (CLC bio, Aarhus, DK), selecting all contigs larger than 500 bp and at least 10X coverage. I then used a custom Perl script (Daisie Huang, University of British Columbia) to BLAST (Altschul et al., 1990) contigs against local databases of three reference plastomes (Gentianales: *Asclepias syriaca*, NC_022432.1; Fabales: *Glycine max*, NC_007942.1; Ericacales: *Arbutus unedo*, JQ067650), in order to identify and remove mitochondrial and

nuclear contigs. For *Pyrola minor* and *Voyria caerulea*, I annotated and isolated individual plastid genes in CLC-produced contigs using DOGMA (Wyman et al., 2004), manually inspecting gene and exon boundaries in Sequencher 4.2.2. (Gene Codes Corporation, Ann Arbor, US) using *Arbutus unedo* (JQ067650) or *Asclepias syriaca* (NC_022432.1) as reference sequences, respectively. For all other taxa, I assembled CLC-produced contigs into full or nearly full circular plastomes, by bridging gaps and confirming contig overlap using Sanger-based DNA sequencing. I designed custom primers for amplification and Sanger sequencing using Primer3 (Untergrasser et al. 2007; Koressaar and Remm 2007) (see Table S2 for primer sequences), performing amplifications using Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific, USA) and sequencing using BigDye Terminator v.3.1 (Applied Biosystems, Inc. Foster City, USA). I performed amplification following the general methodology in Graham and Olmstead (2000) with minor modifications: (1) initial denaturation at 98° C for 5 min; (2) 40 cycles of the following: denaturation at 98° C for 20 s, annealing at 60° C for 30 s, extension at 72° C for 2 min; (3) final extension at 72° C for 5 min. For cycle sequencing, I followed the methodology in Graham and Olmstead (2000) for 25 cycles, with some modifications: (1) denaturation at 96° C for 10 s; (2) annealing at 50° C for 5 s; (3) extension at 60°C for 4 min. Sequencing reactions were run on an Applied Biosystems 3730S 48-capillary DNA analyzer (Applied Biosystems, Inc., Foster City, USA). I produced final whole or partial plastome sequences by assembling Illumina contigs and Sanger sequences in Sequencher 4.2.2, and deduced and annotated gene and exon boundaries using DOGMA and Sequencher, using *Asclepias syriaca*, *Asclepias nivea* (NC_022431), *Glycine max* (NC_007942.1) or *Arbutus unedo* as reference sequences. I used OGDRAW (Lohse et al., 2013) to prepare plastome figures.

2.4 Whole-plastome rearrangements

I used Mauve 3.2.1 (Darling et al., 2004) to predict gene-order rearrangement in the plastomes of mycoheterotrophs with respect to photosynthetic relatives, omitting the second copy of the inverted repeated for these analyses. This program identifies regions of homology shared between at least two sequences in an alignment (called locally colinear blocks; LCBs) using a combination of string-matching, local alignment and breakpoint analysis, and positions LCBs using progressive alignment (CLUSTALW; Thompson et al., 1994). Minimum string lengths ('seeds lengths') and LCB calculation parameters can be optimized by the user: I used a seed length of 21 bp to minimize spurious matches, and allowed minimum LCBs to be calculated automatically.

2.5 Concatenated alignment construction

I performed alignments on individual genes, excluding introns and intergenic regions (and initially included pseudogenes, see below) to prepare a final fully concatenated matrix. I compiled the plastid gene sets I generated (Table 1) with a set of taxa chosen from a publicly available green-plant-wide matrix (Ruhfel et al. 2014) and plastid-genome sequences available from GenBank (Supplementary Table S1). To do this I exported new sequences in FASTA format, and generated single-gene, multi-taxon files using custom Python scripts (Dave Tack, University of British Columbia). These files represent 78 protein-coding and four ribosomal DNA (rDNA) loci. Missing genes for individual taxa (see below) in individual alignments were represented as blanks. The protein-coding set of genes includes the loci typically present in angiosperms, but I excluded *ycf*1 due to alignment difficulty. For each gene I produced automated DNA sequence alignments using MAFFT (Katoh et al., 2002), inspecting the output

and manually adjusting it where necessary, following alignment criteria laid out in Graham et al., (2000). I performed these alignment steps (automated alignment and manual adjustment) using Mesquite v. 3.03+ and v. 3.4 (Maddison and Maddison, 2014, 2015). I used the default settings for MAFFT, although for the gene *ycf*2 I used the 'linsi' option, a more computationally intensive and thorough search approach. I removed introns from split genes, and staggered difficult-to-align regions, as described in Saarela and Graham (2010). Genes obtained from the Rufhel et al. (2014) matrix were pre-trimmed in various ways (i.e., at their 5'- and 3'-ends, and for introns and poorly-aligned regions).

I combined these individual gene alignments into a single concatenated matrix, and prepared two versions of it. One version excluded all or most pseudogenes (see below). This combined 'ORF-only' matrix (ORF = open reading frame) comprised 81,732 bp (for reference, derived from 57,507 bp sequence data in *Exochaenium oliganthum*). I also translated the 78 protein-coding genes in the ORF-only matrix using Mesquite, and constructed a concatenated 25,528 amino-acid residue matrix from this. For newly sequenced taxa, the ORF-only concatenated matrix included four genes with a single reading frame interruption compared to reference taxa (Table 2), which I retained as they may reflect sequencing errors or RNA edit sites e.g.(e.g. Freyer et al., 1997; Kugita et al., 2003; Hoffmann et al., 2009); thus, this matrix may include several genes with recent loss of function. However, in three of these four genes, other subunit genes have multiple reading frame interruptions; and so a more likely situation is that there is a lag in the accumulation of reading frame interruptions in the subunits with only a single interruption. I therefore retained a version of the concatenated matrix that included these and other more obvious pseudogenes. I used this to assess whether their inclusion had an effect on phylogenetic inference. Where the 5'- or 3'-end of a putative pseudogene was not readily

alignable, I trimmed this portion from the alignment. I based the pseudogene status of published sequences on their respective GenBank annotations (Table S1). This 84,567 bp matrix that included pseudogenes was derived from 66,820 bp sequence data for *Exochaenium oliganthum*, for reference.

To ensure that copy-paste or other editing errors were not introduced during data compilation or manual alignment adjustment, I examined the DNA matrix using the following approaches. I excluded all taxa except those retrieved from the Ruhfel et al. (2014) matrix (91 taxa remained) and re-aligned these sequences using MAFFT against all sequences for the corresponding taxon set in the original matrix, for all genes simultaneously (distinguishing realigned and original data in the taxon names). I then ran a heuristic parsimony analysis of this 182-taxon matrix. I consistently found that the original and realigned sequences were sister taxa, and had no differences in terminal branch length between them. For all other taxa, I exported concatenated gene sequences for each taxon and used Sequencher 4.2.2. (Gene Codes Corporation, Ann Arbor, US) to compare them to the original individual taxon files. No obvious editing errors were found using these two error-checking methods.

2.6 Phylogenetic inference

I analyzed the ORF-only data using maximum likelihood and parsimony methods. I ran a heuristic parsimony search in PAUP version 4.0a145 (Swofford, 2003), using tree-bisection-reconnection (TBR) branch swapping, with 10 random stepwise addition replicates, and holding one tree at each step. I performed several different ML searches using RAxML v. 7.4.2 (Stamatakis, 2006), conducting 20 independent searches for the best tree in each case. I also performed bootstrapping analyses to assess the strength of branch support for trees (Felsenstein,

1985). For the parsimony analysis, I ran 500 bootstrap replicates with 10 random stepwise addition replicates. For the ML analyses, I ran 500 rapid bootstrap replicates (Stamatakis et al., 2008) using the same DNA or amino-acid substitution models and partitioning schemes used in searches for the best tree (see below). I considered branches with 95% or better bootstrap support as well-supported, and branches with <70% bootstrap support poorly-supported, following Zgurski et al. (2008). All nucleotide ML analyses were performed on the CIPRES portal (Miller et al., 2010). The amino-acid analysis was performed using the RAxML graphical front-end interface (Silvestro and Michalak, 2012).

For DNA-based ML analyses, I ran both unpartitioned and partitioned analyses. The latter considered codon positions within each protein-coding gene ('GxC' or gene by codon partitioning scheme). To decide on the partitions for the GxC analysis, I designated an initial 238 partitions for the concatenated matrix (derived from the first, second or third codon positions of each protein-coding gene, and four unique partitions representing the four rDNA loci). I then used PartitionFinder version 1.1.1. (Lanfear et al., 2012) to pool partitions that did not have significantly different substitution models or model parameters using the Bayesian Information selection criterion (BIC). For this analysis, branch lengths were linked and only the substitution models implemented in RAxML were explored using the relaxed hierarchical clustering algorithm, as described in Lanfear et al. (2014). I searched the top 5% of schemes expected to improve likelihood scores. I ran a partition-finder analysis of the version of the concatenated matrix with pseudogenes included, in the same manner. I also ran a partition-finder analysis for the concatenated amino-acid matrix using PartitionFinderProtein version 1.1.1 (Lanfear et al., 2012), starting with the 78 protein-coding genes, and otherwise using the settings described above. The ORF-only DNA matrix yielded a partition-scheme with 64 final partitions, and

recovered the GTR+ Γ or GTR+I+Γ DNA substitution models as the best fit for individual data partitions (Table S3). The version of the concatenated matrix that included obvious pseudogenes yielded a partition-scheme with 67 final partitions, and recovered the GTR+ Γ or GTR+I+Γ DNA substitution models as the best fit for individual partitions. PartitionFinder also identified GTR+ Γ as the best DNA substitution model for the unpartitioned ML analysis of the matrix. The partition-finder analysis of the amino-acid matrix found 37 partitions, with best models that included variants of the JTT, JTTF, CPREV, MTMAM or LG substitution models (see Table S3 final data partitioning schemes). I applied the optimal models for each data partition in the various partitioned likelihood analyses.

Chapter 3: Results

3.1 Plastome characteristics

I assembled complete, circular plastome sequences for seven species, including three full mycoheterotrophs (*Exochaenium oliganthum*, *Voyria clavata*, and *Epirixanthes pallida*), two partial mycoheterotrophs (*Bartonia virginica* and *Obolaria virginica*) and two autotrophs (*Exacum affine* and *Polygala arillata*) (Table 3, Figs. S1-S7). I also recovered a nearly complete assembly for a partial mycoheterotroph (*Orthilia secunda*, which likely has only a single gap; Fig. 1, Table S4). Four others are presented here only as gene sets based on more incomplete assemblies, including two full mycoheterotrophs (*Voyria caerulea* and *Epirixanthes elongata*), one partial mycoheterotroph (*Pyrola minor*) and an autotroph (*Salomonia cantoniensis*) (Table S4).

3.1.1 Polygalaceae

The largest new plastome belongs to *Polygala arillata* (Polygalaceae) (Table 3), with a length of 164,747 bp. It also has the largest inverted repeat (IR) region among those recovered here (36,168 bp, comprising 23 genes that extend from *rpl*2 to *ndh*I; Fig. 1, S1, Table 4). For comparison, the IR region of *Exacum affine* (in Gentianaceae) comprises 20 genes and is 26,239 bp in length, spanning from a point 300 bp into *rps*3 to 1,086 bp into *ycf*1. The plastome of *Epirixanthes pallida* is intermediate among the full mycoheterotrophs presented here in terms of plastome length and gene content. *Epirixanthes pallida* is the sole fully assembled species that has lost an IR; however, it retains two ~12 kb direct repeats composed of genes found in the IR of *P. arillata* (Figs. 1, S2). A partial assembly for *Ep. elongata* suggests that it has a very reduced plastome and several repeated regions based on depth of sequencing (Fig. S8). Sectors

of the assembly had read depth varying eight-fold: the lowest coverage contig (~200X read depth) includes loci for *trn*E-UUC, *trn*Y-GUA and *mat*K, and the highest coverage contigs span the rDNA operon (~1700X read depth). A partial assembly of *Salomonia cantoniensis* (an autotrophic member of Polygalaceae; not shown) is consistent with it having a quadripartite structure. *Polygala arillata* and *S. cantoniensis* have three copies of *trn*Q-UUG. These disjunct genomic locations may have resulted from a translocation or a series of inversions, as a single copy is found adjacent to the RNA polymerase operon in the large single copy region (the ancestral arrangement) and two copies are located in the inverted repeat regions; two copies are found in *Ep. pallida*, one in each direct repeat, and the LSC copy of the gene is not present (Figs. 1, S2).

3.1.2 Gentianaceae

The plastome of the full mycoheterotroph *Exochaenium oliganthum* is comparable in size to autotrophic *Exacum affine*, and is slightly larger than the plastomes of the two partial mycoheterotroph species, *Bartonia virginica* and *Obolaria virginica*. It also retains more genes with intact open reading frames than the latter two species (Table 3). *Bartonia virginica* and *O. virginica* have substantially smaller small single copy (SSC) regions than *Exa. affine* and *Exo. oliganthum*, which may be attributed to gene loss and shifts in SSC/IR boundaries (Table 3, 4 Figs. 1, S3-S6). The smallest, fully assembled plastome I recovered in the current study belongs to the fully mycoheterotrophic *Voyria clavata*, which is 31,724 bp in length and has 25 unique genes with uninterrupted reading frame, specifically four rDNA genes, four tRNA genes and 17 protein coding genes (Table 3). This species retains as single-copy genes 13 of the 20 genes

found in the IR of *Exa. affine*, but has a novel IR region corresponding to a block of five genes located in the large single copy (LSC) of *Exa. affine* (Figs. 1, S7).

3.1.3 Ericaceae

Although incomplete, the plastome of the partial mycoheterotroph *Orthilia secunda* appears to be comparable in length and gene content to partial mycoheterotrophs in Gentianaceae (Table S4, Fig. 1). The partial assembly of *O. secunda* is consistent with a quadripartite structure, and it retains a larger SSC region than *Arbutus unedo* (Fig. 1, Table 4).

3.1.4 Mauve-based inferences of genome rearrangement

Gene order in autotrophic relatives of mycoheterotrophs is modified in Polygalaceae (*Polygala arillata* and *Salomonia cantoniensis*, the latter based on incomplete assemblies; not shown) and Ericaceae (*Arbutus unedo*; Martínez-Alberola et al., 2013), compared to the putative ancestral angiosperm gene order (Jansen et al., 2007). The latter order is represented here by *Exacum affine* (Gentianaceae) and tobacco (*Nicotiana tabacum*, NC_001879), which have the same gene order (Fig. 2). Ignoring often substantial deletions, this ancestral gene order has been largely conserved in the fully mycoheterotrophic Gentianaceae examined here (*Voyria* and *Exochaenium*; Fig. 3). Using Mauve alignment, I identified three colinear blocks among Gentianaceae sequences, comprising the large single copy (LSC) region through *ycf*2 in the inverted repeat (IR), the IR region after *ycf*2 through *ndh*D in the small single copy (SSC), and the rest of the SSC, respectively. There are no rearrangements (which would generally appear as crossed lines in the figure) apart from simple inversions. In comparison to gene order in *Exacum,* two colinear blocks are homologous but in reverse orientation: a block composed of IR genes is

reversed in *Voyria clavata*, and in *Bartonia* there is an inverted three-gene block in its contracted SSC (Fig. 3).

The mycoheterotrophic Polygalaceae and Ericaceae are substantially more rearranged (Fig. 4). Eleven and thirteen colinear blocks are identified in Polygalaceae and Ericaceae Mauve-based alignments of full mycoheterotrophs compared to their autotrophic relatives, respectively. In Polygalaceae, rearrangements and inversions are distributed across the plastome of *Epirixanthes pallida*. A ~12 kb region of the IR in *P. arillata* is directly repeated in *Ep. pallida* ('b' in Fig. 4). Rearrangements are concentrated in the LSC in Ericaceae; the IR of *O. secunda* is a single colinear block. What is reconstructed as an inversion in the SSC of *Orthilia secunda* may be better accounted for as an expansion of IR into the SSC in *Arbutus unedo* (see Fig. 1).

3.2 Gene content

Gene retentions, losses and putative pseudogenizations for protein-coding loci are discussed in more detail below for each family.

3.2.1 Polygalaceae

Most genes coding for subunits of the plastid NAD(P)H complex, photosystems I and II, and cytochrome b_6/*f* complex are lost or interrupted in the plastome of *Epirixanthes pallida*, but plastid-encoded subunits of the ATP synthase complex and *rbc*L (which codes for the large subunit of Rubisco) have been retained (Table 2). All genes of the plastid-encoded RNA polymerase (PEP) are interrupted by premature stop codons. *Epirixanthes pallida* retains all 30 plastid-encoded transfer RNA genes. Although not a complete circle, I recovered 21 unique genes with uninterrupted reading frame in the assembly of *Ep. elongata*, specifically three

rDNA, five transfer RNA and 13 protein coding (Table 2, 3). I also did not retrieve two rDNA loci in the gene set of *Salomonia cantoniensis*. Loss of rDNA loci is not documented in any plant, regardless of trophic status, so I presume these taxa retain these small genes but that they were not assembled into the Illumina contigs. The gene coding for the ATP-dependent caseinolytic protease (*clp*P), a small subunit ribosomal protein (*rps*16) and a large subunit ribosomal protein (*rpl*22) have been deleted in the plastomes of *Polygala arillata* and *Ep. pallida*, and translation initiation factor A (*inf*A) is a pseudogene in both species (Table 2). These genes were not recovered in the gene sets of *Salomonia cantoniensis* or *Ep. elongata*. It is not clear if *acc*D, the gene that codes for the beta subunit of acetyl-CoA carboxylase, is retained in Polygalaceae. I recovered a ~1200 bp open reading frame in *P. arillata*, *S. cantoniensis* and *Ep. pallida,* and a ~500 bp truncated putative pseudogene in *Ep. elongata*: *acc*D lacks introns, and is ~1400-1600 bp in Gentianaceae and Fabaceae, for comparison. BLAST searches using the 1395 bp *acc*D locus from *Ceratonia siliqua* (Fabaceae, NC_026678) matched only to subregions of the intact reading frame in *P. arillata, S. cantoniensis* and *Ep. pallida*, recovering matches for 49%, 52% and 46% of the query length, respectively. Protein-translated BLAST searches yielded similar match lengths.

3.2.2 Gentianaceae

The autotrophic *Exacum affine* retains open reading frames for all loci typically found in angiosperm plastomes (Table 2). All photosynthesis-related genes have been deleted in the plastome of *Voyria clavata*, except for a truncated *rbc*L pseudogene (Table 3). Although not a complete circle, all photosynthesis genes retrieved from the assembly of *Voyria caerulea* (nine genes) have interrupted reading frames except for a single locus encoding a subunit of the ATP

synthase complex. *Voyria clavata* retains four transfer RNA loci, and I recovered 13 transfer RNA loci in the *V. caerulea* gene set. *Mat*K, which codes for the group IIa intron maturase (MATK) is also deleted from the *V. clavata* plastome, and I did not recover it in the partial assembly of *V. caerulea*. The two *Voyria* plastomes do, however, retain several intact genes with group IIa introns (i.e., *clp*P, *rpl*2, and *rps*12) (Table 2). There are no gene deletions in the plastome of *Exochaenium oliganthum*, but there are reading frame interruptions in several genes that code for key components of photosystems I and II (Table 2). These include *psa*A, which has no detectable start codon and multiple premature stop codons, and *psb*A, which has a single nucleotide deletion resulting in a frame shift. The third exon of the photosystem I assembly protein, *ycf*3, is also deleted in this species, and there are multiple reading-frame interruptions in the sequence of *ccs*A. The plastid NAD(P)H-dehydrogenase (*ndh*) loci in the full mycoheterotrophs and in the two partial mycoheterotroph species, *B. virginica* and *O. virginica*: all have interrupted reading frames or deletions, for at least some of the genes (Table 2). Genes related directly to photosynthesis (photosystems I and II, the cytochrome b_6/f complex, *rbc*L and ATP synthesis) all have intact reading frames in the two partial mycoheterotroph species, with two exceptions in *O. virginica*. First, the gene coding for a component of photosystem II, *psb*M, is deleted in this species, and second, the c-type cytochrome biogenesis protein, *ccs*A, may also be a pseudogene for it, as it has a single base deletion resulting in a frame shift.

3.2.3 Ericaceae

Most *ndh* genes are interrupted by premature stop codons and non-triplet indels in *Orthilia secunda* and *Pyrola minor.* All other photosynthesis genes are retained with uninterrupted reading frames. It is not clear whether *acc*D has been retained in *O. secunda* and *P. minor* (*acc*D

is a pseudogene in the autotroph *Arbutus unedo*, for reference; Table 2). I recovered what appears to be a ~850 bp fragment of the 3' end of *acc*D with an uninterrupted reading frame in *O. secunda*, and a ~1,600 bp ORF in *P. minor*. BLAST searches using the 1542 bp *acc*D locus from a close relative in which *acc*D is clearly retained (*Camellia crapnelliana,* NC_024541.1) match only to subregions of the intact reading frame in *O. secunda* and *P. minor*, recovering matches for 64% and 34% of the query length, respectively. Protein-translated BLAST searches yielded similar match lengths. *Clp*P, a pseudogene in *A. unedo*, was not recovered from *P. minor* or *O. secunda.*

3.3 Plastid phylogenomics of mycoheterotrophic eudicots

I inferred no major topological differences in core eudicot relationships across the various analyses (Figs. 5-6, S9-S12). Ericaceae, Gentianaceae, and Polygalaceae comprised monophyletic lineages in all phylogenetic analyses, with consistently strong bootstrap support (Figs. 5-6, S9-S12). Within Polygalaceae, I recovered a clade comprising *Epirixanthes* and *Salomonia* as the sister group of *Polygala*, with strong support across all analyses (Figs. 5, S9-S12). A clade comprising Polygalaceae and Fabaceae, the only representatives of Fabales here, was recovered with strong support in all analyses (Figs. 5, S9-S12).

In Gentianaceae, *Exochaenium* and *Exacum* are inferred to be sister taxa, the partial mycoheterotrophs *Obolaria* and *Bartonia* are sister groups*,* and the two species of *Voyria* also formed a clade, all with strong support (Figs. 6, S9-S12). In the ORF-only ML analyses, I inferred *Exochaenium-Exacum* to be the sister group of *Obolaria-Bartonia*, but with poor support (Figs. 6, S9-S11). In the ML analysis that included obvious pseudogenes, *Exochaenium-Exacum* is the sister group of a clade comprising *Obolaria-Bartonia* and *Voyria,* with strong

support (Fig. S12). I recovered two equally parsimonious trees that differed in whether *Voyria* or *Exochaenium-Exacum* was the sister-group to *Obolaria-Bartonia*, and these relationships collapsed in the strict consensus (Figs. S11). The order Gentianales is monophyletic (considering the three of five families included here): Rubiaceae were inferred to be the sister group to Gentianaceae and Apocynaceae at the current taxon sampling, with strong support across analyses (Figs. 6, S9-S12).

Within Ericaceae, *Pyrola* and *Orthilia* are consistently strongly supported as sister groups across all analyses (Figs. 6, S9-S12), as are *Rhododendron* and *Vaccinium*. The position of *Arbutus* differed between the DNA and amino-acid based analyses. In the DNA-based analyses, *Pyrola-Orthilia* is the sister group of *Arbutus*, an arrangement with moderate to strong bootstrap support (98-100% ML; 81% for parsimony), and this overall clade is the sister group of a clade comprising *Rhododendron* and *Vaccinium*, also with strong support (Figs. 6, S9, S11-S12). In contrast, in the amino-acid based likelihood analysis, *Arbutus* is instead inferred to be the sister group of a clade comprising *Pyrola-Orthilia* and *Rhododendron-Vaccinium*, although this arrangement had poor support (Figs. S10). The order Ericales is inferred to be monophyletic (with only four of ~20 families sampled) with strong support across analyses (Figs. 6, S9-S12). Within Ericales, I recovered a clade comprising Ericaceae and Actinidiaceae as the sister group to Theaceae at the current taxon sampling, with Primulaceae then the sister group of the clade formed by those three families; this arrangement had strong support across analyses (Figs. 6, S9-S12).

Chapter 4: Discussion

4.1 Plastid phylogenomics of eudicot mycoheterotrophs

Plastid genomes have only recently begun to be used for phylogenetic inference with full mycoheterotrophs, because it was assumed that too many genes (or the entire genome) would be lost to allow this, or that retained genes would be evolving too rapidly (e.g., Cronquist, 1988, p 467; Merckx et al., 2009). Rate elevation can be problematic if it leads to long-branch attraction in phylogenetic inference (Felsenstein, 1978; Hendy and Penny, 1989). I did not perform a formal rate analysis here, although the mycoheterotrophs examined here appear to have comparable rates of evolution to other eudicots, or moderately elevated rates (based on visual comparison of branch lengths to their sister groups, and to other close green relatives in the same or related orders of eudicots; Figs. 5, 6, see also Figs. S9-S12). Recent phylogenetic studies using retained plastid gene sets demonstrate that even highly reduced and rapidly evolving plastid genomes allow inferences of phylogenetic relationships for mycoheterotrophs that are well supported and consistent with studies based on genes from mitochondria or the nucleus (e.g., for Corsiaceae and Triuridaceae; Lam et al., 2015; Mennes et al., 2015a). The phylogenetic inferences made here are congruent with other studies using non-plastid data (for Ericaceae, Kron et al., 2002, Braukmann and Stefanović, 2012; for Gentianaceae, Merckx et al., 2013b; for Polygalaceae, Bello et al., 2012, Mennes et al., 2015b) where there are overlapping sets of taxa, and disagree only where one or both studies have poor branch support, as with the sister group of Pyroleae (*Pyrola* and *Orthilia* here) in Ericaceae (Bidartondo and Bruns, 2001; Kron et al., 2002; Braukmann and Stefanović, 2012), and the family-level arrangement of Exaceae (represented by *Exacum* and *Exochaenium*) versus Voyrieae (represented by *Voyria*) (Merckx et al., 2013b). In

addition, my phylogenetic inferences are generally not affected by the use of different phylogenetic criteria (parsimony and likelihood), the use of partitioned vs. unpartitioned likelihood analysis, or by the use of DNA vs. amino-acid substitution models (Figs. 5-6, S9-S12). It also does not seem to matter whether pseudogenes are included in analysis or not (cf. Figs. 5-6, S9-S11 and S12), although in a few cases my data resolve relationships that are unclear elsewhere, such as whether Exaceae or Voyrieae is the sister group to Gentianeae (Figs. 6, S9-S12). A clade comprising Exaceae and Gentianeae was recovered as the sister group to Voyrieae for all analyses where pseudogenes were excluded, but with weak support (Figs. 6, S9-S11). The partitioned analysis that included pseudogenes resolved Exaceae as sister-group to Gentianeae and Voyrieae with strong support (Fig. S12), which is congruent with inferences made by Merckx et al. (2013b) using a non-plastid data set.

4.2 Models of plastid genome degradation in heterotrophic plants

As the need to acquire nutrition via photosynthesis declines and ceases in heterotrophs, purifying selection to maintain genes with protein products involved in the photosynthetic apparatus should be relaxed and eventually released. Barrett and Davis (2012) and Barrett et al. (2014) developed two closely related models of plastid genome evolution in heterotrophic plants that predict an ordered series of plastid gene loss and genome reduction, proposing that the extent of plastome reduction is correlated with the degree and recency of dependence on non-photosynthetically derived nutrition. Once photosynthetic function is lost (and photosynthesis genes begin to be lost or degraded), other associated genes may follow, such as the plastid-encoded RNA polymerase genes that are thought to be necessary for photosynthesis-related gene expression (Hajdukiewicz et al., 1997; Zhelyazkova et al., 2012). The most reduced plastomes

may eventually begin to lose genes with roles in the plastid genetic apparatus and other non-photosynthetic metabolic ('housekeeping') functions, as the importance of the plastid organelle for plant survival diminishes. Eventually, most housekeeping functions may be lost, streamlined or replaced by analogous functions provided by non-homologous genes residing in other genomic compartments (or homologous but successfully transferred genes), with only a core of genetic apparatus genes retained in the service of residual but essential non-photosynthetic plastid-encoded genes (e.g., Barbrook et al., 2006; Delannoy et al., 2011).

The partial and full mycoheterotrophs that I sequenced in the eudicots display nearly the full range of large-scale genome modifications. Two partially mycoheterotrophic taxa (*Obolaria* and *Bartonia*, Gentianaceae) have nearly full-sized plastid genomes (~146 kb) but have extensive degradation in NAD(P)H dehydrogenase genes, which has apparently also happened in the two partially mycoheterotrophic Ericaceae based on the plastid gene sets that I was able to recover (Tables 2, 3). Focusing on the full mycoheterotrophs, *Exochaenium oliganthum* (Gentianaceae) has a genome size typical of green plants (~151 kb) with minimal detectable pseudogenization, *Epirixanthes pallida* (Polygalaceae) has a more reduced genome (~94 kb) with nearly all genes with protein products involved in the light reactions of photosynthesis either degraded or lost but many still retained as pseudogenes, and *Voyria clavata* (Gentianaceae) has a substantially truncated genome (~31 kb), with nearly all photosynthesis and many housekeeping genes also lost (Tables 2, 3).

4.3 Loss and retention of plastid gene products

Below I briefly discuss the significance of gene losses and retentions in these different lineages in terms of the protein complexes and other gene products that the plastid loci code for. I present

each in the approximate order of loss proposed by the Barrett and Davis (2012) and Barrett et al. (2014). It should be noted that the fully mycoheterotrophic lineages examined here likely each represent evolutionarily independent losses of photosynthesis, at least concerning the genera *Exochaenium* and *Voyria* (both Gentianaceae) and *Epirixanthes* (Polygalaceae) (Merckx and Freudenstein, 2010; Merckx et al., 2013a; Merckx et al., 2013b; Mennes et al., 2015a) . However, I also examined two species each in two fully mycoheterotrophic genera, *Voyria* (Gentianaceae) and *Epirixanthes* (Polygalaceae). In both genera, it is most parsimonious to assume that the two species in them are the result of a common loss of photosynthesis (one loss in each genus), as all other species in each genus are also fully mycoheterotrophic (Merckx et al., 2013a). Therefore, both genera provide an opportunity to examine the different rates and possibly different routes of genome degradation that follow an homologous origin of heterotrophy, as has been done elsewhere for Orobanchaceae (Wicke et al., 2013), *Epipogium* (Schelkunov et al., 2015). and *Corallorhiza* (Barrett et al., 2014) (the latter are distinct lineages of full mycoheterotrophs in Orchidaceae). In addition, *Bartonia* and *Obolaria* in Gentianaceae, and *Orthilia* and *Pyrola* in Ericaceae, each provide examples of pairs of related taxa in the early stages of mycoheterotrophy (all four taxa appear to be both photosynthetic and partially mycoheterotrophic based on isotopic evidence; see Cameron and Bolin, 2010 for Gentianaceae, and Tedersoo et al., 2007 and Zimmer et al., 2007 for Ericaceae). It is not known whether partial mycoheterotrophy is homologous within each of these pairs, although *Bartonia* and *Obolaria* may both be closely related to each other within Gentianaceae (both belong to subtribe Swertiinae; Struwe, 2014), as are *Orthilia* and *Pyrola* in Ericaceae (both belong to tribe Pyroleae; Kron et al., 2002). .

4.3.1 Loss of NAD(P)H dehydrogenase in early-transitional mycoheterotrophs

The plastid NAD(P)H dehydrogenase complex is associated with cyclic electron transport and is thought to provide protection from photooxidative damage (Martín and Sabater, 2010; Shikanai, 2015). The complex may be nonessential or less essential in the absence of environmental stress (e.g. light, nutrient or CO_2; Peltier and Cournac, 2002). A functional plastid NAD(P)H complex would not be needed in non-photosynthetic plants, and so it is not surprising that it is functionally lost in all full mycoheterotrophs (e.g., Table 3). However, all plastomes of partially heterotrophic plants (hemiparasites and partial mycoheterotrophic) sequenced to date also exhibit pseudogenization or loss of all or some *ndh* genes, supporting the loss or at least non-functionality of the NAD(P)H dehydrogenase complex. These include photosynthetic hemiparasites in Convolvulaceae (Funk et al., 2007; McNeal et al., 2007), Santalales (Petersen et al., 2015), and Orobanchaceae (Wicke et al., 2013), and partial mycoheterotrophs in Orchidaceae (Zimmer et al., 2008; Barrett et al., 2014). For mycoheterotrophs, the commonality of this loss in both partially and fully heterotrophic taxa led to the hypothesis that *ndh* genes are the initial functional group (and thus their gene products the first protein complex) to be lost or degraded before full mycoheterotrophy, and thus before the loss of photosynthesis (e.g. Barrett and Davis, 2012; Wicke et al., 2013; Barrett et al., 2014). This is consistent with what I found in Ericaceae and Gentianaceae, as the four partial mycoheterotrophs that I surveyed all have degradation of the genes coding for the plastid NAD(P)H dehydrogenase complex, despite the retention of all or most of the other plastid-encoded genes. The loss or non-functionality of this complex in partial (photosynthetic) mycoheterotrophs may reflect less photooxidative stress in understory plants that do not obtain all of their nutrition from sunlight (Barrett et al., 2014). Stable isotope signatures for the partial mycoheterotrophs and congeners sequenced here are enriched in ^{15}N

and ^{13}C, but at an intermediate level between full mycoheterotrophs and autotrophs, pointing to incomplete reliance on fungal nutrition (Tedersoo et al., 2007; Zimmer et al., 2007; Cameron and Bolin, 2010). Degradation of the NAD(P)H complex may have occurred repeatedly in different lineages of photosynthetic orchids (Wu et al., 2010; Yang et al., 2013; Kim et al., 2015; Ruhlman et al., 2015). It remains to be shown how many of these plants are partial mycoheterotrophs at maturity, although isotopic evidence suggests this is the case in several orchids (Gebauer and Meyer, 2003; Bidartondo et al., 2004; Tedersoo et al., 2007; Zimmer et al., 2007, 2008).

Some authors have proposed functional replacement by nuclear copies of plastid *ndh* genes as a possible explanation for plastid-encoded NAD(P)H degradation (e.g. Braukmann et al., 2009; Wu et al., 2010; Blazier et al., 2011). However, Ruhlman et al. (2015) found no evidence of expressed nuclear copies of plastid-encoded *ndh* genes or functional nuclear-encoded components in taxa where plastid-encoded subunits are lost or degraded (although a secondary nuclear-encoded plastid *ndh* complex has been hypothesized in at least *Arabidopsis*; Peltier and Cournac, 2002, see Wicke et al. 2011). It should also be noted that other photosynthetic (and likely non-mycoheterotrophic) lineages of plants have also experienced loss of this plastid protein complex (i.e., Gnetales and Pinaceae, Braukmann et al., 2009; Wu et al., 2009; some Geraniaceae, Blazier et al., 2011; some Lentibulariaceae Wicke et al., 2014; four lineages of Alismatales, Iles et al., 2013; Peredo et al., 2013; Ross et al., 2015; some Cactaceae, Sanderson et al., 2015). Thus, loss or non-functionality of the complex is not necessarily an indicator that a transition to heterotrophy has occurred, or is likely to happen.

4.3.2 Loss and retention of photosynthesis-related genes

Full mycoheterotrophs generally have degraded or lost plastid-encoded genes related to photosynthesis (Wicke et al. 2011), comprising the photosystem I and II complexes and assembly factors (*ycf*3 and *ycf*4), cytochrome b_6/f complex, Rubisco and CO_2 uptake (*cem*A), and ATP synthase. This is the case even for taxa in the relatively early stages of plastome reduction, such as the coralroot orchids *Corallorhiza* (Orchidaceae) and the liverwort *Aneura mirabilis* (Aneuraceae) (Wickett et al., 2008; Barrett and Davis, 2012; Barrett et al., 2014). In Gentianaceae, pseudogenization within these genes is minimal in *Exochaenium oliganthum*, but there is a near-complete loss of these genes in *Voyria* (Table 3). Most photosynthesis-related genes are deleted or have reading frame interruptions in *Epirixanthes pallida* (Polygalaceae), consistent with loss of photosynthetic function; however, six plastid-encoded ATP synthase genes and the Rubisco large subunit (*rbc*L) have been retained in this species with open reading frames (Table 2). These genes were not recovered in the gene set assembled for *Ep. elongata*, pointing to a loss of this complex compared to its congener (this needs to be confirmed by completing the plastid genome for this species). The ATP synthase genes and *rbc*L are also retained as open reading frames in *Exo. oliganthum* (Gentianaceae), although their retention here is may be less surprising given the relatively minor extent of photosynthesis gene reduction in its plastome, and the recency of loss of photosynthesis in it (Merckx et al., 2013b).

Although a lag is expected before reading frames are interrupted in photosynthesis genes following the initial functional loss of photosynthesis (Leebens-Mack and DePamphilis, 2002) the retention of ATP synthase genes and Rubisco in multiple independent mycoheterotrophic lineages after other photosynthesis genes are degraded is noteworthy, and points to probable secondary (non-photosynthetic) functions for them (Wickett et al., 2008). Their retention in some

Gentianaceae and Polygalaceae here adds to the comparable published cases in non-photosynthetic representatives of *Corallorhiza*, in the liverwort *Aneura mirabilis,* and in the monocot *Petrosavia stellaris* (Petrosaviaceae) (Wickett et al., 2008; Barrett and Davis, 2012; Barrett et al., 2014; Logacheva et al., 2014); note, though, that the *Corallorhiza* species lack an open reading frame for *rbc*L. The plastid-encoded ATP synthase genes are also retained in some holoparasitic *Cuscuta* species (Convolvulaceae, their retention there may reflect cryptic photosynthesis during seedling establishment; Machado and Zetsche, 1990) and in some representatives of holoparasitic Orobanchaceae, and intact *rbc*L genes have also been identified in non-photosynthetic representatives in both families (Delavault et al., 1995; Randle and Wolfe, 2005; Funk et al., 2007; McNeal et al., 2007; Wicke et al., 2013). The complete suite of ATP synthase genes are also found in the plastome of a heterotrophic alga, where they are apparently transcribed (Knauf and Hachtel, 2002). Although ATP synthase is directly involved in the production of ATP used in the carbon fixing reactions of photosynthesis (reviewed in Walker, 2012), repeated retention of these genes in some heterotrophs prompted Barrett and Davis (2012) and Barrett et al. (2014) to propose that ATP synthase genes are at least initially retained after the loss of photosynthesis. They may therefore act as a landmark for an intermediate level of genome degradation. Rubisco may follow the same general pattern of delayed loss. However, as there may be no linkage between the proposed secondary functions of these protein complexes (see below), we propose that these two complexes may subsequently be lost in either order.

An explanation for the retention of (putatively) functional ATP synthase has not yet been put forward, but a continued need for plastid ATP production from a non-photosynthetically driven proton gradient, or a need for ATP hydrolysis in heterotroph plastids, have both been proposed (Wicke et al., 2013). Involvement in additional metabolic pathways may also explain

why some non-photosynthetic heterotrophs retain a putatively functional *rbc*L gene (see McNeal et al., 2007; Wickett et al., 2008; Wicke et al., 2013; Logacheva et al., 2014). In addition to its primary role in the Calvin cycle, Rubisco is known to catalyze a glycolysis-bypassing lipid synthesis pathway in white turnip (*Brassica napus*, Brassicaceae), although this reaction is thought to require functioning photosynthetic machinery (Schwender et al., 2004). Rubisco is also involved in the production of serine and glycine via the glycolate pathway of the C2 cycle (Tolbert, 1997) and is expressed at low levels in the non-photosynthetic seeds of the castor bean (*Ricinus communis,* Euphorbiaceae), although its function there is unclear (Osmond et al., 1975). It would be worthwhile to determine whether these or related biosynthetic pathways are maintained in heterotrophs with (otherwise) degraded photosynthesis genes.

In the partial mycoheterotrophs considered here (Table 2), the retention of all or most of the photosynthetic genes and isotopic evidence are both consistent with retention of a functional photosynthetic apparatus. However, in *Obolaria virginica* (Gentianaceae), two genes with products involved in photosynthesis are lost (*psb*M) or have reading frame interruption (*ccs*A). Despite these losses, isotopic evidence and visible photosynthetic tissue support the retention of photosynthesis in this species (Cameron and Bolin, 2010). Barrett et al. (2014) also found interruption of reading frames in *psb*M (and *psa*I) in putatively partial mycoheterotrophic species of *Corallorhiza.* These two genes have roles in the assembly and stability of photosystems II and I, respectively. In *psb*M deficient mutants, the movement of electrons around the PSII complex and stability of component dimers is diminished, but functional (Umate et al., 2007; Kawakami et al., 2011). Xu et al. (1995) demonstrated that *psa*I provides structural stability to photosystem I, but defective mutants had only marginally declined efficiency in the affected photosystem. As

such, it may be possible for *O. virginica* and *Corallorhiza* partial mycoheterotrophs to photosynthesize to some degree in the absence of these subunits.

The c-type cytochrome biogenesis protein (*ccs*A) is widely retained in the plastomes of photosynthetic plants (see Fajardo et al. (2013) for a possible exception), but Peterson et al. (2015) found that the gene was pseudogenized in the photosynthetic hemiparasite *Viscum alba*. The protein product *ccs*A is responsible for heme attachment to c-type cytochromes, which are essential components of the photosynthetic electron transport chain (reviewed in Wicke et al., 2011). Mutations in this gene in *Chlamydomonas reinhardtii* resulted in non-photosynthetic phenotypes, which was attributed to the failure to synthesize some forms of cytochromes (Xie and Merchant, 1996). However, Saint-Marcoux et al. (2009) demonstrated that heme delivery to b_6-type cytochromes is mediated by a different protein, and suggested that cytochrome b_6f may be assembled in the absence of functional *ccs*A. *Ccs*A is among the few photosynthesis genes with reading frame interruptions in the plastome of the full mycoheterotroph *Exo. oliganthum* (Gentianaceae), suggesting that it may be lost relatively early in plastid genome degradation (Table 2).

4.3.3 Loss of plastid-encoded RNA polymerase (PEP) genes

Plastid-encoded RNA polymerase (PEP) is coded for by four *rpo* genes in the plastid genome (Table 3), and is thought to perform the majority of transcription, at least in photosynthetic leaves (Zhelyazkova et al., 2012; reviewed in Liere et al., 2011). The complex may not be essential when photosynthesis genes are lost in full mycoheterotrophs, and nuclear-encoded RNA polymerase (NEP) may perform plastid gene transcription for non-photosynthetic genes that are usually or partly transcribed by PEP (most plastid genes have NEP and PEP promoters,

Liere et al., 2011). For example, functional replacement by nuclear gene products has been given as an explanation for the loss of plastid-encoded RNA polymerase (PEP) genes in 'holoparasitic' (but cryptically photosynthetic) *Cuscuta* species (reviewed in Krause, 2008). Berg et al. (2004) demonstrated that nuclear-encoded RNA polymerase (NEP) performs the plastid gene transcription that is usually performed by PEP in *Cuscuta* species. Barrett and Davis (2012) initially proposed a two-stage loss (photosynthesis genes and then PEP genes), but Barrett et al. (2014), instead proposed the concerted loss of photosynthesis genes (excluding ATP synthase) and PEP, as they found no evidence that full mycoheterotrophs with recent lost of photosynthesis have retained uninterrupted *rpo* genes with degraded (or lost) photosynthesis genes (see also Wicke et al., 2013). However, the plastome of *Exo. oliganthum* (Gentianaceae) provides a probable example of the latter (photosynthesis genes degrading before PEP), as its *rpo* genes are retained and are still present in open reading frame (Figs. 1, S4; Table 2). Thus, this provides initial support for the two-stage hypothesis proposed by Barrett and Davis (2012). My finding should be followed up with a functional study of PEP gene activity in *Exo. oliganthum*.

4.3.4 Loss of ribosomal protein and tRNA genes

Land-plant plastomes encode some of the components of the plastid translational machinery, forming complete complexes with nuclear-encoded products. Complete plastid ribosomes are formed by 58-62 ribosomal proteins and four ribosomal RNA subunits, and among these the plastomes of land plants commonly encode all four ribosomal RNA subunits (rDNA genes) and 21 ribosomal proteins (12 small subunit or *rps* genes; 9 large subunit or *rpl* genes) (Palmer, 1985; Wicke et al., 2011; Sugiura, 2014). Ribosomal DNA loci are highly conserved (Palmer, 1985; Harris et al., 1994), and they are retained in all sequenced heterotrophic plant plastomes to

date, including the completely assembled mycoheterotrophs presented here (see Barrett et al., 2014; Lam et al., 2015; Petersen et al., 2015; Table 2). Plastid-encoded ribosomal protein genes are rarely lost in autotrophs (but see Jansen et al., 2007, 2011; Fajardo et al., 2013; Martínez-Alberola et al., 2013), but some have been deleted or found as pseudogenes in the relatively more degraded plastomes of some heterotrophic plants (Delannoy et al., 2011; Wicke et al., 2013; Lam et al., 2015; Schelkunov et al., 2015). *Exochaenium oliganthum* (Gentianaceae) retains all 21 plastid-encoded ribosomal proteins in open reading frame, and *Ep. pallida* (Polygalaceae) a slightly smaller set of 18, although it has two losses in common with autotrophic relatives. A set of eleven plastid-encoded ribosomal proteins is retained across mycoheterotrophic plants: *rps*2, 3, 4, 7, 8, 11, 14 and *rpl*2, 14, 16 and 36 (see Lam et al., 2015). These genes are retained in all fully assembled plastomes presented here, plus an additional four: *rps*12, 18, 19 and *rpl*20 (Table 2).

Land-plant plastomes generally retain loci for 30 transfer RNA (tRNA or *trn* genes). A complete, or nearly complete, set of these is retained in heterotrophs with relatively less degraded plastomes (e.g. *Aneura mirabilis*, Wickett et al., 2008; *Petrosavia stellaris*, Logacheva et al., 2014). *Exochaenium oliganthum* (Gentianaceae) and *Ep. pallida* (Polygalaceae) retain complete sets, although the latter has lost many more photosynthesis genes than the former (Table 3). The loss of many transfer RNA genes is typical of highly reduced mycoheterotrophs and parasites (Delannoy et al., 2011; Wicke et al., 2013; Lam et al., 2015; Schelkunov et al., 2015). The two *Voyria* (Gentianaceae) species and *Ep. elongata* (Polygalaceae) retain few of the thirty transfer RNA genes normally coded for by the plastome (Tables 3, S4: four tRNA genes in *V. caerulea*, 13 in *V. clavata* and five in *Ep. elongata*, although note that the latter two are based on incomplete assemblies). The loss of some tRNA genes might be compensated for by import

from the cytosol (e.g. Alkatib et al., 2012) or 'superwobbling' (Rogalski et al., 2008), although a few of them may not be replaceable by either means (Barbrook et al., 2006). Among these is *trn*E-UUC, whose gene product (glutamyl tRNA) has a secondary role outside translation, in heme biosynthesis (Jahn et al., 1992), and possibly in the regulation of nuclear-encoded plastid RNA synthase (NEP) (Hanaoka et al., 2005, but see Bohne et al., 2009). Barbrook et al. (2006) proposed that the interaction of glutamyl tRNA with multiple enzymes involved in the production of heme makes a replacement by a cytosolic product unlikely. Howe and Purton, (2007) gave a related explanation for the retention of plastid-encoded formylmethionyl-tRNA (*trnf*M-CAU), which has a role in initiating translation in plastids and possibly some mitochondria (Barbrook et al., 2006). It has been suggested that the need to be recognized by multiple enzymes limits the likelihood of replacement (Barbrook et al., 2006; Howe and Purton, 2007; Delannoy et al., 2011); presumably this would require independent adjustment to replacement, by each enzyme that interacts with the tRNA. Barbrook et al. (2006) proposed that the indispensability of these two plastid-encoded transfer RNAs could explain the retention of plastomes in non-photosynthetic organisms, which they call the "essential tRNAs hypothesis." In total, four tRNA genes are retained in all species presented here: *trn*W-CCA, *trn*I-CAU, *trn*fM-CAU, *trn*E-UUC (Table 2). The latter three are retained in all sequenced heterotrophic plants to date, and lend support to the essential tRNA hypothesis.

4.3.5 Loss of other plastid genes of known and unknown function

Plastids are not just photosynthetic organelles. They are the site of additional essential cellular functions including fatty acid, amino acid and tetrapyrrole biosynthesis, pigment production and the conversion of inorganic nitrogen to useful forms (reviewed in Ernes and Neuhaus, 2005). In

addition to genes involved in the plastid genetic apparatus, loci that are retained in the most degraded heterotrophic plastomes encode proteins with roles in essential non-photosynthetic metabolism, including protein turnover and import or intron removal. MATK, the only plastid-encoded group IIa intron maturase (Zoschke et al., 2010), is coded for by *mat*K, a locus retained in nearly all plant plastid genomes. However, the *mat*K gene has been deleted from the plastome of the mycoheterotrophic orchids *Rhizanthella gardneri*, *Epipogium aphylla* and *E. roseum*, and from some holoparasitic *Cuscuta* species (Funk et al., 2007; McNeal et al., 2007; Delannoy et al., 2011; Braukmann et al., 2013; Schelkunov et al., 2015). *Voyria clavata* (Gentianaceae) has lost *mat*K, and I also did not recover it in *V. caerulea* (based on the gene set assembled for this without a full circular genome). It is also likely a pseudogene in *Ep. elongata* (Polygalaceae), as a non-triplet deletion ~900 bp into the reading frame results in a frame shift. In *Cuscuta*, the loss of *mat*K coincides with the loss of all group IIa introns (McNeal et al., 2009). In contrast, *Epipogium* and *Rhizanthella* retain loci with group IIa introns, and at least two of these genes (*rpl*2 and *rps*12) are thought to be targeted by MATK (Zoschke et al., 2010, although the retention of the third exon of *rps*12, and its MATK targeted intron, is uncertain in *Epipogium;* Schelkunov et al., 2015). This parallels the situation in *Voyria* and *Ep. elongata* where *rpl*2 and *rps*12 are retained with group IIa introns intact. Delannoy et al. (2011) demonstrated in *Rhizanthella* that *rpl*2 is correctly spliced, suggesting that an alternative splicing factor facilitates intron removal from their RNA transcripts. Furthermore, *rpl*2 is one of the plastid-encoded ribosomal protein loci retained across heterotrophic land plants, and therefore is likely functional in *Voyria* and *Ep. elongata.* My finding could be followed up with selection tests to ascertain whether these genes in *mat*K-deleted plastomes are under the same selective regime as homologous genes in *mat*K-retaining plastomes.

The gene coding for the beta subunit of Acetyl-CoA carboxylase (*acc*D) is retained in all sequenced plastomes of heterotrophic plants. The protein product is assembled with nuclear-encoded subunits to form a complex that catalyzes the formation of essential components of fatty acids (Ohlrogge and Browse, 1995; Sasaki and Nagano, 2004). However, losses have been documented in autotrophs, where functional transfer of plastid-encoded *acc*D to the nucleus has been proposed and demonstrated in some lineages (Straub et al., 2011; Rousseau-Gueutin et al., 2013; Sabir et al., 2014). All Gentianaceae plastomes sampled here retain the genes coding for the beta subunit of Acetyl-CoA carboxylase (*acc*D), including the very reduced plastome of *Voyria clavata* (Gentianaceae) and gene set of *V. caerulea* (Table 2). In the gene set of *Ep. elongata* (Polygalaceae) I recovered a ~500 bp truncated *acc*D that is likely a pseudogene. The functional status of the *acc*D gene is otherwise unclear in the Polygalaceae and Ericaceae representatives presented here, regardless of trophic category. Only subregions (~30-60%) of the open reading frames recovered in autotroph and mycoheterotroph representatives of these families presented here BLAST to the homologous gene in relatives, where *acc*D is clearly retained (*Ceratonia siliqua*, Fabaceae, and *Camellia crapnelliana*, Theaceae). In contrast *C. siliqua* and *C. crapnelliana* BLAST to *Nicotiana tabacum* (Solanaceae) with 94% and 99% query cover, respectively. A regulatory role has been proposed to explain the general retention of the plastid-encoded subunit of *acc*D (Bungard, 2004; Delannoy et al., 2011), but the nuclear relocation of the gene in some autotrophs suggests that this plastome-specific role is not essential, in at least some lineages. Nevertheless, the retention of long open-reading frames for *acc*D-like genes in these two families, despite substantial sequence change, suggests the retention of function of some kind, which warrants further investigation.

*Clp*P is a plastid-encoded subunit of the Clp protease (or ATP-dependent caseinolytic protease), which has roles in protein turnover and processing, but has also been linked to isopyrenoid and tetrapyrrole biosynthesis, and fibrillins (lipid-body stabilizing molecules) (Kim et al., 2009; Stanne et al., 2009; Krause, 2012). The gene is found in most, but not all heterotrophic plant plastomes, and is considered essential for plant development (Kuroda and Maliga, 2003); a gene with a similar protein product (*clp*C) is retained in the reduced plastome of apicomplexan parasites (reviewed in Sato, 2011). As with *acc*D, this gene is deleted from the plastomes of several lineages of autotrophs, where a nuclear gene presumably codes for the protein, although this has not been demonstrated (Jansen et al., 2007; Straub et al., 2011). All Gentianaceae plastomes retain the *clp*P locus, but it is deleted from the plastomes of fully assembled Polygalaceae, and was not recovered in the genes sets of Ericaceae representatives.

The plastid-encoded translation initiation factor *inf*A has been lost independently many times in land plants (Wicke et al. 2011), and transfer to the nucleus has been demonstrated in several eudicot (asterid and fabid) lineages lacking the plastid locus (Millen et al., 2001; Jansen et al., 2007). Loss of *inf*A may be associated with heterotrophy in Gentianaceae, as it is retained in autotrophic *Exacum* but lost (or found with reading frame interruptions) in four of the five heterotrophs (Table 3). The reading frame is uninterrupted in Ericaceae representatives, and thus may be functional. Fully assembled Polygalaceae retain *inf*A loci with multiple reading frame interruptions, and I did not recover the gene in the gene sets of *Salomonia cantoniensis* or *Ep. elongata*.

*Ycf*1 and *ycf*2 are large hypothetical chloroplast reading frames for which reading frame interruptions have lethal consequences in tobacco (Drescher et al., 2000), yet their precise functions remain uncertain. These loci are retained in most land plants and some heterotrophs,

but have been deleted or pseudogenized in a few autotrophic lineages (Downie et al., 1994; Jansen et al., 2007). Sequence similarity in the binding domain of *ycf*1 to genes in the CDC48 family prompted Wolfe (1994) to suggest a cell membrane-related function for the gene. Recently Kikuchi et al. (2013) demonstrated association of its gene product with a nuclear-encoded inner-envelope membrane translocon complex (TOC/TIC machinery), and Nakai (2015) proposed renaming *ycf*1 as *tic*214 (but see de Vries et al, 2015). Less is known about *ycf*2, but drought-stress expression profiling suggests a role in water-use efficiency (Ruiz-Nieto et al., 2015). I recovered the *ycf*1and *ycf*2 loci with open reading frames in all fully assembled Gentianaceae species except *Voyria*: *ycf*1 is deleted from *V. clavata* and may or may not be retained in *V. caerulea* (as I recovered an incomplete gene without reading frame interruption). *Ycf*2 is severely truncated in both *Voyria* species (Table 2). Both genes are likely functional in autotrophic Polygalaceae, but there are truncated, probable pseudogenes of them in *Ep. pallida* and I did not recover either locus in *Ep. elongata*. Braukmann and Stefanović (2012) noted lack of *ycf*2 probe hybridization in Arbutoideae and Pyroleae, based on a survey of plastome gene content in Ericaceae. I did not recover a *ycf*2 locus in *Orthilia*, and the gene is absent in *Arbutus unedo* (Martínez-Alberola et al., 2013). However, I did recover a severely truncated locus in the *Pyrola* gene set.

4.4 Structural rearrangement and the inverted repeat

Sequenced plastid genomes of heterotrophs (parasites and mycoheterotrophs) have a range of levels of genome rearrangement, from those that are essentially colinear with green relatives despite gene loss (e.g. *Sciaphila densiflora*, Triuridaceae, Lam et al. 2015; *Corallorhiza* spp., Orchidaceae, Barrett et al. 2014; *Aneura mirabilis,* Aneuraceae, Wickett et al., 2009; *Epifagus*

virginiana, Orobanchaceae Wolfe et al., 1992) to highly rearranged ones (e.g. *Orobanche crenata*, Orobanchaceae, Wicke et al. 2013; *Petrosavia stellaris*, Petrosaviaceae, Logacheva et al., 2014), and including several intermediate levels of rearrangement (e.g., some *Cuscuta*, Convolvulaceae, Funk et al., 2007; some Orobanchaceae, Wicke et al., 2013). It is not well understood whether there are general processes affecting genome structure in mycoheterotrophs, and so the newly sequenced genomes here provide additional independent data points for addressing this issue.

Gene loss is associated with considerable changes in genome structure in many of the mycoheterotroph genomes included here (Fig. 1, Table 2), and inverted repeat boundaries have also shifted in some cases (Figs. 1, S1-S7, Table 4). Inverted repeat boundary shifts are reasonably common at the IR/large single copy boundary in autotrophic lineages, but shifts at the IR/small single copy boundary (as found here for autotrophs *Polygala* and *Arbutus*; note that their inverted repeats extend four and eleven genes further into what is the small single copy region in *Nicotiana*, for example, Figs. 2, Table 4) are less common (Zhu et al., 2015).

Setting aside frequent genome compaction due to gene loss, and the typically minor shifts in IR boundaries, the completely (or nearly completely) sequenced plastid genomes of the eudicot mycoheterotrophs examined here (Fig. 1, Table 4) generally do not appear to be evolving in a substantially different manner to their closest green relatives. For example, in Gentianaceae, all four partially and fully mycoheterotrophic taxa (two partial mycoheterotrophs, *Bartonia* and *Obolaria*, and two full mycoheterotrophs, *Exochaenium* and *Voyria*) have colinear or nearly colinear genomes with an autotrophic member of Gentianaceae (*Exacum*; Fig. 3), that in turn is colinear with the plastome of tobacco (*Nicotiana*, Solanaceae; Fig. 2), which has a gene order that is similar to most other angiosperms (Palmer, 1985; Palmer and Stein, 1986; Jansen et al.,

2007). One major rearrangement in *Voyria clavata* concerns the boundaries of the inverted repeats (and hence single copy regions), which no longer span genes found in the inverted repeat regions of other members of the family. *Voyria* also has a single inversion compared to the other Gentianaceae (the right-hand LCB in Fig. 3). Nonetheless, gene order is otherwise largely conserved (one of the *Voyria* inverted repeat copies falls in a locally colinear block, LCB; note that the other is not shown in Fig. 3), and this minimal pattern of genome restructuring is comparable to that of *Sciaphila* (Triuridaceae; Lam et al., 2015), *Rhizanthella* (Orchidaceae; Delannoy et al., 2011) and *Epifagus* (Orobanchaceae; Wolfe et al., 1992) in terms of having retained colinearity despite extensive gene loss.

I inferred multiple plastid genome rearrangements in *Epirixanthes* (Polygalaceae) and *Orthilia* (Ericaceae) compared to their close green relatives (*Polygala* and *Arbutus*, respectively; Figs. 3). However, in both cases, their green relatives also have fairly substantial rearrangements compared to tobacco (*Nicotiana*, Solanaceae; Fig. 2), so it may difficult to distinguish any effects of mycoheterotrophy from other processes leading to genome rearrangement in these taxa. In Ericaceae, two fully assembled plastomes have been published (*Arbutus unedo,* Martínez-Alberola et al., 2013; *Vaccinium macrocarpon,* Fajardo et al., 2013). Both show multiple major rearrangements in comparison to tobacco (see *Arbutus unedo* in Fig. 2). Unusually, the inverted repeat regions of *Arbutus* and *Vaccinium* have expanded to encompass nearly all of the ancestral small single copy region in both autotrophs. Martínez-Alberola et al. (2013) noted that among 37 asterid plastomes sampled, only *Arbutus unedo, Vaccinium macrocarpon* and two other species had tandem repeats larger than 150 bp ('megasatellites'), which are associated with rearrangement in pathogenic yeast (Thierry et al., 2008). Dispersed repeats may contribute to plastid genome rearrangements in other taxa (Downie and Palmer, 1992; Cosner et al., 1997; Cai

et al., 2008; Haberle et al., 2008), and it is possible that dispersed repeat proliferation is a characteristic of Ericaceae plastomes, including the mycoheterotrophs here, though I did not attempt to characterize this possibility here. However, given the relatively modest level of plastome degradation observed in *Orthilia secunda* and the similar level of rearrangement found in fully autotrophic Ericaceae (Fig. 2), it is likely that the number of rearrangements are attributable to shared characteristics of the family, and not to trophic status.

Dispersed repeats may also explain substantial plastid genome rearrangements in some taxa in an inverted-repeat-lacking clade (IRLC) of legumes (Fabaceae), the completely autotrophic sister group of Polygalaceae (e.g., Cai et al. 2008; Schwarz et al., 2015) although it has also been suggested that the lack of an inverted repeat also contributes to genome instability (Palmer et al., 1987; Milligan et al., 1989; Cai et al., 2008; Sabir et al., 2014). Outside this clade of legumes, other members of Fabaceae are largely conserved in plastid genome structure, although several inversions and gene losses have been documented in subfamily Papillionoideae (Schwarz et al., 2015). The loss of the plastid inverted repeat (IR) in *Epirixanthes*, and its switch to a mycoheterotrophic nutritional mode, may not contribute to genome rearrangement in this taxon (Figs. 4, S2), as *Polygala arillata* is autotrophic and retains an inverted repeat, and yet also has substantial rearrangements compared to *Nicotiana* (Fig. 2), Nonetheless, the loss of an IR in *Epirixanthes* provides an intriguing parallel to the IRLC in the sister group of Polygalaceae. The gain of a single large direct repeat in *Epirixanthes* is also unusual and noteworthy (Figs. 4, S2). Large repeats are thought to be selected against as destabilizing elements in plastomes that cause aberrant recombination (Gray et al., 2009; Maréchal and Brisson, 2010). As with Ericaceae, it would be useful to explore the possibility that dispersed repeats have contributed to genome rearrangements in autotrophic and mycoheterotrophic Polygalaceae.

4.5 Conclusion

A rationale for the retention of genetic apparatus genes is the continued need to express plastid genes with putatively essential roles that are not involved in photosynthesis (e.g., *acc*D, *clp*P, *trn*E) (Delannoy et al., 2011; Krause, 2008). As independent losses of *acc*D and *clp*P have occurred in multiple lineages of photosynthetic plants, including species sampled here, the endpoint of plastome reduction may vary by lineage in a manner that is unrelated to heterotrophy. Some essential plastid genes may not be readily replaceable in non-photosynthetic plants (e.g., *trn*E), and it is not yet clear if any land plants have completely lost their plastomes (see Molina et al., 2014, for a possible exception), although this is known in some heterotrophic protists (Janouškovec et al., 2015). The patterns of gene loss characterized here are generally consistent with the trajectory hypothesized by Barrett and Davis (2012) model: plastid NAD(P)H dehydrogenase is likely lost before the loss of photosynthesis in partially mycoheterotrophic plants, most photosynthesis genes are then lost after the initial switch to full mycoheterotrophy, and plastid-encoded RNA polymerase genes are lost next. ATP synthase subunit genes and *rbc*L appear to repeatedly linger after the loss of photosynthesis, likely because of secondary non-photosynthetic roles that they play in the plastid. I propose here that they may be lost in either order after the loss of most photosynthesis genes (this is a modification of the hypothesis of Barrett and Davis, 2012). In the late stages of full mycoheterotrophy, multiple genes in the plastid translation apparatus are lost from the plastome (the most extreme example here is *Voyria*, Gentianaceae), although a core set of ribosomal protein, rDNA and tRNA genes is retained in all mycoheterotrophs examined here. Other non-photosynthetic genes may be lost in a more sporadic manner in the later stages of gene loss, and may include some surprising losses

(e.g., of *mat*K, given that some group IIa introns are retained). Future work should follow these observations up with physiological studies to assess gene-product functionality (for example to determine the possible functions of plastid ATP synthase and Rubisco in the full mycoheterotrophs that retain them). The full and partial mycoheterotrophs sampled here also vary considerably in terms of plastome size and gene content, from extremely reduced to only marginally degraded, and from substantially rearranged plastomes to those that are nearly colinear with green relatives. These differences do not appear to be related to the loss of photosynthesis or the loss of the plastid inverted repeat regions. Because substantial diversity was uncovered among close relatives that represent the same loss of photosynthesis (in *Voyria* and *Epirixanthes*), it would be useful to continue sampling in these genera and other eudicot mycoheterotrophs. Despite gene loss and moderate rate elevation, the plastid gene sets recovered here are shown to be useful for inferring phylogenetic relationships of the mycoheterotrophic eudicots.

Table 1. Specimen source information; herbarium abbreviations follow Thiers (2015)

Trophic status[1]	Family	Species	Specimen voucher [Collector number (herbarium)]
Full MH	Gentianaceae	*Exochaenium oliganthum* (Gilg.) Kissling	Sainge s.n. (YA)
Full MH	Gentianaceae	*Voyria caerulea* Aubl.	Merckx 244 (L)
Full MH	Gentianaceae	*Voyria clavata* Splitg.	Merckx 224 (L)
Full MH	Polygalaceae	*Epirixanthes elongata* Blume	Hsu 17814 (FLAS)
Full MH	Polygalaceae	*Epirixanthes pallida* T. Wendt	Merckx & Mennes CM001 (L)
Partial MH	Ericaceae	*Orthilia secunda* (L.) House	No voucher[1]
Partial MH	Ericaceae	*Pyrola minor* L.	No voucher[2]
Partial MH	Gentianaceae	*Bartonia virginica* (L.) Britton, Sterns & Poggenb.	Bertin 6708 (MASS)
Partial MH	Gentianaceae	*Obolaria virginica* L.	Stefanovic SS-04-103 (TRT)
Full autotroph	Gentianaceae	*Exacum affine* Balf.f. ex Regel	Darby s.n. (UBC)
Full autotroph	Polygalaceae	*Polygala arillata* Buch.-Ham. ex D. Don.	Larsen 46516 (FLAS)
Full autotroph	Polygalaceae	*Salomonia cantoniensis* Lour.	Nosuro 9830009 (FLAS)

[1] See Beatty and Provan (2010)
[2] See Beatty et al. (2010)

Table 2. Plastid gene content across newly sequenced taxa of Gentianaceae, Polygalaceae and Ericaceae (the *Arbutus unedo* plastome is from Martínez-Alberola et al., 2013). Full mycoheterotrophs are bolded, and partial mycoheterotrophs are underlined. An asterisk (*) indicates that a full plastid genome was assembled. Genes with open-reading frames are indicated by '+' (incompletely recovered genes with open reading frames by '(+)'). Gene absence (loci for which remnants could not be detected in full genomes, or that could not be retrieved in plastid gene set assemblies) is indicated with a dash ('-'). Probable pseudogenes (loci with multiple internal stop codons, see text) are indicated as 'ψ'. Loci with single reading frame interruption included in ORF-only matrix (there are four) are indicated with '#'. Genes found intact in all fully assembled species are indicated in bold font.

	Gentianaceae						Polygalaceae				Ericaceae		
	**Exacum affine*	**Bartonia virginica*	**Obolaria virginica*	****Exochaenium oliganthum***	***Voyria caerulea***	****Voyria clavata***	**Polygala arillata*	*Salomonia cantoniensis*	****Epirixanthes pallida***	***Epirixanthes elongata***	**Arbutus unedo*	*Orthilia secunda*	*Pyrola minor*
NAD(P)H dehydrogenase													
*ndh*A	+	-	ψ	ψ	-	-	+	+	-	-	+	ψ	ψ
*ndh*B	+	ψ	ψ	+	ψ	-	+	+	ψ	-	+	+	+
*ndh*C	+	-	ψ	+	-	-	+	+	ψ	-	+	+	+
*ndh*D	+	ψ	ψ	ψ	-	-	+	+	-	-	+	ψ	ψ
*ndh*E	+	-	ψ	ψ	-	-	+	+	-	-	+	ψ	#
*ndh*F	+	-	ψ	ψ	-	-	+	+	ψ	-	+	ψ	ψ
*ndh*G	+	-	ψ	+	-	-	+	+	-	-	+	ψ	ψ
*ndh*H	+	ψ	ψ	ψ	-	-	+	+	ψ	-	+	ψ	ψ
*ndh*I	+	-	-	+	-	-	+	+	-	-	+	ψ	ψ
*ndh*J	+	ψ	-	#	-	-	+	+	-	-	+	#	ψ
*ndh*K	+	-	ψ	ψ	-	-	+	+	ψ	-	+	+	+
Photosystem (PS) I													
*psa*A	+	+	+	ψ	-	-	+	(+)	ψ	-	+	+	+
*psa*B	+	+	+	+	ψ	-	+	(+)	ψ	-	+	+	+
*psa*C	+	+	+	+	-	-	+	+	ψ	-	+	+	+
*psa*I	+	+	+	+	-	-	+	+	-	-	+	+	+
*psa*J	+	+	+	+	-	-	+	+	+	-	+	+	+
Photosystem (PS) II													
*psb*A	+	+	+	ψ	-	-	+	+	ψ	-	+	+	+
*psb*B	+	+	+	+	-	-	+	(+)	-	-	+	+	+
*psb*C	+	+	+	+	ψ	-	+	+	ψ	-	+	+	+
*psb*D	+	+	+	+	ψ	-	+	+	ψ	-	+	+	+
*psb*E	+	+	+	+	-	-	+	+	-	-	+	+	+
*psb*F	+	+	+	+	-	-	+	+	-	-	+	+	+
*psb*H	+	+	+	+	-	-	+	+	-	-	+	+	+
*psb*I	+	+	+	+	-	-	+	+	ψ	-	+	+	+
*psb*J	+	+	+	+	-	-	+	+	-	-	+	+	+
*psb*K	+	+	+	+	-	-	+	+	-	-	+	+	+
*psb*L	+	+	+	+	-	-	+	+	-	-	+	+	+
*psb*M	+	+	-	+	-	-	+	+	+	-	+	+	+
*psb*N	+	+	+	+	-	-	+	+	-	-	+	+	+
*psb*T	+	+	+	+	-	-	+	+	-	-	+	+	+
*psb*Z	+	+	+	+	-	-	+	+	ψ	-	+	+	+
PSI assembly factors													
*ycf*3	+	+	+	ψ	-	-	+	+	ψ	-	+	+	+
*ycf*4	+	+	+	+	-	-	+	+	-	-	+	+	+
Cytochrome b_6/f complex													
*pet*A	+	+	+	+	-	-	+	+	-	-	+	+	+
petB	+	+	+	+	-	-	+	(+)	-	-	+	+	+
*pet*D	+	+	+	+	ψ	-	+	+	ψ	-	+	+	+
petG	+	+	+	+	-	-	+	+	+	-	+	+	+
*pet*L	+	+	+	+	-	-	+	+	-	-	+	+	+
*pet*N	+	+	+	+	-	-	+	+	-	-	+	+	+
Rubisco													
*rbc*L	+	+	+	+	-	ψ	+	+	+	-	+	+	+
ATP synthase													
*atp*A	+	+	+	+	ψ	-	+	+	+	-	+	+	+
*atp*B	+	+	+	+	-	-	+	+	+	-	+	+	+
*atp*E	+	+	+	+	-	-	+	+	+	-	+	+	+
*atp*F	+	+	+	+	ψ	-	+	+	+	-	+	+	+
*atp*H	+	+	+	+	+	-	+	+	+	-	+	+	+
*atp*I	+	+	+	+	ψ	-	+	+	+	-	+	+	+
Other photosynthesis proteins													
*cem*A	+	+	+	+	-	-	+	+	-	-	+	+	+
*ccs*A	+	+	ψ	ψ	-	-	+	+	ψ	-	+	+	(+)

	*Exacum affine	*Bartonia virginica	*Obolaria virginica	***Exochaenium oliganthum**	***Voyria caerulea**	***Voyria clavata**	*Polygala arillata	*Salomonia cantoniensis*	***Epirixanthes pallida**	***Epirixanthes elongata**	*Arbutus unedo	*Orthilia secunda*	*Pyrola minor*
RNA Polymerase													
*rpo*A	+	+	+	+	ψ	-	+	+	ψ	-	+	+	+
*rpo*B	+	+	+	+	-	-	+	θ	ψ	-	+	+	+
*rpo*C1	+	+	+	+	-	-	+	+	ψ	-	+	+	+
*rpo*C2	+	+	+	+	-	-	+	+	ψ	-	+	+	+
Proteins of other function													
*acc*D	+	+	+	+	θ	+	+	+	+	ψ	ψ	θ	+
*clp*P	+	+	+	+	+	+	-	-	-	-	ψ	-	-
*inf*A	+	ψ	ψ	+	-	ψ	ψ	-	ψ	-	+	+	+
*mat*K	+	+	+	+	-	-	+	+	+	#	+	+	+
Proteins of unknown function													
*ycf*1	+	+	+	+	θ	-	+	θ	ψ	-	ψ	-	-
*ycf*2	+	+	+	+	ψ	ψ	+	ψ	ψ	-	-	-	ψ
Ribosomal proteins													
***rpl*2**	+	+	+	+	+	+	+	+	+	+	+	θ	+
***rpl*14**	+	+	+	+	+	+	+	+	+	+	+	+	+
***rpl*16**	+	+	+	+	+	+	+	+	+	+	+	+	+
***rpl*20**	+	+	+	+	+	+	+	+	+	-	+	+	+
*rpl*22	+	+	+	+	ψ	ψ	-	-	-	-	+	+	+
*rpl*23	+	+	+	+	ψ	ψ	+	+	+	-	+	ψ	ψ
*rpl*32	+	+	+	+	-	-	+	+	ψ	-	+	+	+
*rpl*33	+	ψ	+	+	ψ	-	+	+	+	-	+	+	+
***rpl*36**	+	+	+	+	-	+	+	+	+	+	+	+	+
rps2	+	+	+	+	+	+	+	+	+	+	+	+	+
***rps*3**	+	+	+	+	+	+	+	+	+	+	+	+	+
***rps*4**	+	+	+	+	+	+	+	+	+	+	+	+	θ
***rps*7**	+	+	+	+	+	+	+	+	+	θ	+	+	+
***rps*8**	+	+	+	+	+	+	+	+	+	θ	+	+	+
***rps*11**	+	+	+	+	+	+	+	+	+	ψ	+	+	+
***rps*12**	+	+	+	+	+	+	+	+	+	θ	+	+	+
***rps*14**	+	+	+	+	+	+	+	θ	+	+	+	+	+
*rps*15	+	+	+	+	-	-	+	+	+	+	+	+	+
*rps*16	+	-	ψ	+	-	-	-	-	-	-	ψ	+	+
***rps*18**	+	+	+	+	+	+	+	+	+	+	+	+	+
***rps*19**	+	+	+	+	-	+	+	+	+	+	+	-	+

	*Exacum affine	*Bartonia virginica	*Obolaria virginica	***Exochaenium oliganthum**	***Voyria caerulea**	***Voyria clavata**	*Polygala arillata	*Salomonia cantoniensis*	***Epirixanthes pallida**	***Epirixanthes elongata**	*Arbutus unedo	*Orthilia secunda*	*Pyrola minor*
Ribosomal DNA genes													
***rrn*4.5**	+	+	+	+	+	+	+	-	+	+	+	+	+
***rrn*5**	+	+	+	+	+	+	+	-	+	-	+	+	+
***rrn*16**	+	+	+	+	θ	+	+	+	+	+	+	+	+
***rrn*23**	+	+	+	+	+	+	+	θ	+	+	+	+	+
Transfer RNA genes													
trnA-UGC	+	+	+	+	-	-	+	+	+	-	+	+	+
trnC-GCA	+	+	+	+	+	-	+	+	+	-	+	+	+
trnD-GUC	+	+	+	+	+	-	+	+	+	+	+	+	+
trnE-UUC	+	+	+	+	+	+	+	+	+	+	+	+	+
trnF-GAA	+	+	+	+	+	-	+	+	+	-	+	+	+
trnfM-CAU	+	+	+	+	+	+	+	+	+	-	+	+	+
trnG-GCC	+	+	+	+	-	-	+	+	+	-	+	+	+
trnG-UCC	+	+	+	+	θ	-	+	+	+	-	+	+	+
trnH-GUG	+	+	+	+	-	-	+	+	+	-	+	+	+
trnI-CAU	+	+	+	+	+	+	+	+	+	+	+	+	+
trnI-GAU	+	+	+	+	-	-	+	+	+	-	+	+	+
trnK-UUU	+	+	+	+	-	-	+	+	+	-	+	+	+
trnL-CAA	+	+	+	+	+	-	+	+	+	-	+	+	+
trnL-UAA	+	+	+	+	-	-	+	+	+	-	+	+	+
trnL-UAG	+	+	+	+	-	-	+	+	+	-	+	+	+
trnM-CAU	+	+	+	+	-	-	+	+	+	-	+	+	+
trnN-GUU	+	+	+	+	+	-	+	+	+	-	+	+	-
trnP-UGG	+	+	+	+	+	-	+	+	+	-	+	+	+
trnQ-UUG	+	+	+	+	-	-	+	+	+	-	+	+	+
trnR-ACG	+	+	+	+	-	-	+	-	+	-	+	+	+
trnR-UCU	+	+	+	+	-	-	+	+	+	-	+	+	+
trnS-GCU	+	+	+	+	-	-	+	+	+	-	+	+	+
trnS-GGA	+	+	+	+	-	-	+	+	+	-	+	+	+
trnS-UGA	+	+	+	+	-	-	+	+	+	-	+	+	+
trnT-GGU	+	+	+	+	-	-	+	+	+	-	+	+	+
trnT-UGU	+	+	+	+	-	-	+	+	+	-	+	+	+
trnV-GAC	+	+	+	+	+	-	+	θ	+	-	+	+	+
trnV-UAC	+	+	+	+	-	-	+	+	+	-	+	+	+
trnW-CCA	+	+	+	+	+	+	+	+	+	+	+	+	+
trnY-GUA	+	+	+	+	+	-	+	+	+	+	+	+	+

Table 3. Species with fully assembled plastid genomes. Gent. = Gentianaceae; Poly. = Polygalaceae.

Family	Species	No. raw reads	X-Cov[1]	Length (bp)	LSC (bp)	SSC (bp)	IR (bp)	No. genes with intact reading frame[2]	No. rDNA genes	No. tRNA genes
Gent.	*Bartonia virginica*	18,522,856	657.23	145,525	80,530	3,491	30,752	65	4	30
Gent.	*Exacum affine*	10,824,214	1248.89	154,164	83,770	17,916	26,239	79	4	30
Gent.	*Exochaenium oliganthum*	18,882,270	143.99	151,797	81,921	17,512	26,182	68	4	30
Gent.	*Obolaria virginica*	17,075,844	152.25	145,825	79,411	10,014	28,158	64	4	30
Gent.	*Voyria clavata*	14,927,268	917.83	31,724	18,603	9,987	1,567	17	4	4
Poly.	*Epirixanthes pallida*	27,608,108	322.54	96,420	n.a.	n.a.	n.a.	29	4	30
Poly.	*Polygala arillata*	9,171,790	355.35	164,747	83,668	8,743	36,168	76	4	30

[1] Mean depth of coverage (based on remapping original reads to fully assembled plastome sequence)
[2] Genes found in the inverted repeat are counted once

Table 4. Inverted repeat (IR) boundary shifts in eudicot mycoheterotrophs and autotrophic relatives. Following Zhu et al. (2015) the last full gene included in the IR at the SSC and LSC boundaries is indicated (genes that are partially duplicated in IR are not shown here, but see Fig. 1). Numbers in parentheses indicate the number of genes that have been expanded (exp.) into (+) or contracted (cont.) out from (-) the ancestral angiosperm IR boundaries, compared to autotrophic relatives.

Family	Species	Trophic status	IR/SSC boundary (No. genes exp./cont.)	IR/LSC boundary (No. genes exp./cont.)
Ericaceae	*Arbutus unedo*	autotroph	*trn*L (+11)	*trn*I-CAU (-2)
Ericaceae	*Orthilia secunda*	partial MH	*trn*L (+4)	*trn*I-CAU (-2)
Gentianaceae	*Exacum affine*	autotroph	*trn*N-GUU	*rpl*22 (+2)
Gentianaceae	*Obolaria virginica*	partial MH	*trn*N-GUU	*rpl*2
Gentianaceae	*Bartonia virginica*	partial MH	*rps*15 (+2)	*rpl*2
Gentianaceae	*Exochaenium oliganthum*	full MH	*trn*N-GUU	*rpl*22 (+2)
Gentianaceae	*Voyria clavata*	full MH	*rps*11 (n.a.)	*rps*8 (n.a.)
Polygalaceae	*Polygala arillata*	autotroph	*ndh*I (+4)	*rpl*2
Polygalaceae	*Epirixanthes pallida*	full MH	n.a.	n.a.

Figure 1. Linearized plastome maps of photosynthetic and mycoheterotrophic representatives of Ericaceae (the *Arbutus unedo* plastome is from Martínez-Alberola et al., 2013), Gentianaceae and Polygalaceae. Blue horizontal bars are 10 kb increments. The inverted repeat region is indicated with grey bars. Genes are colour-coded by function (see caption). Genes with introns are indicated with an asterisk (*), and putative pseudogenes in red text. The *Orthilia secunda* draft assembly is a single contig, split into two fragments to match the orientation of the other plastome maps; the arrow points to the true ends of the assembly. A 12 kb direct repeat in *Epirixanthes pallida* is indicated with blue bars.

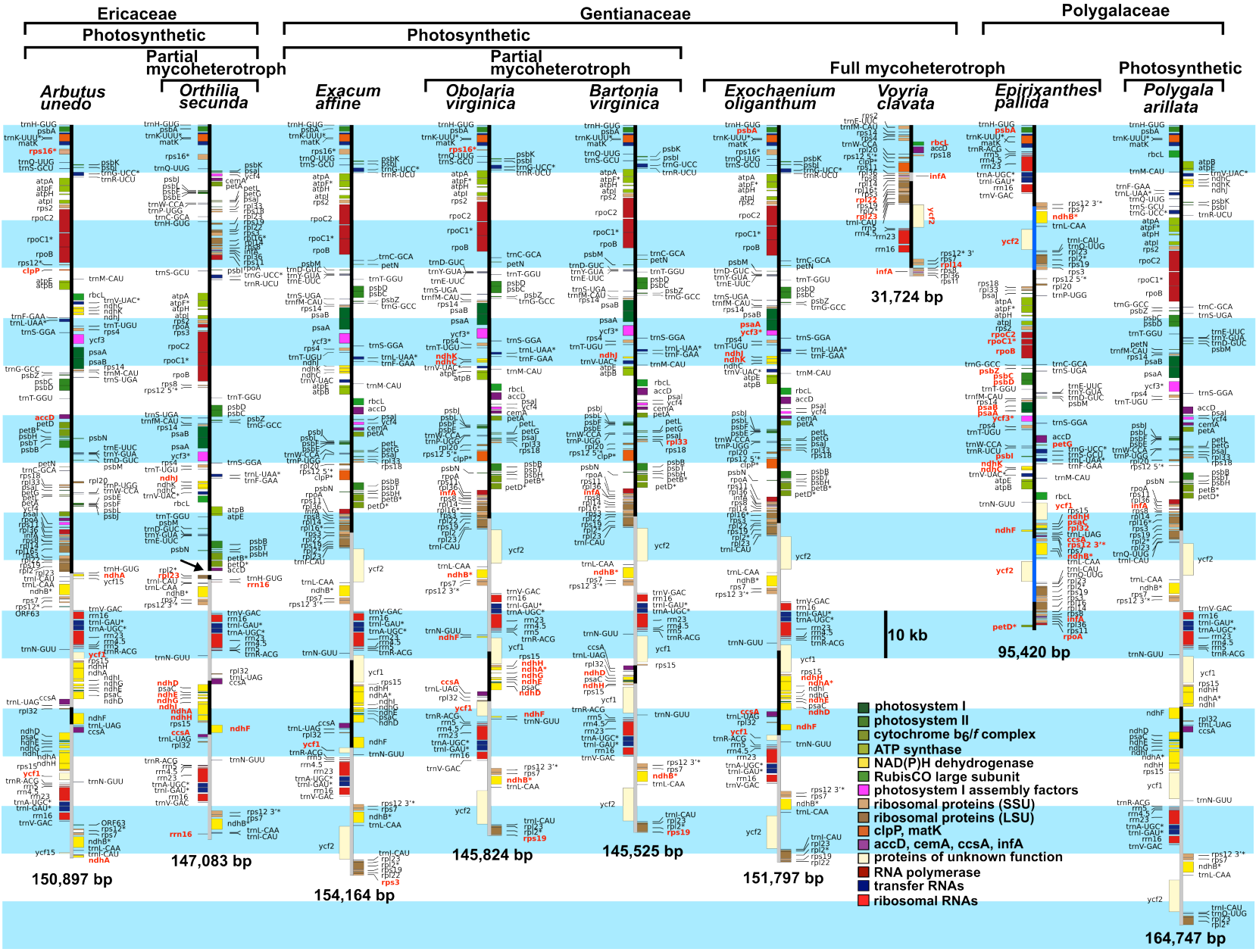
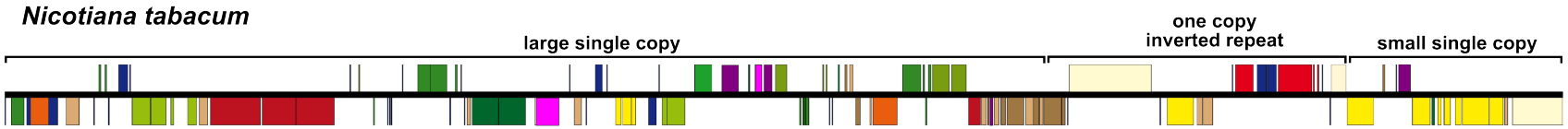
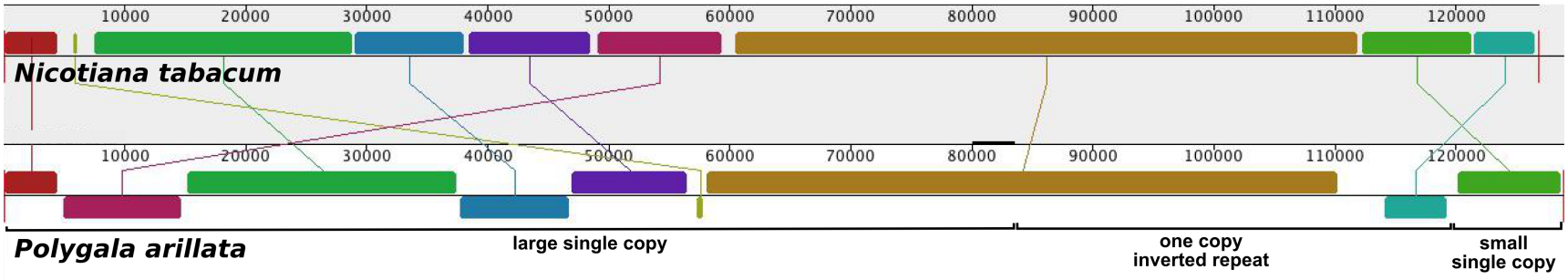

Ericaceae
Photosynthetic
Partial mycoheterotroph
Arbutus unedo
Orthilia secunda
Gentianaceae
Photosynthetic
Partial mycoheterotroph
Exacum affine
Obolaria virginica
Bartonia virginica
Full mycoheterotroph
Exochaenium oliganthum
Voyria clavata
Polygalaceae
Photosynthetic
Epirixanthes pallida
Polygala arillata
150,897 bp
147,083 bp
154,164 bp
145,824 bp
145,525 bp
151,797 bp
31,724 bp
95,420 bp
164,747 bp
10 kb
photosystem I
photosystem II
cytochrome b6/f complex
ATP synthase
NAD(P)H dehydrogenase
RubisCO large subunit
photosystem I assembly factors
ribosomal proteins (SSU)
ribosomal proteins (LSU)
clpP, matK
accD, cemA, ccsA, infA
proteins of unknown function
RNA polymerase
transfer RNAs
ribosomal RNAs

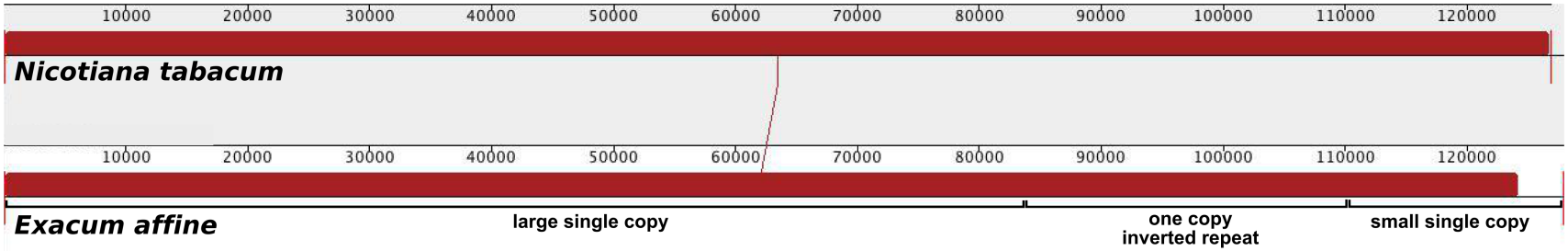
Figure 2. Pairwise Mauve-based alignments of *Nicotiana tabacum* (NC_001879) with autotrophic representatives of Polygalaceae, Gentianaceae and Ericaceae. A linear map of the *N. tabacum* reference sequence appears first. A single copy of the inverted repeat region was included in each comparison. Coloured blocks are homologous regions with shared gene order between two or more genomes, referred to as 'locally colinear blocks' (LCB). LCBs appearing above the central line are colinear and in the same orientation as the reference sequence. LCBs below align in reverse complement. Coloured lines link blocks of homology shared between taxa.



Polygalaceae



Gentianaceae



Ericaceae

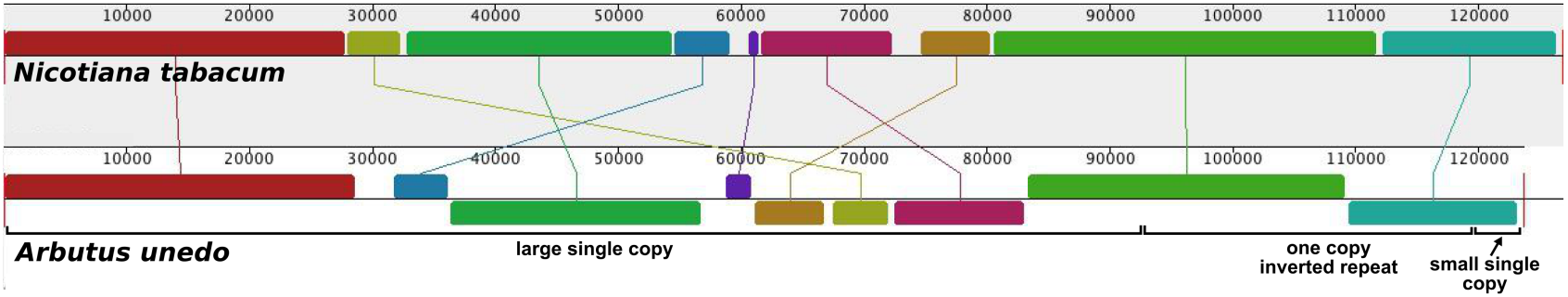


Figure 3. Mauve-based alignments of Gentianaceae plastomes (a linear map of autotrophic *Exacum affine* appears first for reference; this genome is colinear with *Nicotiana tabacum,* see Fig. 2). A single copy of the inverted repeat region was included in each comparison. Coloured blocks are homologous regions with shared gene order between two or more genomes, referred to as 'locally colinear blocks' (LCB). LCBs appearing above the central line are colinear and in the same orientation as the reference sequence. LCBs below align in reverse complement. Coloured lines link blocks of homology shared between taxa.

Gentianaceae

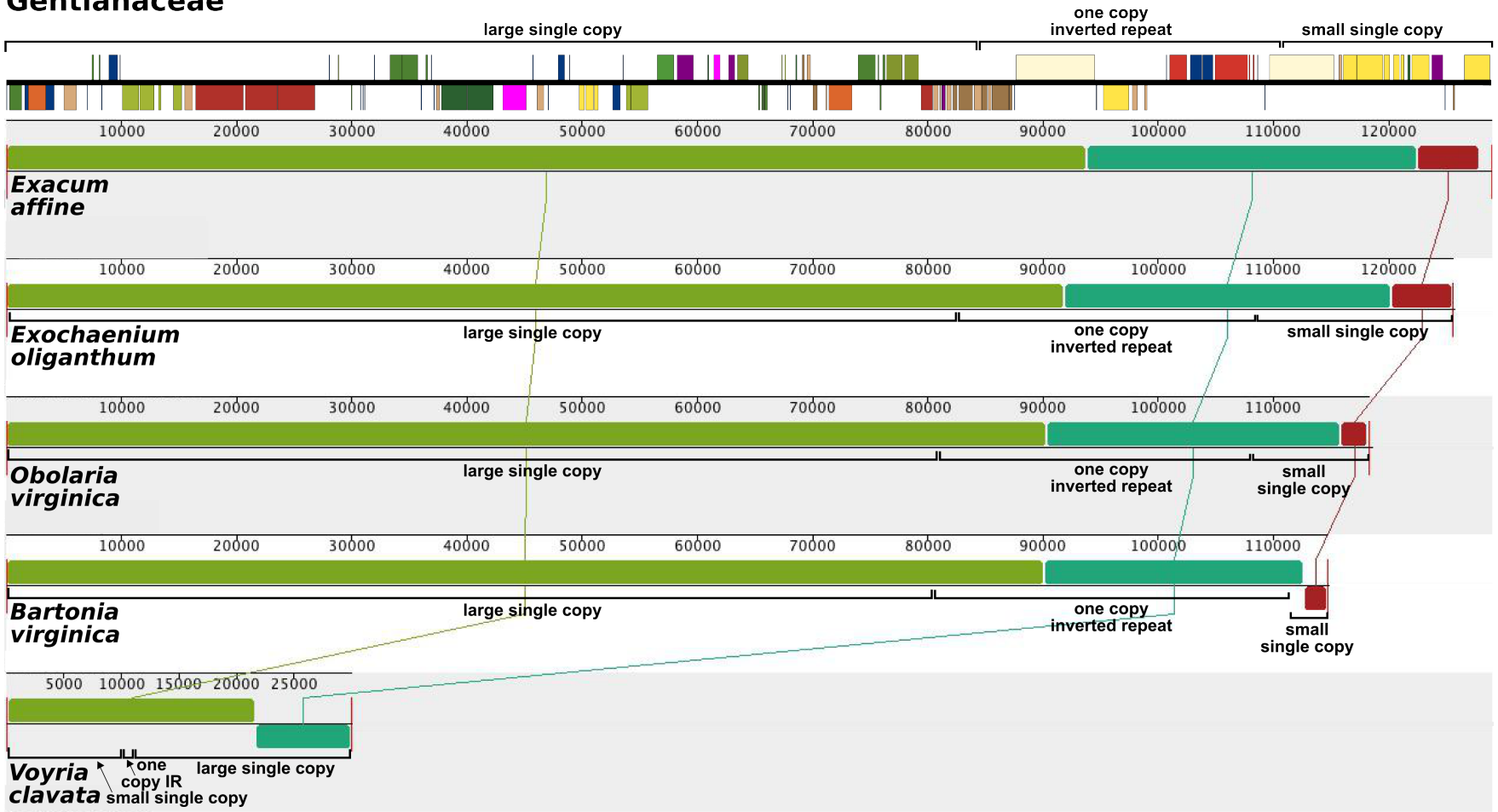


Figure 4. Mauve-based alignments of Polygalaceae and Ericaceae plastomes (a linear map of autotrophic *Polygala arillata*, Polygalaceae, or *Arbutus unedo*, Ericaceae, appears above the respective comparisons, for reference; these genomes are rearranged compared to *Nicotiana tabacum,* see Fig. 2). A single copy of the inverted repeat region was included in each comparison. Coloured blocks are homologous regions with shared gene order between two or more genomes, referred to as 'locally colinear blocks' (LCB). LCBs appearing above the central line are colinear and in the same orientation as the reference sequence. LCBs below align in reverse complement. Coloured lines link blocks of homology shared between taxa. Polygalaceae: 'a' and 'c' are regions of *P. arillata* sequence that are deleted from *E. pallida* plastome; 'b' corresponds to the inverted repeat region of *P. arillata*, which appears in direct repeat in the *E. pallida* plastome. Ericaceae: 'd'-'g' are intragenic regions that do not align under the set parameters.

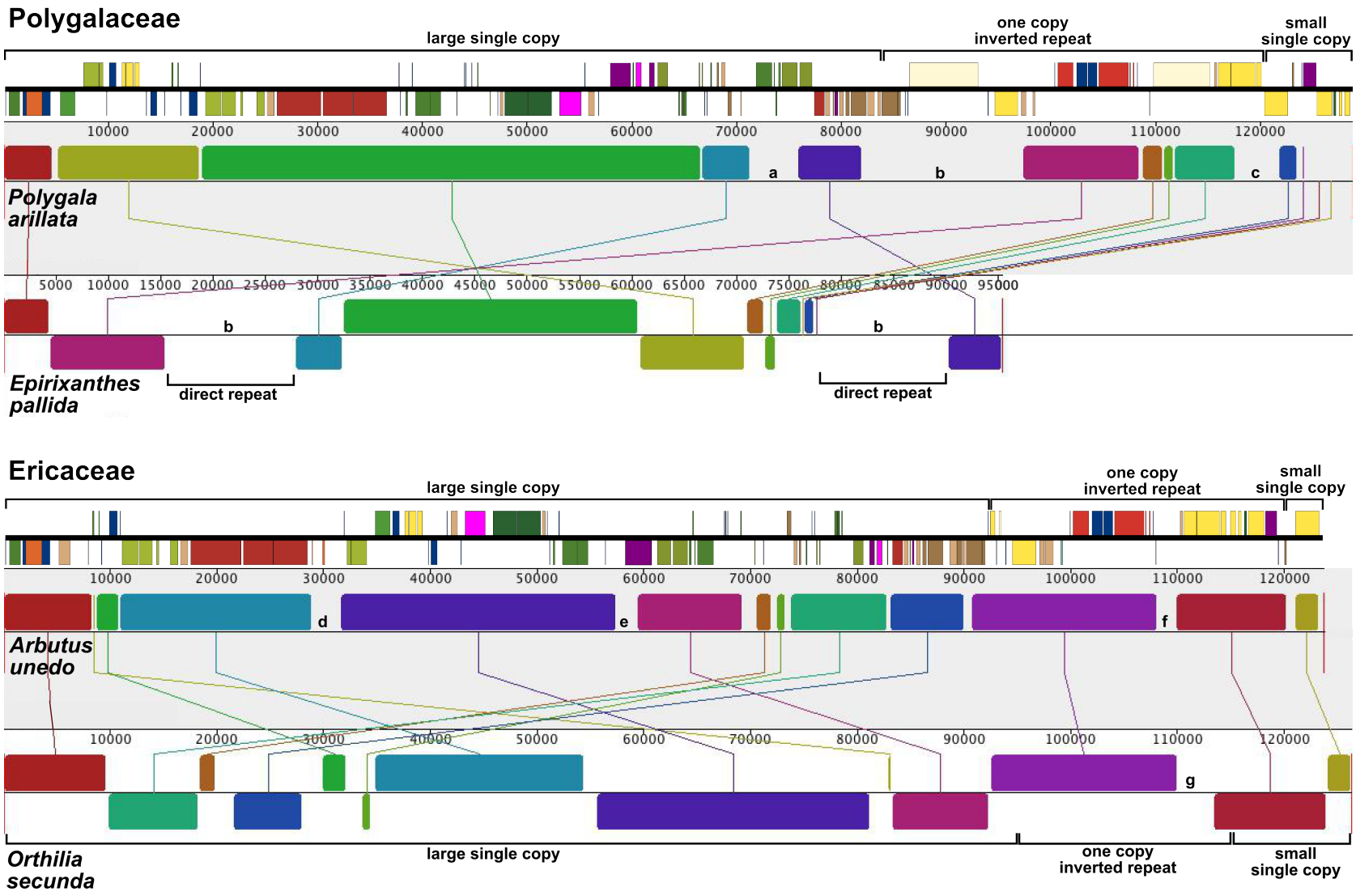

Polygalaceae
large single copy
one copy inverted repeat
small single copy
10000 20000 30000 40000 50000 60000 70000 80000 90000 100000 110000 120000
a
b
c
Polygala arillata
5000 10000 15000 20000 25000 30000 35000 40000 45000 50000 55000 60000 65000 70000 75000 80000 85000 90000 95000
b
b
Epirixanthes pallida
direct repeat
direct repeat
Ericaceae
large single copy
one copy inverted repeat
small single copy
10000 20000 30000 40000 50000 60000 70000 80000 90000 100000 110000 120000
d
e
f
Arbutus unedo
10000 20000 30000 40000 50000 60000 70000 80000 90000 100000 110000 120000
g
Orthilia secunda
large single copy
one copy inverted repeat
small single copy

Figure 5. A portion of angiosperm phylogeny inferred in a likelihood analysis of 82 plastid coding regions using the "GxC" partitioning scheme based on an 'ORF-only' alignment (see text and Table S3 for details); this portion of the tree shows rosid relationships. Eudicot families where mycoheterotrophy has evolved are indicated in blue. Log likelihood score of best tree: -1,506,318.267. Bootstrap support values are indicated beside branches; thick lines indicate 100% bootstrap support; '--' indicates <50% bootstrap support. The scale bar indicates estimated substitutions per site.

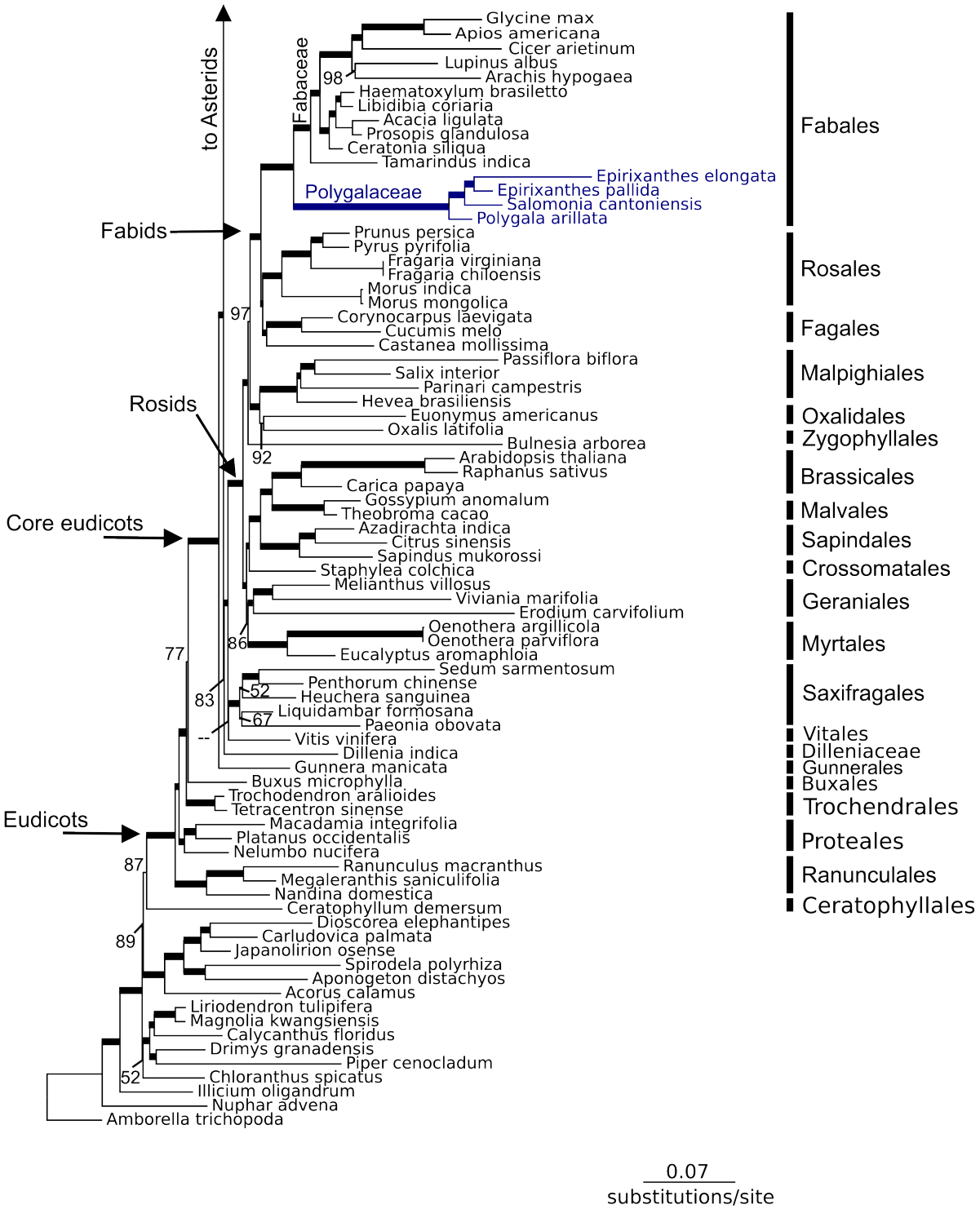
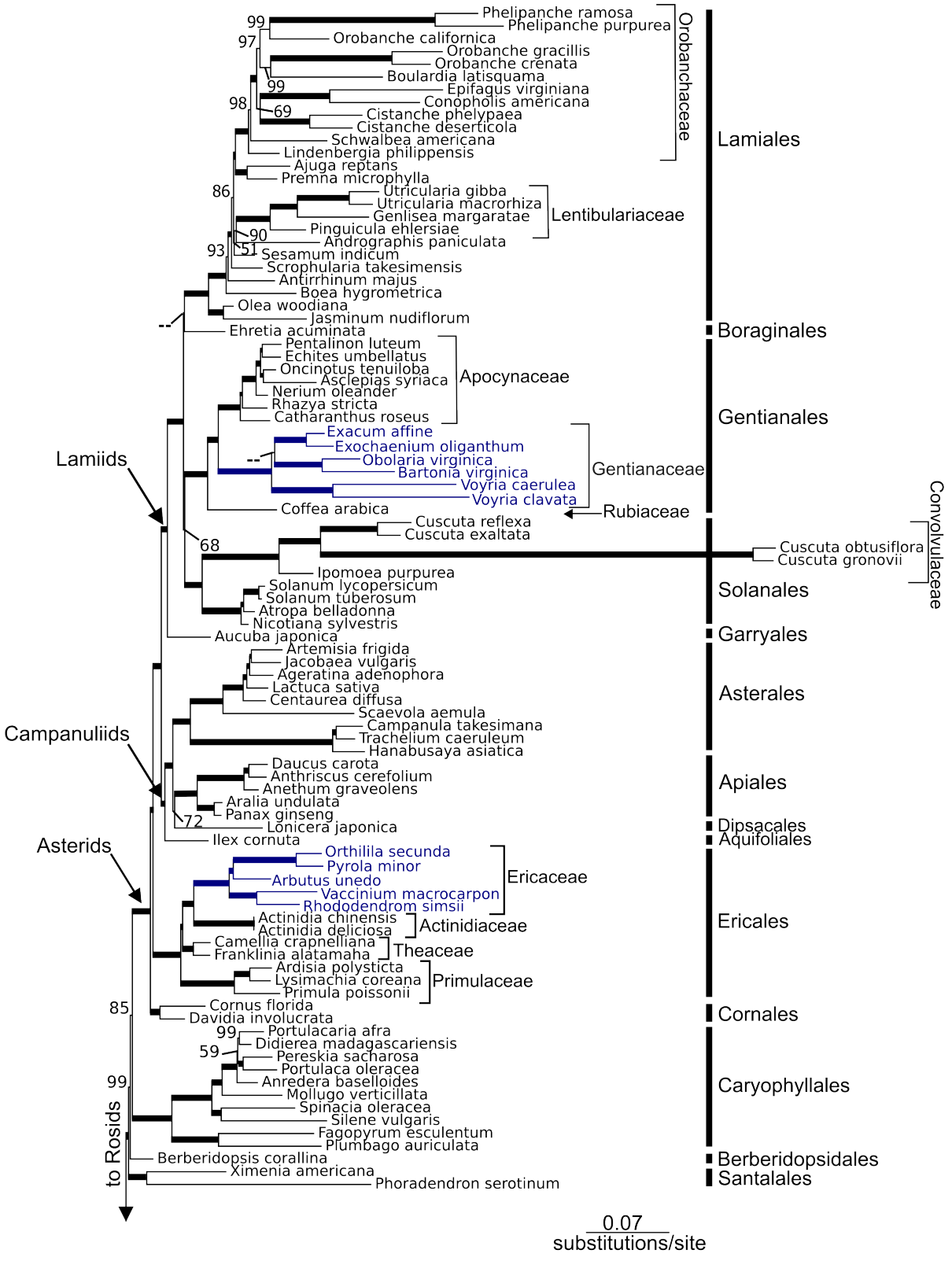

to Asterids
Fabaceae
Glycine max
Apios americana
Cicer arietinum
Lupinus albus
Arachis hypogaea
Haematoxylum brasiletto
Libidibia coriaria
Acacia ligulata
Prosopis glandulosa
Ceratonia siliqua
Tamarindus indica
Polygalaceae
Epirixanthes elongata
Epirixanthes pallida
Salomonia cantoniensis
Polygala arillata
Fabids
Prunus persica
Pyrus pyrifolia
Fragaria virginiana
Fragaria chiloensis
Morus indica
Morus mongolica
Corynocarpus laevigata
Cucumis melo
Castanea mollissima
Passiflora biflora
Salix interior
Parinari campestris
Hevea brasiliensis
Rosids
Euonymus americanus
Oxalis latifolia
Bulnesia arborea
Arabidopsis thaliana
Raphanus sativus
Carica papaya
Gossypium anomalum
Theobroma cacao
Core eudicots
Azadirachta indica
Citrus sinensis
Sapindus mukorossi
Staphylea colchica
Melianthus villosus
Viviania marifolia
Erodium carvifolium
Oenothera argillicola
Oenothera parviflora
Eucalyptus aromaphloia
Sedum sarmentosum
Penthorum chinense
Heuchera sanguinea
Liquidambar formosana
Paeonia obovata
Vitis vinifera
Dillenia indica
Gunnera manicata
Buxus microphylla
Trochodendron aralioides
Tetracentron sinense
Eudicots
Macadamia integrifolia
Platanus occidentalis
Nelumbo nucifera
Ranunculus macranthus
Megaleranthis saniculifolia
Nandina domestica
Ceratophyllum demersum
Dioscorea elephantipes
Carludovica palmata
Japanolirion osense
Spirodela polyrhiza
Aponogeton distachyos
Acorus calamus
Liriodendron tulipifera
Magnolia kwangsiensis
Calycanthus floridus
Drimys granadensis
Piper cenocladum
Chloranthus spicatus
Illicium oligandrum
Nuphar advena
Amborella trichopoda
98
97
92
86
77
83
52
67
--
87
89
52
Fabales
Rosales
Fagales
Malpighiales
Oxalidales
Zygophyllales
Brassicales
Malvales
Sapindales
Crossomatales
Geraniales
Myrtales
Saxifragales
Vitales
Dilleniaceae
Gunnerales
Buxales
Trochendrales
Proteales
Ranunculales
Ceratophyllales
0.07
substitutions/site

Figure 6. A portion of angiosperm phylogeny inferred in a likelihood analysis of 82 plastid coding regions using the "GxC" partitioning scheme based on an 'ORF-only' alignment (see text and Table S3 for details); this portion of the tree shows asterid relationships. Eudicot families where mycoheterotrophy has evolved are indicated in blue. Log likelihood score of best tree: -1,506,318.267. Bootstrap support values are indicated beside branches; thick lines indicate 100% bootstrap support; '--' indicates <50% bootstrap support. The scale bar indicates estimated substitutions per site.


Phelipanche ramosa
Phelipanche purpurea
Orobanche californica
Orobanche gracillis
Orobanche crenata
Boulardia latisquama
Epifagus virginiana
Conopholis americana
Cistanche phelypaea
Cistanche deserticola
Schwalbea americana
Lindenbergia philippensis
Ajuga reptans
Premna microphylla
Utricularia gibba
Utricularia macrorhiza
Genlisea margaratae
Pinguicula ehlersiae
Andrographis paniculata
Sesamum indicum
Scrophularia takesimensis
Antirrhinum majus
Boea hygrometrica
Olea woodiana
Jasminum nudiflorum
Ehretia acuminata
Pentalinon luteum
Echites umbellatus
Oncinotus tenuiloba
Asclepias syriaca
Nerium oleander
Rhazya stricta
Catharanthus roseus
Exacum affine
Exochaenium oliganthum
Obolaria virginica
Bartonia virginica
Voyria caerulea
Voyria clavata
Coffea arabica
Cuscuta reflexa
Cuscuta exaltata
Cuscuta obtusiflora
Cuscuta gronovii
Ipomoea purpurea
Solanum lycopersicum
Solanum tuberosum
Atropa belladonna
Nicotiana sylvestris
Aucuba japonica
Artemisia frigida
Jacobaea vulgaris
Ageratina adenophora
Lactuca sativa
Centaurea diffusa
Scaevola aemula
Campanula takesimana
Trachelium caeruleum
Hanabusaya asiatica
Daucus carota
Anthriscus cerefolium
Anethum graveolens
Aralia undulata
Panax ginseng
Lonicera japonica
Ilex cornuta
Orthilila secunda
Pyrola minor
Arbutus unedo
Vaccinium macrocarpon
Rhododendrom simsii
Actinidia chinensis
Actinidia deliciosa
Camellia crapnelliana
Franklinia alatamaha
Ardisia polysticta
Lysimachia coreana
Primula poissonii
Cornus florida
Davidia involucrata
Portulacaria afra
Didierea madagascariensis
Pereskia sacharosa
Portulaca oleracea
Anredera baselloides
Mollugo verticillata
Spinacia oleracea
Silene vulgaris
Fagopyrum esculentum
Plumbago auriculata
Berberidopsis corallina
Ximenia americana
Phoradendron serotinum
Orobanchaceae
Lentibulariaceae
Apocynaceae
Gentianaceae
Rubiaceae
Convolvulaceae
Ericaceae
Actinidiaceae
Theaceae
Primulaceae
Lamiales
Boraginales
Gentianales
Solanales
Garryales
Asterales
Apiales
Dipsacales
Aquifoliales
Ericales
Cornales
Caryophyllales
Berberidopsidales
Santalales
Lamiids
Campanuliids
Asterids
to Rosids
99
97
98
99
69
86
90
51
93
--
--
68
72
85
99
59
99
0.07
substitutions/site

Bibliography

ALKATIB, S., T.T. FLEISCHMANN, L.B. SCHARFF, and R. BOCK. 2012. Evolutionary constraints on the plastid tRNA set decoding methionine and isoleucine. *Nucleic Acids Research* 40: 6713–6724.

ALTSCHUL, S.F., W. GISH, T. PENNSYLVANIA, and U. PARK. 1990. Basic Local Alignment Search Tool. *Journal of Molecular Biology* 215: 403–410.

BARBROOK, A.C., C.J. HOWE, and S. PURTON. 2006. Why are plastid genomes retained in non-photosynthetic organisms? *Trends in Plant Science* 11: 101–108.

BARRETT, C.F., and J.I. DAVIS. 2012. The plastid genome of the mycoheterotrophic *Corallorhiza striata* (Orchidaceae) is in the relatively early stages of degradation. *American Journal of Botany* 99: 1513–1523.

BARRETT, C.F., J. V. FREUDENSTEIN, J. LI, D.R. MAYFIELD-JONES, L. PEREZ, J.C. PIRES, and C. SANTOS. 2014. Investigating the path of plastid genome degradation in an early-transitional clade of heterotrophic orchids, and implications for heterotrophic angiosperms. *Molecular Biology and Evolution* 31: 3095–3112.

BEATTY, G.E., M. PHILIPP, and J. PROVAN. 2010. Unidirectional hybridization at a species' range boundary: implications for habitat tracking. *Diversity and Distributions* 16: 1–9.

BEATTY, G.E., and J. PROVAN. 2010. Refugial persistence and postglacial recolonization of North America by the cold-tolerant herbaceous plant *Orthilia secunda*. *Molecular Ecology* 19: 5009–5021.

BELLO, M.A., P.J. RUDALL, and J. A. HAWKINS. 2012. Combined phylogenetic analyses reveal interfamilial relationships and patterns of floral evolution in the eudicot order Fabales. *Cladistics* 28: 393–421.

BERG, S., K. KRAUSE, and K. KRUPINSKA. 2004. The *rbc*L genes of two Cuscuta species, C. gronovii and C. subinclusa, are transcribed by the nuclear-encoded plastid RNA polymerase (NEP). *Planta* 219: 541–6.

BIDARTONDO, M.I. 2005. The evolutionary ecology of myco-heterotrophy. *New Phytologist* 167: 335–352.

BIDARTONDO, M.I., and T.D. BRUNS. 2001. Extreme specificity in epiparasitic Monotropoideae (Ericaceae): widespread phylogenetic and geographical structure. *Molecular Ecology* 10: 2285–2295.

BIDARTONDO, M.I., B. BURGHARDT, G. GEBAUER, T.D. BRUNS, and D.J. READ. 2004. Changing partners in the dark: isotopic and molecular evidence of ectomycorrhizal liaisons between forest orchids and trees. *Proceedings of the Royal Society B: Biological Sciences* 271: 1799–1806.

BLAZIER, C.C., M.M. GUISINGER, and R.K. JANSEN. 2011. Recent loss of plastid-encoded *ndh* genes within *Erodium* (Geraniaceae). *Plant Molecular Biology* 76: 263–272.

BOHNE, A., A. WEIHE, and T. BÖRNER. 2009. Transfer RNAs inhibit Arabidopsis phage-type RNA polymerases. 63–69.

BRAUKMANN, T., M. KUZMINA, and S. STEFANOVIĆ. 2013. Plastid genome evolution across the genus *Cuscuta* (Convolvulaceae): Two clades within subgenus *Grammica* exhibit extensive gene loss. *Journal of Experimental Botany* 64: 977–989.

BRAUKMANN, T., and S. STEFANOVIĆ. 2012. Plastid genome evolution in mycoheterotrophic Ericaceae. *Plant Molecular Biology* 79: 5–20.

BRAUKMANN, T.W.A., M. KUZMINA, and S. STEFANOVIĆ. 2009. Loss of all plastid *ndh* genes in Gnetales and conifers: extent and evolutionary significance for the seed plant phylogeny.

Current Genetics 55: 323–337.

BUNGARD, R. A. 2004. Photosynthetic evolution in parasitic plants: Insight from the chloroplast genome. *BioEssays* 26: 235–247.

CAI, Z., M. GUISINGER, H.G. KIM, E. RUCK, J.C. BLAZIER, V. MCMURTRY, J. V. KUEHL, ET AL. 2008. Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. *Journal of Molecular Evolution* 67: 696–704.

CAMERON, D.D., and J.F. BOLIN. 2010. Isotopic evidence of partial mycoheterotrophy in the Gentianaceae: *Bartonia virginica* and *Obolaria virginica* as case studies. *American Journal of Botany* 97: 1272–1277.

COSNER, M.E., R.K. JANSEN, J.D. PALMER, and S.R. DOWNIE. 1997. The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. *Current Genetics* 31: 419–429.

CRONN, R., A. LISTON, M. PARKS, D.S. GERNANDT, R. SHEN, and T. MOCKLER. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research* 36: e122.

CRONQUIST, A. 1988. The Evolution and Classification of Flowering Plants. 2nd ed. The New York Botanical Garden, Bronx, New York.

CUMMINGS, M.P., and N.A. WELSCHMEYER. 1998. Pigment composition of putatively achlorophyllous angiosperms. *Plant Systematics and Evolution* 210: 105–111.

DARLING, A.C.E., B. MAU, F.R. BLATTNER, and N.T. PERNA. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* 14: 1394–403.

DELANNOY, E., S. FUJII, C. COLAS DES FRANCS-SMALL, M. BRUNDRETT, and I. SMALL. 2011. Rampant gene loss in the underground orchid *Rhizanthella gardneri* highlights evolutionary constraints on plastid genomes. *Molecular Biology and Evolution* 28: 2077–2086.

DELAVAULT, P., V. SAKANYAN, and P. THALOUARN. 1995. Divergent evolution of two plastid genes, *rbc*L and *atp*B, in a non-photosynthetic parasitic plant. *Plant Molecular Biology* 29: 1071–9.

DOWNIE, S.R., D.S. KATZ-DOWNIE, K.H. WOLFE, P.J. CALIE, and J.D. PALMER. 1994. Internal plasticity and multiple gene loss during angiosperm evolution. *Current Genetics* 25: 367–378.

DOWNIE, S.R., and J.D. PALMER. 1992. Use of chloroplast DNA rearrangments in reconstructing plant phylogeny. *In* P. S. Soltis, D. E. Soltis, and J. J. Doyle [eds.], Molecular Systematics of Plants, 14–35. Springer US, New York.

DOYLE, J., and J. DOYLE. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *The Phytochemical Bulletin* 19: 11–15.

DRESCHER, A., R. STEPHANIE, T. CALSA, H. CARRER, and R. BOCK. 2000. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. *Plant Journal* 22: 97–104.

ERNES, M.J., and H.E. NEUHAUS. 2005. Metabolism and transport in non-photosynthetic plastids. *Journal of Experimental Botany* 48: 1995–2005.

FAJARDO, D., D. SENALIK, M. AMES, H. ZHU, S. A. STEFFAN, R. HARBUT, J. POLASHOCK, ET AL. 2013. Complete plastid genome sequence of *Vaccinium macrocarpon*: Structure, gene content, and rearrangements revealed by next generation sequencing. *Tree Genetics and Genomes* 9: 489–498.

FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* 27: 401–410.

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

FREYER, R., M.C. KIEFER-MEYER, and H. KÖSSEL. 1997. Occurrence of plastid RNA editing in all major lineages of land plants. *Proceedings of the National Academy of Sciences of the United States of America* 94: 6285–6290.

FUNK, H.T., S. BERG, K. KRUPINSKA, U.G. MAIER, and K. KRAUSE. 2007. Complete DNA sequences of the plastid genomes of two parasitic flowering plant species, *Cuscuta reflexa* and *Cuscuta gronovii*. *BMC Plant Biology* 7: 45.

GEBAUER, G., and M. MEYER. 2003. 15N and 13C natural abundance of autotrophic and myco-heterotrophic orchids provides insight into nitrogen and carbon gain from fungal association. *New Phytologist* 160: 209–223.

GRAHAM, S.W., and R.G. OLMSTEAD. 2000. Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *American Journal of Botany* 87: 1712–1730.

GRAHAM, S.W., P. A. REEVES, A.C.E. BURNS, and R.G. OLMSTEAD. 2000. Microstructural changes in noncoding chloroplast DNA: interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference. *International Journal of Plant Sciences* 161: S83–S96.

GRAY, B.N., B. A AHNER, and M.R. HANSON. 2009. Extensive homologous recombination between introduced and native regulatory plastid DNA elements in transplastomic plants. *Transgenic Research* 18: 559–572.

HABERLE, R.C., H.M. FOURCADE, J.L. BOORE, and R.K. JANSEN. 2008. Extensive rearrangements

in the chloroplast genome of Trachelium caeruleum are associated with repeats and tRNA genes. *Journal of Molecular Evolution* 66: 350–361.

HAJDUKIEWICZ, P.T., L.A. ALLISON, and P. MALIGA. 1997. The two RNA polymerases encoded by the nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids. *The EMBO Journal* 16: 4041–8.

HANAOKA, M., K. KANAMARU, M. FUJIWARA, H. TAKAHASHI, and K. TANAKA. 2005. Glutamyl-tRNA mediates a switch in RNA polymerase use during chloroplast biogenesis. *EMBO Reports* 6: 545–550.

HARRIS, E.H., J.E. BOYNTON, and N.W. GILLHAM. 1994. Chloroplast ribosomes and protein synthesis. *Microbiological Reviews* 58: 700–754.

HENDY, M.D., and D. PENNY. 1989. A framework for the quantitative study of evolutionary trees. *Systematic Zoology* 38: 297–309.

HOFFMANN, S., C. OTTO, S. KURTZ, C.M. SHARMA, P. KHAITOVICH, J. VOGEL, P.F. STADLER, and J. HACKERMÜLLER. 2009. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Computational Biology* 5: e1000502.

HOWE, C.J., and S. PURTON. 2007. The little genome of apicomplexan plastids: its raison d'etre and a possible explanation for the "delayed death" phenomenon. *Protist* 158: 121–133.

HYNSON, N. A, and T.D. BRUNS. 2009. Evidence of a myco-heterotroph in the plant family Ericaceae that lacks mycorrhizal specificity. *Proceedings of the Royal Society B: Biological Sciences* 276: 4053–4059.

HYNSON, N. A, K. PREISS, G. GEBAUER, and T.D. BRUNS. 2009. Isotopic evidence of full and partial myco-heterotrophy in the plant tribe Pyroleae (Ericaceae). *The New phytologist* 182: 719–26.

ILES, W.J.D., S.Y. SMITH, and S.W. GRAHAM. 2013. A well-supported phylogenetic framework for the monocot order Alismatales reveals multiple losses of the plastid NADH dehydrogenase complex and a strong long branch effect. *In* P. Wilkin, and S. J. Mayo [eds.], Early Events in Monocot Evolution, 1–19. Cambridge University Press.

JAHN, D., E. VERKAMP, and D. SÖLL. 1992. Glutamyl-transfer RNA: a precursor of heme and chlorophyll biosynthesis. *Trends in Biochemical Sciences* 17: 215–218.

JANOUŠKOVEC, J., D. V. TIKHONENKOV, F. BURKI, A.T. HOWE, M. KOLÍSKO, A.P. MYLNIKOV, and P.J. KEELING. 2015. Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. *Proceedings of the National Academy of Sciences* 112: 10200–10207.

JANSEN, R.K., Z. CAI, L. A RAUBESON, H. DANIELL, C.W. DEPAMPHILIS, J. LEEBENS-MACK, K.F. MÜLLER, ET AL. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences of the United States of America* 104: 19369–19374.

JANSEN, R.K., C. SASKI, S.-B. LEE, A.K. HANSEN, and H. DANIELL. 2011. Complete plastid genome sequences of three Rosids (Castanea, Prunus, Theobroma): evidence for at least two independent transfers of rpl22 to the nucleus. *Molecular Biology and Evolution* 28: 835–47.

KATOH, K., K. MISAWA, K. KUMA, and T. MIYATA. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30: 3059–3066.

KAWAKAMI, K., Y. UMENA, M. IWAI, Y. KAWABATA, M. IKEUCHI, N. KAMIYA, and J.-R. SHEN. 2011. Roles of PsbI and PsbM in photosystem II dimer formation and stability studied by deletion mutagenesis and X-ray crystallography. *Biochimica et Biophysica Acta (BBA) -*

Bioenergetics 1807: 319–325.

KIKUCHI, S., J. BÉDARD, M. HIRANO, Y. HIRABAYASHI, M. OISHI, M. IMAI, M. TAKASE, ET AL. 2013. Uncovering the protein translocon at the chloroplast inner envelope membrane. *Science* 339: 571–4.

KIM, H.T., J.S. KIM, M.J. MOORE, K.M. NEUBIG, N.H. WILLIAMS, W.M. WHITTEN, and J.-H. KIM. 2015. Seven new complete plastome sequences reveal rampant independent loss of the ndh gene family across orchids and associated instability of the inverted repeat/small single-copy region boundaries. *Plos ONE* 10: e0142215.

KIM, J., A. RUDELLA, V. RAMIREZ RODRIGUEZ, B. ZYBAILOV, P.D.B. OLINARES, and K.J. VAN WIJK. 2009. Subunits of the plastid ClpPR protease complex have differential contributions to embryogenesis, plastid biogenesis, and plant development in Arabidopsis. *The Plant Cell* 21: 1669–1692.

KISSLING, J. 2012. Taxonomy of *Exochaenium* and *Lagenias*: two resurrected genera of tribe Exaceae (Gentianaceae). *Systematic Botany* 37: 238–253.

KNAUF, U., and W. HACHTEL. 2002. The genes encoding subunits of ATP synthase are conserved in the reduced plastid genome of the heterotrophic alga *Prototheca wickerhamii*. *Molecular Genetics and Genomics* 267: 492–7.

KRAUSE, K. 2008. From chloroplasts to "cryptic" plastids: evolution of plastid genomes in parasitic plants. *Current Genetics* 54: 111–121.

KRAUSE, K. 2012. Plastid genomes of parasitic plants: A trail of reductions and losses. *In* Bullerwell C.E. [ed.], Organelle Genetics, 79–103. Springer-Verlag, Berlin Heidelberg.

KRON, K.A., W.S. JUDD, P.F. STEVENS, D.M. CRAYN, and A.A. ANDERBERG. 2002. Phylogenetic classification of Ericaceae: molecular and morpholgical evidence. *The Botanical Review* 68:

335–423.

KUGITA, M., Y. YAMAMOTO, T. FUJIKAWA, T. MATSUMOTO, and K. YOSHINAGA. 2003. RNA editing in hornwort chloroplasts makes more than half the genes functional. *Nucleic Acids Research* 31: 2417–2423.

KURODA, H., and P. MALIGA. 2003. The plastid *clp*P1 protease gene is essential for plant development. *Nature* 425: 86–89.

LAM, V.K.Y., M. SOTO GOMEZ, and S.W. GRAHAM. 2015. The highly reduced plastome of mycoheterotrophic *Sciaphila* (Triuridaceae) is colinear with its green relatives and is under strong purifying selection. *Genome Biology and Evolution* 7: 2220–2236.

LANFEAR, R., B. CALCOTT, S.Y.W. HO, and S. GUINDON. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution* 29: 1695–1701.

LANFEAR, R., B. CALCOTT, D. KAINER, C. MAYER, and A. STAMATAKIS. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evolutionary Biology* 14: 82.

LEAKE, J.R., and D.D. CAMERON. 2010. Physiological ecology of mycoheterotrophy. *New Phytologist* 185: 601–605.

LEEBENS-MACK, J., and C. DEPAMPHILIS. 2002. Power analysis of tests for loss of selective constraint in cave crayfish and nonphotosynthetic plant lineages. *Molecular biology and evolution* 19: 1292–1302.

LIERE, K., A. WEIHE, and T. BÖRNER. 2011. The transcription machineries of plant mitochondria and chloroplasts: composition, function, and regulation. *Journal of Plant Physiology* 168: 1345–1360.

LOGACHEVA, M.D., M.I. SCHELKUNOV, M.S. NURALIEV, T.H. SAMIGULLIN, and A. A. PENIN.

2014. The plastid genome of mycoheterotrophic monocot *Petrosavia stellaris* exhibits both gene losses and multiple rearrangements. *Genome Biology and Evolution* 6: 238–246.

LOGACHEVA, M.D., M.I. SCHELKUNOV, and A. A. PENIN. 2011. Sequencing and analysis of plastid genome in mycoheterotrophic orchid *Neottia nidus-avis*. *Genome Biology and Evolution* 3: 1296–1303.

LOHSE, M., O. DRECHSEL, S. KAHLAU, and R. BOCK. 2013. OrganellarGenomeDRAW--a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research* 41: 1–7.

MACHADO, M. A., and K. ZETSCHE. 1990. A structural, functional and molecular analysis of plastids of the holoparasites *Cuscuta reflexa* and *Cuscuta europaea*. *Planta* 181: 91–96.

MADDISON, W.P., and D.R. MADDISON. 2014. Mesquite: a modular system for evolutionary analyses. Version 3.03+. Available at: http://mesquiteproject.org.

MADDISON, W.P., and D.R. MADDISON. 2015. Mesquite: a modular system for evolutionary analyses. Version 3.04. Available at: http://mesquiteproject.org.

MARÉCHAL, A., and N. BRISSON. 2010. Recombination and the maintenance of plant organelle genome stability. *New Phytologist* 186: 299–317.

MARTÍN, M., and B. SABATER. 2010. Plastid *ndh* genes in plant evolution. *Plant Physiology and Biochemistry* 48: 636–645.

MARTÍNEZ-ALBEROLA, F., E.M. DEL CAMPO, D. LÁZARO-GIMENO, S. MEZQUITA-CLARAMONTE, A. MOLINS, I. MATEU-ANDRÉS, J. PEDROLA-MONFORT, ET AL. 2013. Balanced gene losses, duplications and intensive rearrangements led to an unusual regularly sized genome in Arbutus unedo chloroplasts. *PLoS ONE*. http://doi:10.1371/journal.pone.0079685

MCNEAL, J.R., K. ARUMUGUNATHAN, J. V KUEHL, J.L. BOORE, and C.W. DEPAMPHILIS. 2007.

Systematics and plastid genome evolution of the cryptically photosynthetic parasitic plant genus *Cuscuta* (Convolvulaceae). *BMC Biology* 5: 55.

MCNEAL, J.R., J. V. KUEHL, J.L. BOORE, J. LEEBENS-MACK, and C.W. DEPAMPHILIS. 2009. Parallel loss of plastid introns and their maturase in the genus *Cuscuta*. *PLoS ONE* 4: 1–8.

MENNES, C.B., V.K.Y. LAM, P.J. RUDALL, S.P. LYON, S.W. GRAHAM, E.F. SMETS, and V.S.F.T. MERCKX. 2015. Ancient Gondwana break-up explains the distribution of the mycoheterotrophic family Corsiaceae (Liliales). *Journal of Biogeography* 42: 1123–1136.

MENNES, C.B., M.S. MOERLAND, M. RATH, E.F. SMETS, and V.S.F.T. MERCKX. 2015. Evolution of mycoheterotrophy in Polygalaceae: the case of *Epirixanthes*. *American Journal of Botany* 102: 598–608.

MERCKX, V., F.T. BAKKER, S. HUYSMANS, and E. SMETS. 2009. Bias and conflict in phylogenetic inference of myco-heterotrophic plants: a case study in Thismiaceae. *Cladistics* 25: 64–77.

MERCKX, V., and J. V FREUDENSTEIN. 2010. Evolution of mycoheterotrophy in plants: a phylogenetic perspective. *New Phytologist* 185: 605–609.

MERCKX, V.S.F.T. 2013. Mycoheterotrophy: an introduction. *In* V. S. F. T. Merckx [ed.], Mycoheterotrophy: The Biology of Plants Living on Fungi, 1–17. Springer Science and Business Media, New York.

MERCKX, V.S.F.T., J. V. FREUDENSTEIN, B. CRANDALL-STOTLER, M.J. CHRISTENHUSZ, R.E. STOTLER, N.J. WICKETT, H. MAAS-VAN DE KAMER, and P.J.M. MAAS. 2013. Taxonomy and Classification. *In* V. S. F. T. Merckx [ed.], Mycoheterotrophy:The Biology of Plants Living on Fungi, 19–101. Springer Science and Business Media, New York.

MERCKX, V.S.F.T., J. KISSLING, H. HENTRICH, S.B. JANSSENS, C.B. MENNES, C.D. SPECHT, and E.F. SMETS. 2013. Phylogenetic relationships of the mycoheterotrophic genus *Voyria* and

the implications for the biogeographic history of Gentianaceae. *American Journal of Botany* 100: 712–721.

MILLEN, R.S., R.G. OLMSTEAD, K.L. ADAMS, J.D. PALMER, N.T. LAO, L. HEGGIE, T. A KAVANAGH, ET AL. 2001. Many parallel losses of infA from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. *The Plant Cell* 13: 645–658.

MILLER, M. A, W. PFEIFFER, and T. SCHWARTZ. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *2010 Gateway Computing Environments Workshop, GCE 2010:*1–8.

MILLIGAN, B.G., J.N. HAMPTON, and J.D. PALMER. 1989. Dispersed repeats and structural reorganization in subclover chloroplast DNA. *Molecular Biology and Evolution* 6: 355–368.

MOLINA, J., K.M. HAZZOURI, D. NICKRENT, M. GEISLER, R.S. MEYER, M.M. PENTONY, J.M. FLOWERS, ET AL. 2014. Possible loss of the chloroplast genome in the parasitic flowering plant *Rafflesia lagascae* (Rafflesiaceae). *Molecular Biology and Evolution* 31: 793–803.

NAKAI, M. 2015. YCF1: A green TIC: response to the de Vries et al. commentary. *The Plant cell* 27: 1834–1838.

OHLROGGE, J., and J. BROWSE. 1995. Lipid biosynthesis. *Plant Cell* 7: 957–970.

OSMOND, C.B., T. AKAZAWA, and H. BEEVERS. 1975. Localization and properties of ribulose diphosphate carboxylase from castor bean endosperm. *Plant Physiology* 55: 226–230.

PALMER, J.D. 1985. Comparative organization of chloroplast genomes. *Annual Review of Genetics* 19: 325–354.

PALMER, J.D., B. OSORIO, J. ALDRICH, and W.F. THOMPSON. 1987. Chloroplast DNA evolution among legumes: loss of a large inverted repeat occurred prior to other sequence

rearrangements. *Current Genetics* 11: 275–286.

PALMER, J.D., and D.B. STEIN. 1986. Conservation of chloroplast genome structure among vascular plants. *Current Genetics* 10: 823–833.

PELTIER, G., and L. COURNAC. 2002. Chlororespiration. *Annual Review of Plant Biology* 53: 523–550.

PEREDO, E.L., U.M. KING, and D.H. LES. 2013. The plastid genome of Najas flexilis: adaptation to submersed environments is accompanied by the complete loss of the NDH complex in an aquatic angiosperm. *PloS ONE* 8: e68591.

PETERSEN, G., A. CUENCA, and O. SEBERG. 2015. Plastome evolution in hemiparasitic mistletoes. *Genome Biology and Evolution* 7: 2520–2532.

RANDLE, C.P., and A.D. WOLFE. 2005. The evolution and expression of *rbc*L in holoparasitic sister-genera *Harveya* and *Hyobanche* (Orobanchaceae). *American Journal of Botany* 92: 1575–1585.

ROGALSKI, M., D. KARCHER, and R. BOCK. 2008. Superwobbling facilitates translation with reduced tRNA sets. *Nature, Structural & Molecular Biology* 15: 192–198.

ROSS, T.G., C.F. BARRETT, M. SOTO, V.K.Y. LAM, C.L. HENRIQUEZ, D.H. LES, J.I. DAVIS, ET AL. 2015. Plastid phylogenomics and molecular evolution of Alismatales. *Cladistics*1–19.

ROUSSEAU-GUEUTIN, M., X. HUANG, E. HIGGINSON, M. AYLIFFE, A. DAY, and J.N. TIMMIS. 2013. Potential functional replacement of the plastidic acetyl-CoA carboxylase subunit (*acc*D) gene by recent transfers to the nucleus in some angiosperm lineages. *Plant Physiology* 161: 1918–29.

RUHFEL, B.R., M. A GITZENDANNER, P.S. SOLTIS, D.E. SOLTIS, and J.G. BURLEIGH. 2014. From algae to angiosperms - inferring the phylogeny of green plants (Viridiplantae) from 360

plastid genomes. *BMC Evolutionary Biology* 14: 23.

RUHLMAN, T. A, W.-J. CHANG, J.J. CHEN, Y.-T. HUANG, M.-T. CHAN, J. ZHANG, D.-C. LIAO, ET AL. 2015. NDH expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. *BMC Plant Biology* 15: 1–9.

RUIZ-NIETO, J.E., C.L. AGUIRRE-MANCILLA, J.A. ACOSTA-GALLEGOS, J.C. RAYA-PÉREZ, E. PIEDRA-IBARRA, J. VÁZQUEZ-MEDRANO, and V. MONTERO-TAVERA. 2015. Photosynthesis and chloroplast genes are involved in water-use efficiency in common bean. *Plant Physiology and Biochemistry* 86: 166–173.

SAARELA, J.M., and S.W. GRAHAM. 2010. Inference of phylogenetic relationships among the subfamilies of grasses (Poaceae: Poales) using meso-scale exemplar-based sampling of the plastid genome. *Botany* 88: 65–84.

SABIR, J., E. SCHWARZ, N. ELLISON, J. ZHANG, N. A BAESHEN, M. MUTWAKIL, R. JANSEN, and T. RUHLMAN. 2014. Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. *Plant Biotechnology Journal* 12: 743–54.

SAINT-MARCOUX, D., F.A. WOLLMAN, and C. DE VITRY. 2009. Biogenesis of cytochrome b6 in photosynthetic membranes. *Journal of Cell Biology* 185: 1195–1207.

SANDERSON, M.J., D. COPETTI, A. BURQUEZ, E. BUSTAMANTE, J.L.M. CHARBONEAU, L.E. EGUIARTE, S. KUMAR, ET AL. 2015. Exceptional reduction of the plastid genome of saguaro cactus (*Carnegiea gigantea*): loss of the *ndh* gene suite and inverted repeat. *American Journal of Botany* 102: 1115–1127.

SASAKI, Y., and Y. NAGANO. 2004. Plant acetyl-CoA carboxylase: structure, biosynthesis, regulation, and gene manipulation for plant breeding. *Bioscience, Biotechnology, and Biochemistry* 68: 1175–1184.

SATO, S. 2011. The apicomplexan plastid and its evolution. *Cellular and Molecular Life Sciences* 68: 1285–1296.

SCHELKUNOV, M., V. SHTRATNIKOVA, M. NURALIEV, M. -A. SELOSSE, A. PENIN, and M. LOGACHEVA. 2015. Exploring the limits for reduction of plastid genomes: a case study of the mycoheterotrophic orchids *Epipogium aphyllum* and *Epipogium roseum*. *Genome Biology and Evolution* evv019: 1–29.

SCHWARZ, E.N., T.A. RUHLMAN, J.S.M. SABIR, N.H. HAJRAH, N.S. ALHARBI, A.L. AL-MALKI, C.D. BAILEY, and R.K. JANSEN. 2015. Plastid genome sequences of legumes reveal parallel inversions and multiple losses of rps16 in papilionoids. *Journal of Systematics and Evolution* 53: 458–468.

SCHWENDER, J., F. GOFFMAN, J.B. OHLROGGE, and Y. SHACHAR-HILL. 2004. Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds. *Nature* 432: 779–782.

SHIKANAI, T. 2015. Chloroplast NDH: A different enzyme with a structure similar to that of respiratory NADH dehydrogenase. *Biochimica et Biophysica Acta*. http://dx.doi.org/10.1016/j.bbabio.2015.10.013

SILVESTRO, D., and I. MICHALAK. 2012. raxmlGUI: a graphical front-end for RAxML. *Organisms Diversity & Evolution* 12: 335–337.

STAMATAKIS, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22: 2688–2690.

STAMATAKIS, A., P. HOOVER, and J. ROUGEMONT. 2008. A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology* 57: 758–771.

STANNE, T.M., L.L.E. SJÖGREN, S. KOUSSEVITZKY, and A.K. CLARKE. 2009. Identification of

new protein substrates for the chloroplast ATP-dependent Clp protease supports its constitutive role in Arabidopsis. *The Biochemical Journal* 417: 257–68.

STRAUB, S.C.K., M. FISHBEIN, T. LIVSHULTZ, Z. FOSTER, M. PARKS, K. WEITEMIER, R.C. CRONN, and A. LISTON. 2011. Building a model: developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics* 12: 211.

STRUWE, L. 2014. Classification and evolution in the family Gentianaceae. *In* J. J. Rybcynski, M. R. Davey, and A. Mikula [eds.], The Gentianaceae - Vol. 1: Characterization and Ecology, 1–12. Springer-Verlag, Berlin Heidelberg.

SUGIURA, M. 2014. Plastid mRNA translation. *In* P. Maliga [ed.], Chloroplast Biotechnology: Methods and Protocols, Methods in Molecular Biology, 73–91. Springer Science and Business Media, New York.

TEDERSOO, L., P. PELLET, U. KÕLJALG, and M. A. SELOSSE. 2007. Parallel evolutionary paths to mycoheterotrophy in understorey Ericaceae and Orchidaceae: ecological evidence for mixotrophy in Pyroleae. *Oecologia* 151: 206–217.

THIERRY, A., C. BOUCHIER, B. DUJON, and G.-F. RICHARD. 2008. Megasatellites: a peculiar class of giant minisatellites in genes involved in cell adhesion and pathogenicity in *Candida glabrata*. *Nucleic Acids Research* 36: 5970–82.

THIERS, B. 2015. Index Herbariorium: A global directory of public herbaria and associated staff. *New York Botanical Garden's Virtual Herbarium*. Available at: http://sweetgum.nybg.org/science/ih/.

THOMPSON, J.D., D.G. HIGGINS, and T.J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific

gap penalties and weight matrix choice. *Nucleic Acids Research* 22: 4673–4680.

TOLBERT, N.E. 1997. The C2 oxidative photosynthetic carbon cycle. *Annual Review of Plant Physiology and Plant Molecular Biology* 48: 1–25.

UMATE, P., S. SCHWENKERT, I. KARBAT, C. DAL BOSCO, L. MÌCOHX , S. VOLZ, H. ZER, ET AL. 2007. Deletion of PsbM in tobacco alters the QB site properties and the electron flow within photosystem II. *Journal of Biological Chemistry* 282: 9758–9767.

DE VRIES, J., F.L. SOUSA, B. BÖLTER, J. SOLL, and S.B. GOULD. 2015. YCF1: A Green TIC? *The Plant Cell* 27: 1827–1833.

WALKER, J.E. 2013. The ATP synthase: the understood, the uncertain and the unknown. *Biochemical Society Transactions* 41: 1–16.

WICKE, S., K.F. MÜLLER, C.W. DE PAMPHILIS, D. QUANDT, N.J. WICKETT, Y. ZHANG, S.S. RENNER, and G.M. SCHNEEWEISS. 2013. Mechanisms of functional and physical genome reduction in photosynthetic and nonphotosynthetic parasitic plants of the broomrape family. *The Plant Cell* 25: 3711–25.

WICKE, S., B. SCHÄFERHOFF, C.W. DEPAMPHILIS, and K.F. MÜLLER. 2014. Disproportional plastome-wide increase of substitution rates and relaxed purifying selection in genes of carnivorous Lentibulariaceae. *Molecular Biology and Evolution* 31: 529–545.

WICKE, S., G.M. SCHNEEWEISS, C.W. DEPAMPHILIS, K.F. MÜLLER, and D. QUANDT. 2011. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Molecular Biology* 76: 273–297.

WICKETT, N.J., Y. ZHANG, S.K. HANSEN, J.M. ROPER, J. V. KUEHL, S. A. PLOCK, P.G. WOLF, ET AL. 2008. Functional gene losses occur with minimal size reduction in the plastid genome of the parasitic liverwort *Aneura mirabilis*. *Molecular Biology and Evolution* 25: 393–401.

WOLFE, K.H. 1994. Similarity between putative ATP-binding sites in land plant plastid ORF2280 proteins and the FtsH/CDC48 family of ATPases. *Current Genetics* 25: 379–383.

WOLFE, K.H., C.W. MORDEN, and J.D. PALMER. 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proceedings of the National Academy of Sciences of the United States of America* 89: 10648–10652.

WU, C.S., Y.T. LAI, C.P. LIN, Y.N. WANG, and S.M. CHAW. 2009. Evolution of reduced and compact chloroplast genomes (cpDNAs) in gnetophytes: selection toward a lower-cost strategy. *Molecular Phylogenetics and Evolution* 52: 115–124.

WU, F.-H., M.-T. CHAN, D.-C. LIAO, C.-T. HSU, Y.-W. LEE, H. DANIELL, M.R. DUVALL, and C.-S. LIN. 2010. Complete chloroplast genome of *Oncidium* Gower Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. *BMC Plant Biology* 10: 68.

WYMAN, S.K., R.K. JANSEN, and J.L. BOORE. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20: 3252–3255.

XIE, Z., and S. MERCHANT. 1996. The plastid-encoded *ccs*A gene is required for heme attachment to chloroplast c-type cytochromes. *Journal of Biological Chemistry* 271: 4632–4639.

XU, Q., D. HOPPE, V.P. CHITNIS, W.R. ODOM, J. A GUIKEMA, and P.R. CHITNIS. 1995. Mutational analysis of photosystem I polypeptides in the cyanobacterium Synechocystis sp. PCC 6803. Targeted inactivation of psaI reveals the function of psaI in the structural organization of psaL. *The Journal of Biological Chemistry* 270: 16243–50.

YAGI, Y., and T. SHIINA. 2014. Recent advances in the study of chloroplast gene expression and its evolution. *Frontiers in Plant Science* 5: 61.

YANG, J.-B., M. TANG, H.-T. LI, Z.-R. ZHANG, and D.-Z. LI. 2013. Complete chloroplast genome

of the genus Cymbidium: lights into the species identification, phylogenetic implications and population genetic analyses. *BMC Evolutionary Biology* 13: 84.

ZGURSKI, J.M., H.S. RAI, Q.M. FAI, D.J. BOGLER, J. FRANCISCO-ORTEGA, and S.W. GRAHAM. 2008. How well do we understand the overall backbone of cycad phylogeny? New insights from a large, multigene plastid data set. *Molecular Phylogenetics and Evolution* 47: 1232–1237.

ZHELYAZKOVA, P., C.M. SHARMA, K.U. FORSTNER, K. LIERE, J. VOGEL, and T. BORNER. 2012. The primary transcriptome of barley chloroplasts: numerous noncoding RNAs and the dominating role of the plastid-encoded RNA polymerase. *The Plant Cell* 24: 123–136.

ZHU, A., W. GUO, S. GUPTA, W. FAN, and J.P. MOWER. 2015. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytologist*.

ZIMMER, K., N. A. HYNSON, G. GEBAUER, E.B. ALLEN, M.F. ALLEN, and D.J. READ. 2007. Wide geographical and ecological distribution of nitrogen and carbon gains from fungi in pyroloids and monotropoids (Ericaceae) and in orchids. *New Phytologist* 175: 166–175.

ZIMMER, K., C. MEYER, and G. GEBAUER. 2008. The ectomycorrhizal specialist orchid Corallorhiza trifida is a partial myco-heterotroph. *New Phytologist* 178: 395–400.

ZOSCHKE, R., M. NAKAMURA, K. LIERE, M. SUGIURA, T. BÖRNER, and C. SCHMITZ-LINNEWEBER. 2010. An organellar maturase associates with multiple group II introns. *Proceedings of the National Academy of Sciences of the United States of America* 107: 3245–3250.

Appendices

Supplementary Table S1. Accession information for publically available plastomes included in the angiosperm matrix (for all others, see Ruhfel et al., 2014).

Order	Family	Species	Accession #
Apiales	Araliaceae	*Aralia undulata* Hand.-Mazz	NC_022810
Asterales	Asteraceae	*Artemisia frigida* Eichw.	NC_020607
Asterales	Asteraceae	*Centaurea diffusa* Lam.	NC_024286
Asterales	Campanulaceae	*Campanula takesimana* Nakai	NC_026203
Asterales	Campanulaceae	*Hanabusaya asiatica* (Nakai) Nakai	NC_024732
Brassicales	Brassicaceae	*Raphanus sativus* L.	NC_024469
Caryophyllales	Polygonaceae	*Fagopyrum esculentum* Moench	NC_010776
Ericales	Actinidiaceae	*Actinidia chinensis* Planch.	NC_026690
Ericales	Actinidiaceae	*Actinidia deliciosa* (A.Chev.) C.F. Lian & A.R.Ferguson	NC_026691
Ericales	Ericaceae	*Vaccinium macrocarpon* Aiton	NC_019616
Ericales	Primulaceae	*Ardisia polysticta* Miq.	NC_021121
Ericales	Primulaceae	*Lysimachia coreana* Nakai	NC_026197
Ericales	Primulaceae	*Primula poissonii* Franch.	NC_024543
Ericales	Theaceae	*Camellia crapnelliana* Tutcher	NC_024541
Fabales	Fabaceae	*Acacia ligulata* Benth.	NC_026134
Fabales	Fabaceae	*Apios americana* Medik.	NC_025909
Fabales	Fabaceae	*Arachis hypogaea* L.	NC_026676
Fabales	Fabaceae	*Ceratonia siliqua* L.	NC_026678

Order	Family	Species	Accession #
Fabales	Fabaceae	*Haematoxylum brasiletto* H.Karst.	NC_026679
Fabales	Fabaceae	*Libidibia coriaria* (Jacq.) Schltdl.	NC_026677
Fabales	Fabaceae	*Lupinus albus* L.	NC_026681
Fabales	Fabaceae	*Prosopis glandulosa* Torr.	NC_026683
Fabales	Fabaceae	*Tamarindus indica* L.	NC_026685
Gentianales	Apocynaceae	*Asclepias syriaca* L.	NC_022432
Gentianales	Apocynaceae	Catharanthus roseus (L.) G.Don cultivar Pacifica Punch Halo	NC_021423
Gentianales	Apocynaceae	*Echites umbellatus* Jacq.	NC_025655
Gentianales	Apocynaceae	*Oncinotis tenuiloba* Stapf	NC_025657
Gentianales	Apocynaceae	*Pentalinon luteum* (L.) B.F.Hansen & Wunderlin	NC_025658
Gentianales	Apocynaceae	*Rhazya stricta* Decne.	NC_024292
Geraniales	Melianthaceae	*Melianthus villosus* Bolus	NC_023256
Geraniales	Vivianiaceae	*Viviania marifolia* Cav.	NC_023259
Lamiales	Acanthaceae	*Andrographis paniculata* (Burm.f.) Nees	NC_022451
Lamiales	Lamiaceae	*Ajuga reptans* L.	NC_023102
Lamiales	Lamiaceae	*Premna microphylla* Turcz.	NC_026291
Lamiales	Lentibulariaceae	*Genlisea margaretae* Hutch.	NC_025652
Lamiales	Lentibulariaceae	*Pinguicula ehlersiae* Speta & F. Fuchs	NC_023463

Order	Family	Species	Accession #
Lamiales	Lentibulariaceae	*Utricularia gibba* L.	NC_021449
Lamiales	Lentibulariaceae	*Utricularia macrorhiza* Leconte	NC_025653
Lamiales	Orobanchaceae	*Cistanche deserticola* Y.C.Ma	NC_021111
Lamiales	Orobanchaceae	*Cistanche phelypaea* (L.) Cout.	NC_025642
Lamiales	Orobanchaceae	*Epifagus virginiana* (L.) W.P.C. Barton	NC_001568
Lamiales	Orobanchaceae	*Lindenbergia philippensis* (Cham. & Schltdl.) Benth.	NC_022859
Lamiales	Orobanchaceae	*Orobanche californica* Cham & Schltdl.	NC_025651
Lamiales	Orobanchaceae	*Orobanche crenata* Forssk.	NC_024845
Lamiales	Orobanchaceae	*Orobanche gracilis* Sm.	NC_023464
Lamiales	Orobanchaceae	*Phelipanche purpurea* (Jacq.) Sojak	NC_023132
Lamiales	Orobanchaceae	*Phelipanche ramosa* (L.) Pomel	NC_023465
Lamiales	Orobanchaceae	*Schwalbea americana* L.	NC_023115
Lamiales	Scrophulariaceae	*Boulardia latisquama* F.W.Schultz	NC_025641
Lamiales	Scrophulariaceae	*Scrophularia takesimensis* Nakai	NC_026202
Malphigiales	Chrysobalanaceae	*Parinari campestris* Aubl.	NC_024067
Malphigiales	Salicaceae	*Salix interior* Rowlee	NC_024681
Malvales	Malvaceae	*Gossypium anomalum* Wawra & Peyr.	NC_023213

Order	Family	Species	Accession #
Myrtales	Myrtaceae	*Eucalyptus aromaphloia* Pryor & J.H.Willis	NC_022396
Pandanales	Cyclanthaceae	*Carludovica palmata* Ruiz & Pav.	NC_026786
Proteales	Proteaceae	*Macadamia integrifolia* Maiden & Betche	NC_025288
Rosales	Moraceae	*Morus mongolica* (Bureau) C.K.Schneid.	NC_025772
Rosales	Rosaceae	*Fragaria chiloensis* Auct.	NC_019601
Rosales	Rosaceae	*Fragaria virginiana* Mill.	NC_019602
Sapindales	Meliaceae	*Azadirachta indica* A.Juss.	NC_023792
Sapindales	Sapindaceae	*Sapindus mukorossi* Gaertn.	NC_025554
Saxifragales	Altingiaceae	*Liquidambar formosana* Hance	NC_023092
Saxifragales	Crassulaceae	*Sedum sarmentosum* Bunge	NC_023085
Saxifragales	Paeoniaceae	*Paeonia obovata* Maxim.	NC_026076
Saxifragales	Penthoraceae	*Penthorum chinense* Pursh	NC_023086
Trochodendrales	Trochodendraceae	*Tetracentron sinense* Oliv.	NC_021425

Supplementary Table S2. List of primer sequences used to close gaps and verify overlapping contigs.

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
Ericaceae			
Orthilia secunda	Orsec_291F	GTAAAGGGGGTCTGGGAAAA	Orsec_291R
	Orsec_291R	TTCCTATTTCTTCGCGTTCG	Orsec_291F
	Orsec_34L	CCCCCTTCTATCCACACCTT	Orsec_501L
	Orsec_501L	CTCTGGCCTCTCAGGAATTG	Orsec_34L
	Orsec_81R	TGATGTGGAAATTGGCTCTG	Orsec_L1R
	Orsec_A1F	TCTACCCTTTCCCGTAAGTTGA	Orsec_A1R
	Orsec_A1R	GGAAGGGGTTAAGTGCAACA	Orsec_A1F
	Orsec_A2R	CCCGGTTCAATTGTAATGATG	Orsec_O2F
	Orsec_B1F	TTCCGAGATGGAACTCTTGC	Orsec_B1R
	Orsec_B1R	CAACGAAAGTGACCACGAGA	Orsec_B1F
	Orsec_D1F	CGGCATGCCATCTTCTAAA	Orsec_D1R, Orsec_J1R
	Orsec_D1R	CCAATAATCCAATTGTTCAATCA	Orsec_D1F
	Orsec_F1F	CCAAGGGCTCAAGAATAAACC	Orsec_F1R
	Orsec_F1R	TCCATGATACAGCAGAGCAGA	Orsec_F1F

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Orsec_I1F	AAATTGCTTTGGGTCGTTTG	Orsec_I1R
	Orsec_I1R	AATCCCAATGAAAAGGCAGA	Orsec_I1F
	Orsec_J1R	TGCAACATTGTTAACTCGAGGA	Orsec_D1F
	Orsec_L1R	GCTGCTTGGCCTGTAGTAGG	Orsec_81R
	Orsec_M1R	TGCTCAAACAATCCCAATCA	Orsec_O1R
	Orsec_O1R	CCAATGGCGTTGGCTACTAT	Orsec_M1R
	Orsec_O2F	TGGACAATGAGGAAGACTGC	Orsec_A2R
Gentianaceae			
Bartonia virginica	Bavir_12F	CCCCCAGGATCTATAATTTACTC	Bavir_12R
	Bavir_12R	ATTGGTGAACCAGCAGATCC	Bavir_12F
	Bavir_19L	CCTTGGGGTTATCCTGCACT	Bavir_5R
	Bavir_19L3	ATGTTGGGGTGAACCAGAAA	Bavir_5R
	Bavir_19L4	AAAAGGAGTAAGCTTGGGACA	Bavir_50R2, Bavir_50R3
	Bavir_200F	CACGCAGAGGAACTAGGATTC	Bavir_200R

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Bavir_200R	CCTTGTTGTTCTAGTTGGATGTG	Bavir_200F
	Bavir_218R	ACATCCGTCCCAAGGTATCA	Bavir_C6R, Bavir_C8R
	Bavir_219F	ATCGAACCCGCATCTTCTC	Bavir_219R
	Bavir_219F2	GCATCGTTTCTCCTCCAAAA	Bavir_219R2
	Bavir_219R	TCCCTTGAACCTGTGTATGAAG	Bavir_219F
	Bavir_219R2	AGGCGTAGGTGCTTTTCTTC	Bavir_219F2
	Bavir_37F	ATTGCCTTGGACTTGTCGTT	Bavir_37R
	Bavir_37F2	TACCGGAACAAACGGCTATC	Bavir_37R2
	Bavir_37R	CGCACACACTCTCTTTCCAA	Bavir_37F
	Bavir_37R2	AAGCTAACGATGCGGGTTC	Bavir_37F2
	Bavir_50R2	GCTATGCATGGTTCCTTGGT	Bavir_19L4
	Bavir_50R3	CTGCTGCTATAGAAGTTCCATCT	Bavir_19L4
	Bavir_5R	TAGATGTCGGCCAAAAGCA	Bavir_19L3, Bavir_19L
	Bavir_A1F	GGCCCGAGAATTGATGTGTA	Bavir_A1R
	Bavir_A1R	TTCCCGCTGTTTTCTCATGT	Bavir_A1F

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Bavir_B2F	CGGTCCAGTAGGTCCGTAAA	Bavir_B2R
	Bavir_B2R	CTACCACGTGGAAACGCTCT	Bavir_B2F
	Bavir_C1F	TGGGTAACGGTATTCTGCCTA	Bavir_C1R
	Bavir_C1R	CGTTGCGGTCGGACTCTAT	Bavir_C1F
	Bavir_C6R	TGTTGGTAGCCCAGTTTTCC	Bavir_218R
	Bavir_C8R	GTCCTCCCTACCCACCAATC	Bavir_218R
	Bavir_D1F	CCTGGATACTCGGGTTCAAA	Bavir_D2R
	Bavir_D2R	AACCCCAGGTTAAGCGAGAT	Bavir_D1F
	Bavir_E1F	ACCTGAGAGCGGACAGCTAA	Bavir_E1R
	Bavir_E1R	GTTGTATGCTGCGTTCGAGA	Bavir_E1F
	Bavir_F1F	GTTTGATTCAGCGGGAGAAA	Bavir_F1R
	Bavir_F1R	CTTTGCCAAGGAGAAGATGC	Bavir_F1F
Exacum affine	Exaff_146F	ACCTTTCCGAAGTCCTGGAG	Exaff_146R
	Exaff_146F2	AGATTACGCCCCTACTCTGC	Exaff_146R2
	Exaff_146R	TCGCTATCAACTGCTTGTCC	Exaff_146F

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Exaff_146R2	TCCACAGACGACGAAACTCT	Exaff_146F2
	Exaff_14F2	TCCACGTGGTAGAACCTCCT	Exaff_14R2
	Exaff_14F3	GTAGGCCCCCATCGTCTAGT	Exaff_14R3
	Exaff_14F4	ACTATAGGCGGAGCAATTCG	Exaff_14R4
	Exaff_14L	TAGACGCCCCAGCAACTAAG	Exaff_14R
	Exaff_14R	CACCACCAACTGTAGCAGCA	Exaff_14L
	Exaff_14R2	GGTCCTGAAGCACAAGGAGA	Exaff_14F2
	Exaff_14R3	TGCTAGGGGTGGGATATTTG	Exaff_14F3
	Exaff_14R4	CACCACCAACTGTAGCAGCA	Exaff_14F4
	Exaff_20R	CAATATTCACCGGCCCAAGG	Exaff_63L3
	Exaff_235F	AATGTATCGCCCCATCTCAA	Exaff_235R
	Exaff_235R	GCTGGATCAACCCTTGAAAC	Exaff_235F
	Exaff_236F	AATCGGAATCGTGGGTAGTA	Exaff_236R
	Exaff_236R	TCAAGCTCTGGCAGATGGTA	Exaff_236F
	Exaff_289R	TGAGTTCAACCAAGCCAACC	Exaff_571R

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Exaff_30F	CAACGAATCCGAATGTTTGA	Exaff_30R
	Exaff_30R	GCCGATGATTTGGACGATAC	Exaff_30F
	Exaff_31F	TGCCATGGTTCCTTACTTCG	Exaff_31R
	Exaff_31L	CTTCTTGCTTTATCAAGGGAACAT	Exaff_G1F
	Exaff_31R	GTGGAGAACGGAACCAAGAA	Exaff_31F
	Exaff_38F	AGAGGGACGATTTCGTGAGA	Exaff_38R
	Exaff_38F2	GAAGCTCGGTAAAAGCAACG	Exaff_38R2
	Exaff_38R	TTGAACTAGCCATCCCTTCG	Exaff_38F
	Exaff_38R2	CCCTGGATAAGCTTCACGAC	Exaff_38F2
	Exaff_47L2	CTCCTCGAAGCGATAAACGA	Exaff_84R2
	Exaff_49R	AGCTCCACGCTTTCTTTCCT	Exaff_65L2
	Exaff_4F	AATTCGAGTGGCTGAAGCTG	Exaff_4R
	Exaff_4R	CAGGGTCAAGAACGACGAAT	Exaff_4F
	Exaff_571R	CCTCCCCGTTCAGTGAATTA	Exaff_289R
	Exaff_63L3	TTAAAAGTTGCTCCTGCTACTCA	Exaff_20R

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Exaff_65L2	AAGACATCACGATCCCTTGC	Exaff_49R
	Exaff_70F	GCTCTTATGCCTGCAGAAACA	Exaff_Q1f
	Exaff_76F	GGACGTTACCAAGGCTGAGA	Exaff_76R
	Exaff_76F2	TGCAGTCACTTCTTGTTTCCTG	Exaff_76R2
	Exaff_76R	GTTGGTAACCGACCCAAAGA	Exaff_76F
	Exaff_76R2	GACACATAAGAGCCCGAACC	Exaff_76F2
	Exaff_84R2	AGACGACTGAGCCAACTTGAG	Exaff_47L2
	Exaff_91F3	GAGCGCGAAAAATTGAGC	Exaff_91R3
	Exaff_91R	TGGTTGGTCATATAATCGTGCT	Exaff_E1R
	Exaff_91R3	GACTCGTGTTCTGGCTCGTC	Exaff_91F3
	Exaff_C1F	TCTCACATTCGGCTAGAGCA	Exaff_C1R
	Exaff_C1R	CCAAGGCTTTACCCCAAGAT	Exaff_C1F
	Exaff_D1F	GGTCGAATTTTCCATCTCCA	Exaff_D1R
	Exaff_D1R	ATCGGAGGAGTAGCTGCTGA	Exaff_D1F
	Exaff_E1F	GTTTTGTCTAGTGCCAACAAAGG	Exaff_E1R

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Exaff_E1R	AGAAGGGGTGGAAAGTGAGG	Exaff_91R, Exaff_E1F
	Exaff_F1F	CCCTGGAGAGATGGTTCACT	Exaff_F1R
	Exaff_F1R	ACGACAGAAAGGGGGATTG	Exaff_F1F
	Exaff_G1F	TAGTGGGGGAGTATGGGACA	Exaff_31L, Exaff_G1R
	Exaff_G1R	TCCGTGTCGCTAAATATCCA	Exaff_G1F
	Exaff_H1F	TCCGGCGTAGTTTTATACGG	Exaff_H1R
	Exaff_H1R	ATCCACAAGTACCGGCAGAG	Exaff_H1F
	Exaff_K1F	AAAATCGTGGTTGGGAAGG	Exaff_K1R
	Exaff_K1R	GAGTTGACCGCCAGACCTAC	Exaff_K1F
	Exaff_L1F	TCCTCCCGGAATAAAAGGAT	Exaff_L1R
	Exaff_L1R	GGTTTGCCTTGGTATCGTGT	Exaff_L1F
	Exaff_M1F	CCAAAGATCTCGGTCAGAGC	Exaff_M1R
	Exaff_M1R	CAAGTATGGTCGTCCCCTGT	Exaff_M1F
	Exaff_N1F	GGGTGAACGTACTCGTGAGG	Exaff_N1R
	Exaff_N1R	GCGCTCGTGCTACAGTTAAA	Exaff_N1F

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Exaff_O1F	TCAGAAAAGGGGTGGCTCTA	Exaff_O1R
	Exaff_O1R	TCCATCTCTCCTACCCGTTG	Exaff_O1F
	Exaff_Q1f	CCGATTAGCCGTTGTCATTT	Exaff_70F
	Exaff_R1F	CCACTCCAGTCGTTGCTTTT	Exaff_R1R
	Exaff_R1R	TGGGCGGAACAGGTCTACTA	Exaff_R1F
	Exaff_S1F	GCGTTCTTCGTCTCATCGTT	Exaff_S1R
	Exaff_S1R	GGGGCTTCGACTCTCACATA	Exaff_S1F
	Exaff_T1F	TTGGGGCCTCCTAAAAAGAT	Exaff_T1R
	Exaff_T1R	GCTTAAAGTGCGGGAATATGA	Exaff_T1F
	Exaff_U1F	GCTGGATTATTCGTCACTGC	Exaff_U1R
	Exaff_U1R	GTCGCTTGCCTAACAATCAA	Exaff_U1F
	Exaff_V1F	CTGAGGTACTCGGGTTCCAA	Exaff_V1R
	Exaff_V1R	TCACCCCTTTCACTTCCTTG	Exaff_V1F
	Exaff_W1F	TCCGCCTATAGTTCCTCGAA	Exaff_W1R
	Exaff_W1R	CAGATTGGGGAGGAAGATCA	Exaff_W1F

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Exaff_ycf2F	ACAGACAGAGTTCGAAGGGG	Exaff_ycf2R
	Exaff_ycf2R	TCCAGCTCCGTATCAAGGTC	Exaff_ycf2F
Exochaenium oliganthum	Exoli_1689F	CCCCTTTATTTCACCGGTTT	Ex_ol_1689R
	Exoli_1689R	GTGTGGACCGACGGACTTAC	Ex_ol_1689F
	Exoli_104F	CCCACAGCTTTGCTTTCAAT	Exol_104R
	Exoli_104R	CAAAACTTCTACCCCGAGCA	Exol_104F
	Exoli_90F	TGGGGTGATCTCGTAGTTCC	Exoli_90R
	Exoli_90R	GCCAGGGTAAGGAAGAAAGG	Exoli_90F
	Exoli_A1F	CAAGGTGGTCCTTGCTGATT	Exoli_A1R
	Exoli_A1R	CGAGTCCGCTTATCTCCAAC	Exoli_A1F
	Exoli_B1F	TCGAGCCGTGAAAAAGATTC	Exoli_B1R
	Exoli_B1R	GCCACTACTGGTGAGCCCTA	Exoli_B1F
	Exoli_C1F	GCTGGGGTTGCAAAATAAAA	Exoli_C1R
	Exoli_C1R	CGGACAAAGCAAGAAGGGTA	Exoli_C1F, Exoli_R1R
	Exoli_D1F	CACAATCTGGTTCTTGTTTCCA	Exoli_D1R

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Exoli_D1R	GGCAGAATACCGTCATCCAT	Exoli_D1F
	Exoli_E1F	CAACTGCGAAATAGGCACAA	Exoli_E1R
	Exoli_E1R	GAGGGGGAGTCGATTATTCC	Exoli_E1F
	Exoli_F1F	AGAGCACGTAGGGCTTTGAA	Exoli_F1R
	Exoli_F1R	GAAAAACTGGGTTGCGCTAT	Exoli_F1F
	Exoli_G1F	TAGCACCATGCCAAATGTGT	Exoli_G1R
	Exoli_G1R	TTTGCAGCTTTTGTTGTTGC	Exoli_G1F
	Exoli_H1F	TCCCTTGCCTAACAATCAAA	Exoli_H1R
	Exoli_H1R	GGGATCAGTTGGACCTTTGA	Exoli_H1F
	Exoli_I1F	GCTTCCTCGTTTCACTTTGC	Exoli_I1R
	Exoli_I1R	CCACGCGAAGGGTTTAGTTA	Exoli_I1F
	Exoli_J1F	TAGGGCGTATCGTCCAAATC	Exoli_J1R
	Exoli_J1R	CGGCGATAAGGTGCTAAAAG	Exoli_J1F
	Exoli_K1F	AGCCTTTGCACAATTTGCTT	Exoli_K1R
	Exoli_K1R	GAATGAAAGGCGTCCATTGT	Exoli_K1F

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Exoli_L1F	GCATGGGAACAGGTTCATCT	Exoli_L1R
	Exoli_L1R	CTCTACCCAGGATCCCAACA	Exoli_L1F
	Exoli_M1F	GATCCAACTCACATTCGGCC	Exoli_M1R
	Exoli_M1R	AAATCCGCGGTTCCTAATGG	Exoli_M1F
	Exoli_N1F	AAAAGCACTTGCCATTCGTT	Exoli_N1R
	Exoli_N1R	TTCTTCTCTCCATCGGACCA	Exoli_N1F
	Exoli_O1F	CTTCCTCAGCCAGGCAATAG	Exoli_O1R
	Exoli_O1R	AGTTTGCGAAAGATGCAGGT	Exoli_O1F
	Exoli_P1F	TGGACAAAGGTAAACATCTTGG	Exoli_Q1F
	Exoli_Q1F	AATTTTTCGCAAACCCCTCT	Exoli_P1F
	Exoli_R1R	CGACTCCTCGTGATCGACTT	Exoli_C1R
	Exoli_S1F	CCAAAGATCTCCGTCAGAGC	Exoli_S1R
	Exoli_S1R	TTTGGATTCAAAGCCCTACG	Exoli_S1F
	Exoli_T1F	TGTACAAGGGCGTGCTGTAG	Exoli_T1R
	Exoli_T1R	ACAACGTCGATGAAGACGTG	Exoli_T1F

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Exoli_T2F	GGTTACACCTCCAACCGAAA	Exoli_T2R
	Exoli_T2R	GAAGACGTGTGGGTGCACTA	Exoli_T2F
	Exoli_T3F	TTGGTTTACGCACGAATGAA	Exoli_T3R
	Exoli_T3R	CAATACCCACGCCAAGAAAT	Exoli_T3F
	Exoli_T4F	TTTCATCCACAAACGCAGAG	Exoli_T4R
	Exoli_T4R	CATGCCCAGACGGATAAACT	Exoli_T4F
Obolaria virginica	Obvir_1068L	TGCATTCACACCATTCCAAC	Obvir_E1F
	Obvir_108R	GAACATAGAAAGGCGGGATG	Obvir_221L
	Obvir_10L	GGATTGCCTCACGAAATAGC	Obvir_90R
	Obvir_221L	CGTCGGATGCTGGATATCTT	Obvir_108R
	Obvir_403R	CTGGGTAGCTGACCCTTTGA	Obvir_B1F
	Obvir_90L2	AAAGAAGGATATGCTTGAAATGA	Obvir_A2F, Obvir_A4F
	Obvir_90R	GGATTGCAAGGGTCAGTCAT	Obvir_10L
	Obvir_A2F	CCAAAAACTGCTCAGCAACA	Obvir_90L2, Obvir_A2R
	Obvir_A2R	CCCTCGCCCTAGGTTTTAAT	Obvir_A2F, Obvir_A4F

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Obvir_A4F	GCATCTACCATTATCCCCACA	Obvir_90L2, Obvir_A2R
	Obvir_B1F	GCAAAGCCCTATGGGTTGTA	Obvir_403R
	Obvir_C1F	TCAAGTCCACCACGAAGACA	Obvir_C1R
	Obvir_C1R	GGTTGGGGATTTTGTGAAAG	Obvir_C1F
	Obvir_C2F	GAGGAGGGCCTTGAAAAGTT	Obvir_C2R
	Obvir_C2R	AGCAAGTCAAGTCGCACGTT	Obvir_C2F
	Obvir_D1F	CTTGGCTTGGACAGGTCATT	Obvir_D1R
	Obvir_D1R	GATAGCTCCATGGGCAAAAG	Obvir_D1F
	Obvir_E1F	CCTGAAACCTTGGCACAGAT	Obvir_E1R, 1069L
	Obvir_E1R	TGTCGAATGAGTTTGGAAAGA	Obvir_E1F
	Obvir_ccsAF	CGATGTCAGGGCTTTTAACG	Obvir_ccsAR
	Obvir_ccsAR	TACGATTCGTGTCGGTTCAC	Obvir_ccsAF
	Obvir_ycf2F	CTCCAGGGATGAATCGAAAA	Obvir_ycf2R
	Obvir_ycf2R	AGGGTGCTATTGTTCCTCCA	Obvir_ycf2F
Voyria clavata	Vocla_21F2	CCCAATGCTGTCCTAGTTGA	Vocla_21R2

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Vocla_21R	GCAGCATCCAAAATGCCTAT	Vocla_B1R
	Vocla_21R2	TGTGAATTGCGCGAAAGTAG	Vocla_21F2
	Vocla_6L	GGCTCTACTCCGGGTAAAAA	Vocla_C1F
	Vocla_B1F	TCGATGAACGTTTGATTTTCC	Vocla_B1R
	Vocla_B1R	TCGAAGTAACCTCCTTTGATCC	Vocla_B1F, Vocla_21R
	Vocla_C1F	TGTAGACCCCCGAACAAAAG	Vocla_6L, Vocla_C1R
	Vocla_C1R	AAAAGTGGCTCGGTGGTATG	Vocla_C1F
Polygalaceae			
Epirixanthes elongata	Epelo_15L	GTTCGAGTACCAGGCGCTAC	Epelo_416R
	Epelo_15R	GTAGCGCCTGGTACTCGAAC	Epelo_441L
	Epelo_20L	AGGCCTACGGGTCGTAAACT	Epelo_39R
	Epelo_21L	TCTAGCCCCTCTGGGATGTA	Epelo_847R
	Epelo_21R	GGGGAACTCGAATTTTTGGT	Epelo_416L
	Epelo_2828L	TCTAAGGGTAGCCTGCTCCA	Epelo_416L
	Epelo_348L	TGCACGGCTACACAGAAATC	Epelo_416R

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Epelo_348R	AGGGGCTCAGGACATCTCTC	Epelo_39L, Epelo_847L
	Epelo_39L	CCGTCACACTAGGGAAGCTG	Epelo_348R, Epelo 441R
	Epelo_39R	CATGTCAAGCCCTGGTAAGG	Epelo_20L, Epelo_847L
	Epelo_416L	GTGGGCGTTAGAGCATTGAT	Epelo_2828L, Epelo_21R
	Epelo_416R	CCCCCATACATGGTCTTACG	Epelo_15L
	Epelo_441L	GGGTGATCTATCCAGGACCA	Epelo_15R
	Epelo_441R	GCTACTGGACTCTCGCCATC	Epelo_39L
	Epelo_847L	TCGACGAAGACGTGTAGGTG	Epelo_39R, Epelo_348R
	Epelo_847R	GATCTCGCGGATCTTTCGAT	Epelo_21L
Epirixanthes pallida	Eppal_2427F	ATCTCCCGGATAAGCCTCAC	Eppal_2427R
	Eppal_2427R	TGCCCTGGCTAAACCTATTG	Eppal_2427F
	Eppal_701F	TCTTGATTGGAAGGGACACC	Eppal_701R
	Eppal_701R	GGGCGTTAGAGCATTGAGAG	Eppal_701F
	Eppal_A1F	CATCGGTCCACACAGTTGTC	Eppal_A1R
	Eppal_A1R	AGCGATGGAGTTAGCAATCG	Eppal_A1F

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Eppal_B1F	TGCGTTTTGGGAGCTTCTAT	Eppal_B2R
	Eppal_B2R	GCGCCTAACCCTATGAGTTG	Eppal_B1F
	Eppal_C1F	GAATCCCATGAAGGACGAAA	Eppal_C1R
	Eppal_C1R	ACGGGAATCCCCTTTATTTG	Eppal_C1F
	Eppal_D1F	AGCATGGACCCACTCCTATG	Eppal_D2R
	Eppal_D2R	CACATGGAGCCATCTCCTTA	Eppal_D1F
	Eppal_E1F	TCATTCATGGGCGTTGATAA	Eppal_E1R
	Eppal_E1R	CAGAGCGCAAGCTAGTGATG	Eppal_E1F
	Eppal_F1F	CCGCCATCCTACCTAATGAA	Eppal_F1R
	Eppal_F1R	CTCATCGCCTCGCTTTATCT	Eppal_F1F
	Eppal_G1F	TTCATCGAATACGGCTTTCC	Eppal_K1F, Eppal_G1R
	Eppal_G1R	AGGGGGAAGGGTTAAGGATT	Eppal_G1F
	Eppal_H2F	ACGAAATCGCATTGATAGCC	Eppal_I1F
	Eppal_I1F	TCAACCCACCCTTAGTACCG	Eppal_H2F
	Eppal_I1R	AACTACGAGATCGCCCCTTT	Eppal_J3R

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Eppal_J3R	CGTAGTTCCTACGGGGTGAA	Eppal_I1R
	Eppal_K1F	GGCATGGCATCTTATGAAGG	Eppal_G1F
	Eppal_L1F	TGGAACTCCAACAGGCATAA	Eppal_L2R
	Eppal_L2R	GGATTCAACAAAGACGGTTCA	Eppal_L1F
Polygala arillata	Poari_2F	GAATGAGGAGCCGTATGAGG	Poari_2R
	Poari_2R	TCCCTACGAAATACCAGACGA	Poari_2F
	Poari_A1F	TGATTGGTCGTATAATCGTGGT	Poari_A1R
	Poari_A1R	TGGGACGTTTACCAGTGTCA	Poari_A1F, Poari_C1F
	Poari_B1R	GCGCTAACCTTGGTATGGAA	Poari_B4F
	Poari_B4F	GGAAATCGGCCACATTAAAA	Poari_B1R
	Poari_C1F	TGCTGCAGCTACAAAGTGTG	Poari_A1R
Salomonia cantoniensis	Sacan_326R	CAACCGGTCGAGTAAGATGAG	Sacan_347L
	Sacan_347L	TGCTTCTGGCCTGGATAAAC	Sacan_326R
	Sacan_B1F	CACGGAATGTATTTGCACCA	Sacan_B1R
	Sacan_B1R	TTGGTTCACGGGTACAACCT	Sacan_B1F

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Sacan_C1F	AGCTGTGCTGCTGCTACAAA	Sacan_C1R
	Sacan_C1R	TGGGACGTTTACCAGTGTCA	Sacan_C1F, Sacan_M1F
	Sacan_D1R	GGATTGAGCCGAATACAACC	Sacan_Na1F
	Sacan_E1F	TTAGCGAATTCGTGTGCTTG	Sacan_E1R
	Sacan_E1R	ATCGGCCAAAATAACCATGA	Sacan_E1F
	Sacan_F1F	GCGCTAACCTTGGTATGGAA	Sacan_P1R
	Sacan_F1R	TGGCTAGGTAAGCGTCCTGT	Sacan_F1F
	Sacan_G1F	TCCCCATGAGTTCCAGTCTC	Sacan_G1R
	Sacan_G1F	TCCCCATGAGTTCCAGTCTC	Sacan_G1R
	Sacan_G1R	ATCCAGGATTTGAACGGATG	Sacan_G1F
	Sacan_G1R	ATCCAGGATTTGAACGGATG	Sacan_G1F
	Sacan_H1F	TCGGTTTCCATTTTGGTTGT	Sacan_H1R
	Sacan_H1R	CTACTCAGCCCAGAGCCTTG	Sacan_H1F
	Sacan_I1F	AAGGGGTTTCAAAAACCAAGA	Sacan_I1R
	Sacan_I1R	CTTCGTTTGCAGCAACACTC	Sacan_I1F

Taxon	Primer name	Primer sequence (5’ to 3’)	Primer pair
	Sacan_J1F	CATGCACGGTTTTGAATGAG	Sacan_J1R
	Sacan_J1R	TTCTTGGTTTCGTCCAGTCA	Sacan_J1F
	Sacan_M1F	TGCTTGGTCGTATCATCGTG	Sacan_C1R
	Sacan_Na1F	TGAACAGATCCGGTGAAAAA	Sacan_D1R
	Sacan_Nb1F	TTTCAACTTGCTCTGCTCCT	Sacan_R1F
	Sacan_O1F	AGGGTGTCCGTGACGTGT	Sacan_O1R
	Sacan_O1R	AGGGGTTGTGGATACTGCTG	Sacan_O1F
	Sacan_P1R	TGCTCTATTTCGTTCCTTGG	Sacan_F1F
	Sacan_R1F	CCCGTTCTCTACGTTTTTGC	Sacan_Nb1F
	Sacan_T1F	AGGCCATTTAGTCCATGTCG	Sacan_T1R
	Sacan_T1R	CAGAAAGAGGCTGACCCAAC	Sacan_T1F

Supplementary Table S3. Results of partition-finder analyses, summarizing final partitioning schemes and the optimal DNA or amino-acid substitution models associated with each data partition: (a) 'ORF-only' (open reading frame only) matrix partitioned using the 'GxC' (gene by codon) partitioning scheme; (b) A version of the matrix with pseudogenes included, partitioned using the GxC scheme; (c) Amino-acid matrix, partitioned by gene. Genes are indicated before the underscore; the 'pos' term after the underscore indicates the codon position for protein-coding genes.

Partition	Best Model	Partition subset
a) ORF-only (GxC scheme)		
1	GTR+I+Γ	accD_pos1, clpP_pos1
2	GTR+I+Γ	accD_pos2, ccsA_pos1, ndhF_pos1
3	GTR+I+Γ	accD_pos3, atpE_pos3, infA_pos3, rpl20_pos3, rpoC2_pos3
4	GTR+I+Γ	atpA_pos1, atpI_pos1, petA_pos1, rpoB_pos1, rps12_pos1
5	GTR+I+Γ	atpA_pos2, atpB_pos2, psbB_pos2, rps12_pos
6	GTR+I+Γ	atpA_pos3, atpI_pos3, ndhK_pos3, petA_pos3, petB_pos3, psbB_pos3, psbI_pos3, rps16_pos3
7	GTR+I+Γ	atpB_pos1, psbB_pos1
8	GTR+I+Γ	atpB_pos3, atpF_pos3, ndhJ_pos3, psbC_pos3, rpl33_pos3, rpoC1_pos3
9	GTR+I+Γ	atpE_pos1, atpF_pos1, rpl2_pos3, rps14_pos1, rps19_pos1, rps2_pos1
10	GTR+I+Γ	atpE_pos2, cemA_pos2, ndhC_pos2, ndhE_pos2, petL_pos1, petL_pos2, psaJ_pos1, psbL_pos3, psbT_pos1,
11	GTR+I+Γ	atpF_pos2, psbF_pos3, rpoC2_pos2, rps11_pos2
12	GTR+I+Γ	atpH_pos1, petD_pos1, psaA_pos1, psaB_pos1, psbN_pos1 psbZ_pos1, rpoC1_pos2, rps12_pos3, rps2_pos2, ycf4_pos2
13	GTR+Γ	atpH_pos2
14	GTR+I+Γ	atpH_pos3, psbZ_pos3, rpl36_pos3, rps14_pos3, rps18_pos3, rps4_pos
15	GTR+I+Γ	atpI_pos2, ndhG_pos2, ndhJ_pos2, rpoB_pos2, rps14_pos2
16	GTR+I+Γ	ccsA_pos2
17	GTR+I+Γ	ccsA_pos3, ndhD_pos3, ndhE_pos3, petD_pos3, psaC_pos3, rps15_pos3
18	GTR+I+Γ	cemA_pos1, rpoC2_pos1, rps18_pos2
19	GTR+Γ	cemA_pos3, clpP_pos3
20	GTR+I+Γ	clpP_pos2
21	GTR+I+Γ	infA_pos1, rpl16_pos1, rpl20_pos1, rpl33_pos1, rps11_pos1, rps16_pos1, rps4_pos1

Partition	Best Model	Partition subset
22	GTR+I+Γ	infA_pos2, ndhH_pos2, petA_pos2, rpl16_pos2, rps19_pos2
23	GTR+I+Γ	matK_pos1
24	GTR+Γ	matK_pos2
25	GTR+I+Γ	matK_pos3, rpl16_pos3, rpoA_pos3
26	GTR+I+Γ	ndhA_pos1, ndhD_pos1, psaI_pos2, psaJ_pos2, psbH_pos1, psbJ_pos1, psbM_pos1, psbT_pos2, rps8_pos2
27	GTR+I+Γ	ndhA_pos2, petN_pos3, psaI_pos1, psbE_pos3, psbJ_pos3, rps16_pos2
28	GTR+I+Γ	ndhA_pos3, ndhG_pos3, psbT_pos3, rps3_pos3, rps8_pos3
29	GTR+Γ	ndhB_pos1, petN_pos1, petN_pos2, psbF_pos1, psbI_pos2, psbL_pos1, psbM_pos2, rrn5
30	GTR+I+Γ	ndhB_pos2, psbI_pos1, psbN_pos2
31	GTR+Γ	ndhB_pos3, ndhE_pos1, ndhJ_pos1, rpl23_pos2, rpl23_pos3, rps7_pos3
32	GTR+I+Γ	ndhC_pos1, ndhH_pos1, ndhI_pos1, ndhK_pos2, rps4_pos2
33	GTR+I+Γ	ndhC_pos3, ndhG_pos1, psbD_pos3, psbK_pos3, psbN_pos3, rpl20_pos2, ycf3_pos3
34	GTR+I+Γ	ndhD_pos2, petG_pos1
35	GTR+I+Γ	ndhF_pos2, ndhI_pos2, psbK_pos1
36	GTR+I+Γ	ndhF_pos3
37	GTR+I+Γ	ndhH_pos3, ndhI_pos3, rbcL_pos3, rps11_pos3
38	GTR+I+Γ	ndhK_pos1, rpl14_pos1, rpoC1_pos1, ycf4_pos1
39	GTR+I+Γ	petB_pos1, psbD_pos1
40	GTR+I+Γ	petB_pos2, psbD_pos2
41	GTR+I+Γ	petD_pos2, psbA_pos1, psbA_pos2, psbE_pos1, psbL_pos2
42	GTR+Γ	petG_pos2
43	GTR+I+Γ	petG_pos3, petL_pos3, psaI_pos3, rps15_pos2
44	GTR+I+Γ	psaA_pos2, psbE_pos2
45	GTR+I+Γ	psaA_pos3, psaB_pos3, psbH_pos3, psbM_pos3, rpl14_pos3, rpoB_pos3, rps2_pos3, ycf4_pos3
46	GTR+I+Γ	psaB_pos2, psaC_pos2, psbC_pos2, psbF_pos2, psbJ_pos2, psbK_pos2, psbZ_pos2
47	GTR+I+Γ	psaC_pos1, rpl36_pos1, rpl36_pos2, rps7_pos1, ycf3_pos1
48	GTR+I+Γ	psaJ_pos3
49	GTR+I+Γ	psbA_pos3
50	GTR+I+Γ	psbC_pos1
51	GTR+I+Γ	psbH_pos2
52	GTR+I+Γ	rbcL_pos1
53	GTR+I+Γ	rbcL_pos2

Partition	Best Model	Partition subset
54	GTR+I+Γ	rpl14_pos2, rpl23_pos1, rpl2_pos1, rpl2_pos2, rps7_pos2, ycf3_pos2
55	GTR+Γ	rpl22_pos1
56	GTR+I+Γ	rpl22_pos2
57	GTR+I+Γ	rpl22_pos3, rpl32_pos3
58	GTR+Γ	rpl32_pos1, rpl32_pos2
59	GTR+I+Γ	rpl33_pos2, rpoA_pos1, rps15_pos1, rps18_pos1, rps3_pos1
60	GTR+I+Γ	rpoA_pos2, rps3_pos2
61	GTR+I+Γ	rps8_pos1
62	GTR+I+Γ	rps19_pos3
63	GTR+Γ	rrn16, rrn4_5
64	GTR+I+Γ	rrn23
65	GTR+Γ	ycf2_pos1
66	GTR+Γ	ycf2_pos2
67	GTR+Γ	ycf2_pos3

b) Pseudogenes included (GxC scheme)

1	GTR+I+Γ	accD_pos1, clpP_pos1, rpl32_pos1
2	GTR+I+Γ	accD_pos2
3	GTR+I+Γ	accD_pos3, infA_pos3, rpoC2_pos3
4	GTR+I+Γ	atpA_pos1
5	GTR+I+Γ	atpA_pos2, rpoB_pos2, rpoC1_pos2, rps2_pos2
6	GTR+I+Γ	atpA_pos3, atpI_pos3, ndhG_pos3, ndhK_pos3, petA_pos3, petB_pos3, psbB_pos3, psbI_pos3
7	GTR+I+Γ	atpB_pos1, atpI_pos1, petA_pos1, petA_pos2, rpl16_pos2, rps12_pos1
8	GTR+I+Γ	atpB_pos2, psbB_pos2, psbF_pos2
9	GTR+I+Γ	atpB_pos3, ndhJ_pos3
10	GTR+Γ	atpE_pos1, atpF_pos1, infA_pos2, rpl2_pos3, rps19_pos1
11	GTR+I+Γ	atpE_pos2, rps19_pos2, rps7_pos2
12	GTR+I+Γ	atpE_pos3, ndhA_pos3, psbT_pos3, rpl20_pos3, rps16_pos3, rps3_pos3, rps8_pos3
13	GTR+I+Γ	atpF_pos2, rpoC2_pos2, rps11_pos2, rps18_pos2
14	GTR+I+Γ	atpF_pos3, rpl32_pos2, rpl33_pos3, rpoB_pos3, rpoC1_pos3, rps18_pos3, ycf4_pos3
15	GTR+I+Γ	atpH_pos1, ndhC_pos2, petB_pos1, petD_pos1, psaB_pos1, psbA_pos1, psbA_pos2, psbN_pos1
16	GTR+Γ	atpH_pos2
17	GTR+I+Γ	atpH_pos3, psbH_pos3, psbN_pos3, psbZ_pos3, rpl36_pos3

Partition	Best Model	Partition subset
18	GTR+I+Γ	atpI_pos2, petB_pos2
19	GTR+I+Γ	ccsA_pos1, ndhF_pos1
20	GTR+I+Γ	ccsA_pos2, ndhI_pos2
21	GTR+I+Γ	ccsA_pos3, ndhD_pos3, ndhE_pos3, petD_pos3, psaC_pos3, rps15_pos3
22	GTR+I+Γ	cemA_pos1, psbM_pos1, rpl20_pos2, rpl33_pos2, rpoC2_pos1, rps15_pos1, rps18_pos1
23	GTR+Γ	cemA_pos2, ndhE_pos2, petL_pos1, petL_pos2, psaJ_pos1, psbL_pos3, psbZ_pos1, ycf4_pos2
24	GTR+Γ	cemA_pos3, clpP_pos3
25	GTR+I+Γ	clpP_pos2, petG_pos1
26	GTR+I+Γ	infA_pos1, rpl22_pos2, rpl33_pos1, rpoA_pos1, rpoA_pos2, rps3_pos2
27	GTR+I+Γ	matK_pos1
28	GTR+I+Γ	matK_pos2, ndhG_pos1, petG_pos3, petL_pos3, psaI_pos1, rps15_pos2
29	GTR+I+Γ	matK_pos3, ndhI_pos3, rpoA_pos3
30	GTR+I+Γ	ndhA_pos1, ndhD_pos1, psaI_pos2, psaJ_pos2, psbH_pos1, psbJ_pos1, psbL_pos2, psbT_pos2, rps8_pos2
31	GTR+I+Γ	ndhA_pos2, ndhG_pos2, ndhJ_pos2, ndhK_pos2, rps12_pos2, rps16_pos2, rps4_pos2
32	GTR+I+Γ	ndhB_pos1, petN_pos1, petN_pos2, psaA_pos2, psbC_pos2, psbE_pos2, psbF_pos1, psbI_pos2, psbL_pos1, psbM_pos2, rpl14_pos2, rrn4_5, rrn5, ycf3_pos2
33	GTR+I+Γ	ndhB_pos2, psbI_pos1, psbN_pos2, psbZ_pos2
34	GTR+Γ	ndhB_pos3, rpl23_pos1, rpl23_pos2, rpl23_pos3, rps7_pos3
35	GTR+I+Γ	ndhC_pos1, ndhH_pos1, ndhI_pos1, ndhK_pos1, rpl14_pos1, rps11_pos1, rps12_pos3, rps14_pos2, rps16_pos1, rps4_pos1, ycf4_pos1
36	GTR+I+Γ	ndhC_pos3, psaB_pos3, psbC_pos3
37	GTR+I+Γ	ndhD_pos2, petD_pos2
38	GTR+Γ	ndhE_pos1, ycf2_pos2
39	GTR+I+Γ	ndhF_pos2, psbK_pos1, psbM_pos3
40	GTR+I+Γ	ndhF_pos3
41	GTR+Γ	ndhH_pos2, psbK_pos2
42	GTR+I+Γ	ndhH_pos3, rbcL_pos3, rpl16_pos3, rps11_pos3
43	GTR+Γ	ndhJ_pos1, psaA_pos1, rpl36_pos2, rps7_pos1
44	GTR+I+Γ	petG_pos2, rbcL_pos2
45	GTR+Γ	petN_pos3, psbE_pos3, psbF_pos3, psbJ_pos3
46	GTR+I+Γ	psaA_pos3, rpl14_pos3
47	GTR+I+Γ	psaB_pos2, psaC_pos2, psbD_pos2

Partition	Best Model	Partition subset
48	GTR+I+Γ	psaC_pos1, rpl2_pos1, rpl2_pos2, ycf3_pos1
49	GTR+I+Γ	psaI_pos3, psbD_pos3, psbK_pos3, rps14_pos3, rps2_pos3, rps4_pos3, ycf3_pos3
50	GTR+I+Γ	psaJ_pos3
51	GTR+I+Γ	psbA_pos3
52	GTR+I+Γ	psbB_pos1, psbC_pos1, psbE_pos1, psbT_pos1
53	GTR+I+Γ	psbD_pos1, psbJ_pos2, rrn16
54	GTR+I+Γ	psbH_pos2
55	GTR+I+Γ	rbcL_pos1
56	GTR+Γ	rpl16_pos1, rpl20_pos1
57	GTR+Γ	rpl22_pos1
58	GTR+I+Γ	rpl22_pos3, rpl32_pos3
59	GTR+Γ	rpl36_pos1, ycf2_pos1
60	GTR+I+Γ	rpoB_pos1, rpoC1_pos1, rps14_pos1, rps2_pos1
61	GTR+I+Γ	rps3_pos1, rps8_pos1
62	GTR+I+Γ	rps19_pos3
63	GTR+I+Γ	rrn23
64	GTR+Γ	ycf2_pos3

c) Amino acid scheme (partitioned by gene)

1	JTT+Γ+F	accD, rpl22, rpl32
2	JTT+I+Γ	atpA, atpB
3	CPREV+Γ	atpE
4	JTT+Γ+F	atpF, ndhG
5	CPREV+Γ	atpH
6	JTT+Γ+F	atpI, ndhE, ndhI, ndhJ, petA, petL, rpoB
7	JTT+I+Γ+F	ccsA
8	JTT+I+Γ+F	cemA, psaI, rpoA, rps15, rps18
9	CPREV+Γ	clpP
10	JTT+Γ	infA, rps4
11	JTT+Γ+F	matK
12	JTT+I+Γ+F	ndhA
13	JTT+Γ+F	ndhB, petD, psaA, psaB, psbN, psbZ
14	JTT+I+Γ+F	ndhC, ndhK, psbK, rpl14, rpoC1
15	JTT+I+Γ+F	ndhD, rpoC2, rps3, rps8
16	JTT+I+Γ+F	ndhF
17	JTT+I+Γ	ndhH
18	JTT+I+Γ	petB, psbE
19	CPREV+Γ	petG, petN, psb
20	JTT+I+Γ	psaC, psbJ
21	MTMAM+Γ	psaJ

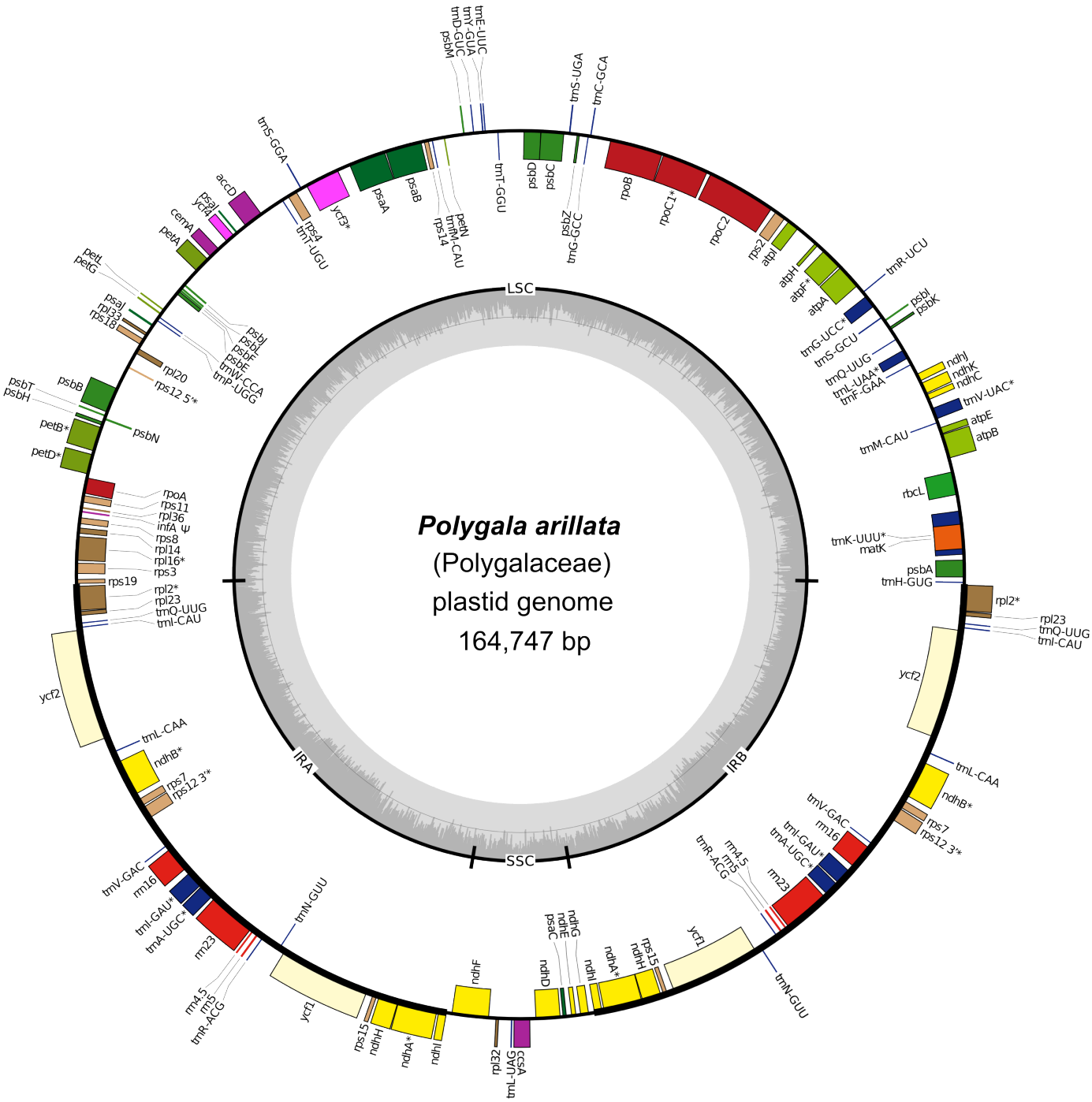
Partition	Best Model	Partition subset
22	CPREV+I+Γ	psbA
23	JTT+I+Γ	psbB, psbL, psbT
24	CPREV+I+Γ	psbC
25	CPREV+I+Γ	psbD
26	CPREV+Γ	psbF, psbM, rpl36, rps12
27	JTT+I+Γ	psbH
28	LG+I+Γ	rbcL
29	JTT+Γ	rpl2, ycf3
30	CPREV+Γ	rpl16
31	CPREV+I+Γ	rpl20
32	JTT+Γ	rpl23, rps7
33	JTT+Γ	rpl33, rps19
34	JTT+Γ+F	rps2, ycf2
35	JTT+Γ	rps11, rps16
36	JTT+Γ	rps14
37	JTT+Γ+F	ycf4

Supplementary Table S4. Species with partially assembled plastid genomes.

Family	Species	No. raw reads	No. contigs	Combined length (bp)	No. genes with intact reading frame[1]	No. rDNA genes[1]	No. tRNA genes[1]
Ericaceae	*Orthilia secunda*[2]	15,484,936	1	145,723	66	4	30
Ericaceae	*Pyrola minor*	8,314,630	17	127,096	68	4	30
Gentianaceae	*Voyria caerulea*	29,597,488	9	46,826	17	4	13
Polygalaceae	*Epirixanthes elongata*[2]	19,295,216	2[3]	16,593[3]	13[3]	3	5
Polygalaceae	*Salomonia cantoniensis*[2]	15,896,526	5	135,290	75	2	30

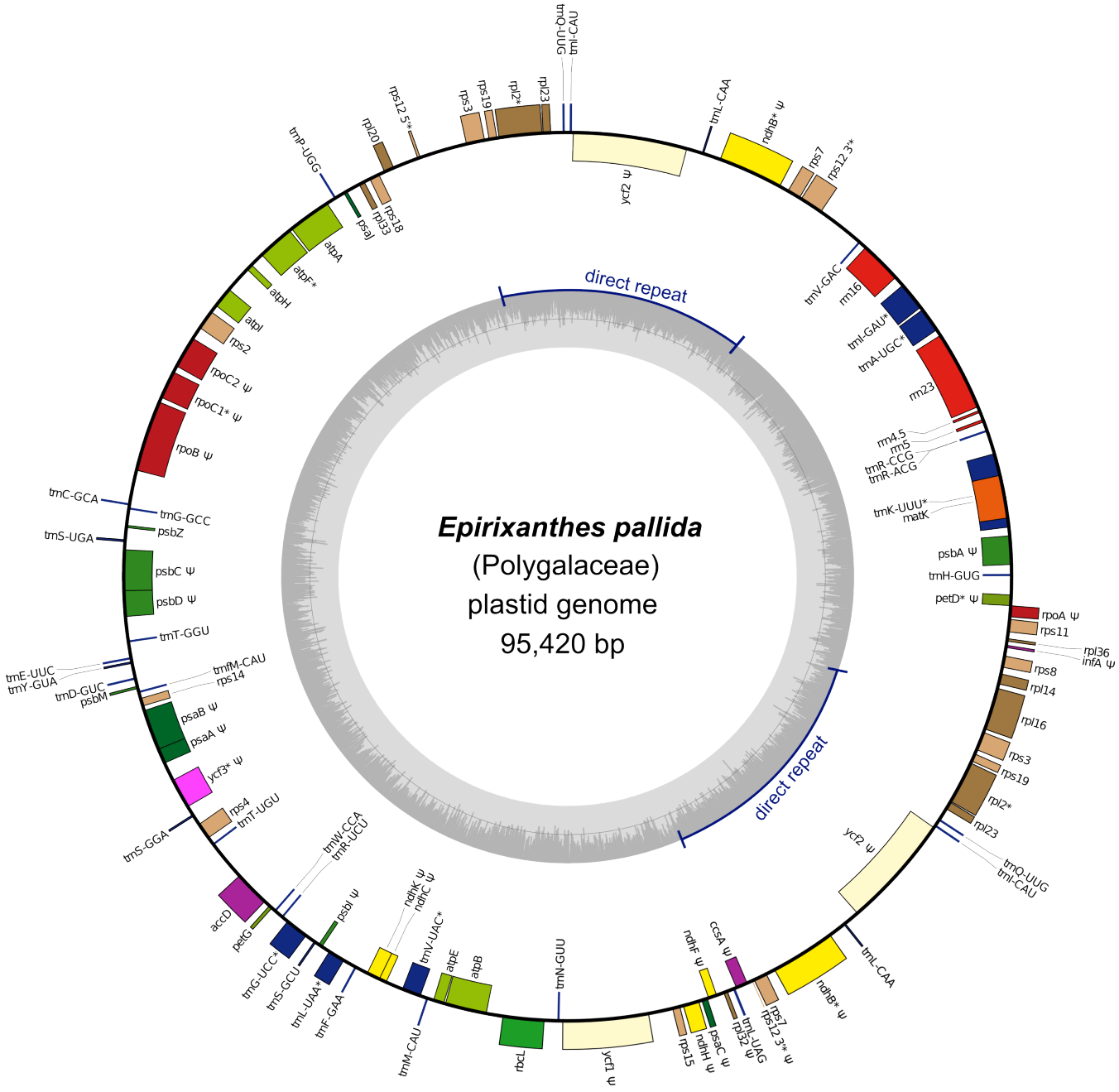
[1] Genes found in the inverted repeat counted once
[2] Species for which additional PCR and Sanger sequencing were used to join *de novo* contigs into larger fragments.
[3] Calculated with all possible plastid sequences, including low-depth regions (see Fig. S8)

Figure S1. Circular plastome map of *Polygala arillata* (Polygalaceae). Genes located inside the circle are transcribed clockwise, those outside are transcribed counterclockwise. The grey circle marks the GC content; the inner circle marks a 50% threshold. Thick black lines indicate inverted repeat (IR) copies. Genes with introns are indicated with asterisks (*). Pseudogenes are marked as 'Ψ'.



- photosystem I
- photosystem II
- cytochrome b_6/f complex
- ATP synthase
- NAD(P)H dehydrogenase
- RubisCO large subunit
- photosystem I assembly factor (ycf3 & ycf4)
- ribosomal proteins (SSU)
- ribosomal proteins (LSU)
- matK
- accD, ccsA, cemA, infA
- hypothetical chloroplast reading frames (ycf1 & ycf2)
- RNA polymerase
- transfer RNAs
- ribosomal RNAs

Figure S2. Circular plastome map of *Epirixanthes pallida* (Polygalaceae). Genes located inside the circle are transcribed clockwise, those outside are transcribed counterclockwise. The grey circle marks the GC content; the inner circle marks a 50% threshold. Thick blue lines indicate direct repeat copies. Genes with introns are indicated with asterisks (*). Pseudogenes are marked as 'Ψ'**.**



- photosystem I
- photosystem II
- cytochrome b_6/f complex
- ATP synthase
- NAD(P)H dehydrogenase
- RubisCO large subunit
- photosystem I assembly factor (ycf3)
- ribosomal proteins (SSU)
- ribosomal proteins (LSU)
- matK
- accD, ccsA
- hypothetical chloroplast reading frames (ycf1 & ycf2)
- RNA polymerase
- transfer RNAs
- ribosomal RNAs

Figure S3. Circular plastome map of *Exacum affine* (Gentianaceae). Genes located inside the circle are transcribed clockwise, those outside are transcribed counterclockwise. The grey circle marks the GC content; the inner circle marks a 50% threshold. Thick black lines indicate inverted repeat (IR) copies. Genes with introns are indicated with asterisks (*). The truncated *ycf*1 pseudogene is marked as 'Ψ'**.**

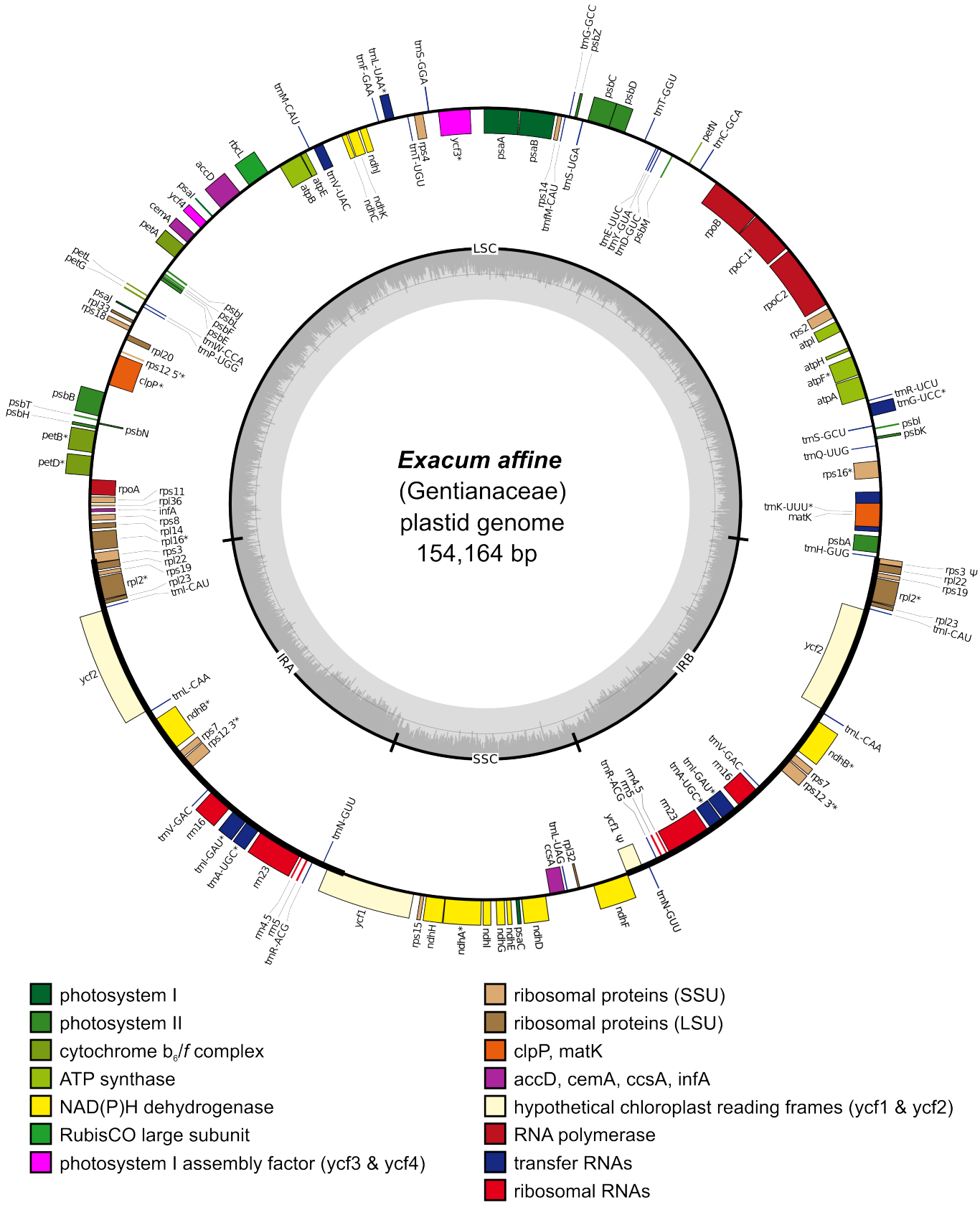

Exacum affine
(Gentianaceae)
plastid genome
154,164 bp
LSC
IRA
IRB
SSC
photosystem I
photosystem II
cytochrome b6/f complex
ATP synthase
NAD(P)H dehydrogenase
RubisCO large subunit
photosystem I assembly factor (ycf3 & ycf4)
ribosomal proteins (SSU)
ribosomal proteins (LSU)
clpP, matK
accD, cemA, ccsA, infA
hypothetical chloroplast reading frames (ycf1 & ycf2)
RNA polymerase
transfer RNAs
ribosomal RNAs

Figure S4. Circular plastome map of *Exochaenium oliganthum* (Gentianaceae). Genes located inside the circle are transcribed clockwise, those outside are transcribed counterclockwise. The grey circle marks the GC content; the inner circle marks a 50% threshold. Thick black lines indicate inverted repeat (IR) copies. Genes with introns are indicated with asterisks (*). Pseudogenes are marked as 'Ψ'**.**

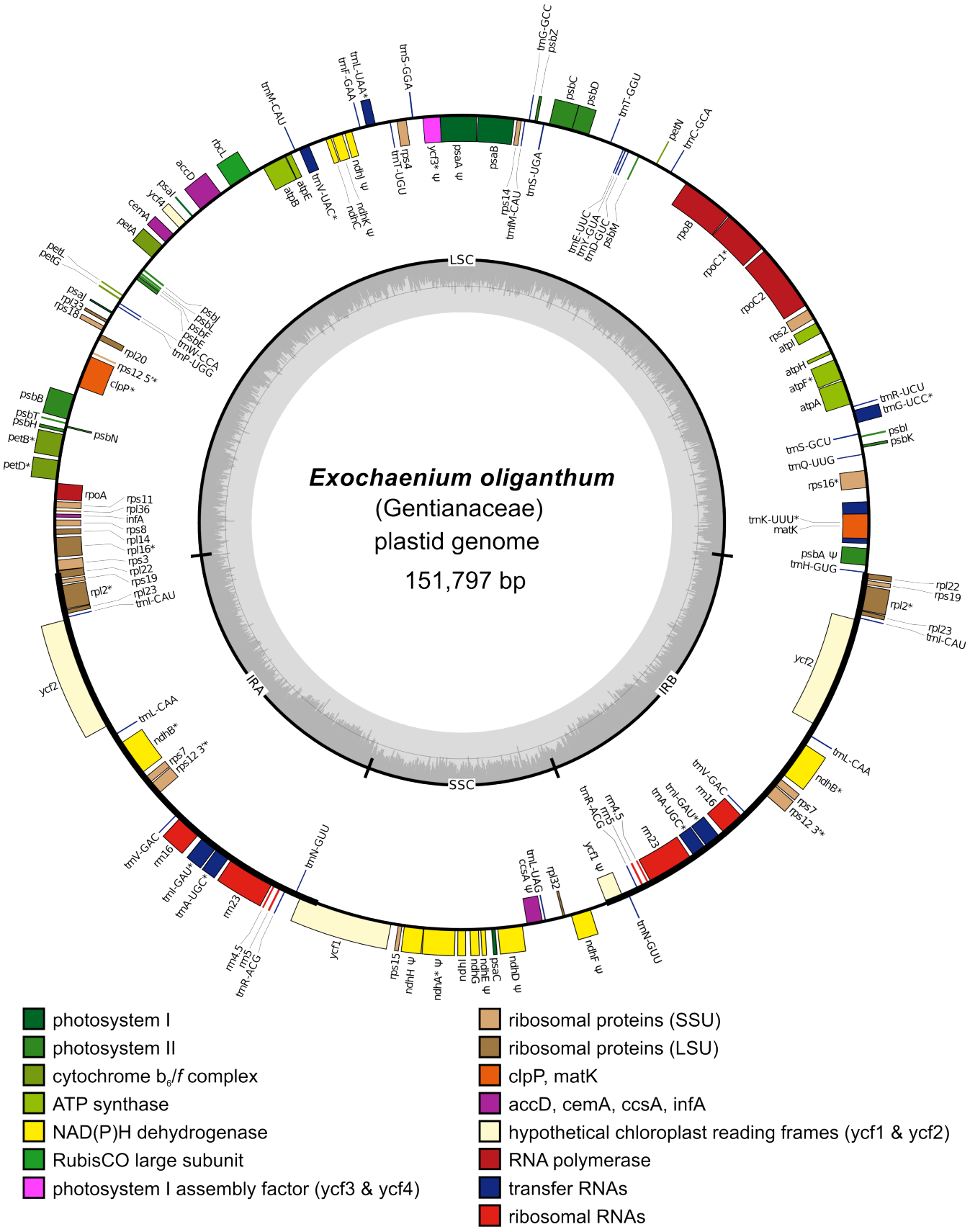

Exochaenium oliganthum
(Gentianaceae)
plastid genome
151,797 bp
LSC
IRA
IRB
SSC
photosystem I
photosystem II
cytochrome b6/f complex
ATP synthase
NAD(P)H dehydrogenase
RubisCO large subunit
photosystem I assembly factor (ycf3 & ycf4)
ribosomal proteins (SSU)
ribosomal proteins (LSU)
clpP, matK
accD, cemA, ccsA, infA
hypothetical chloroplast reading frames (ycf1 & ycf2)
RNA polymerase
transfer RNAs
ribosomal RNAs

Figure S5. Circular plastome map of *Bartonia virginica* (Gentianaceae). Genes located inside the circle are transcribed clockwise, those outside are transcribed counterclockwise. The grey circle marks the GC content; the inner circle marks a 50% threshold. Thick black lines indicate inverted repeat (IR) copies. Genes with introns are indicated with asterisks (*). Pseudogenes are marked as 'Ψ'.

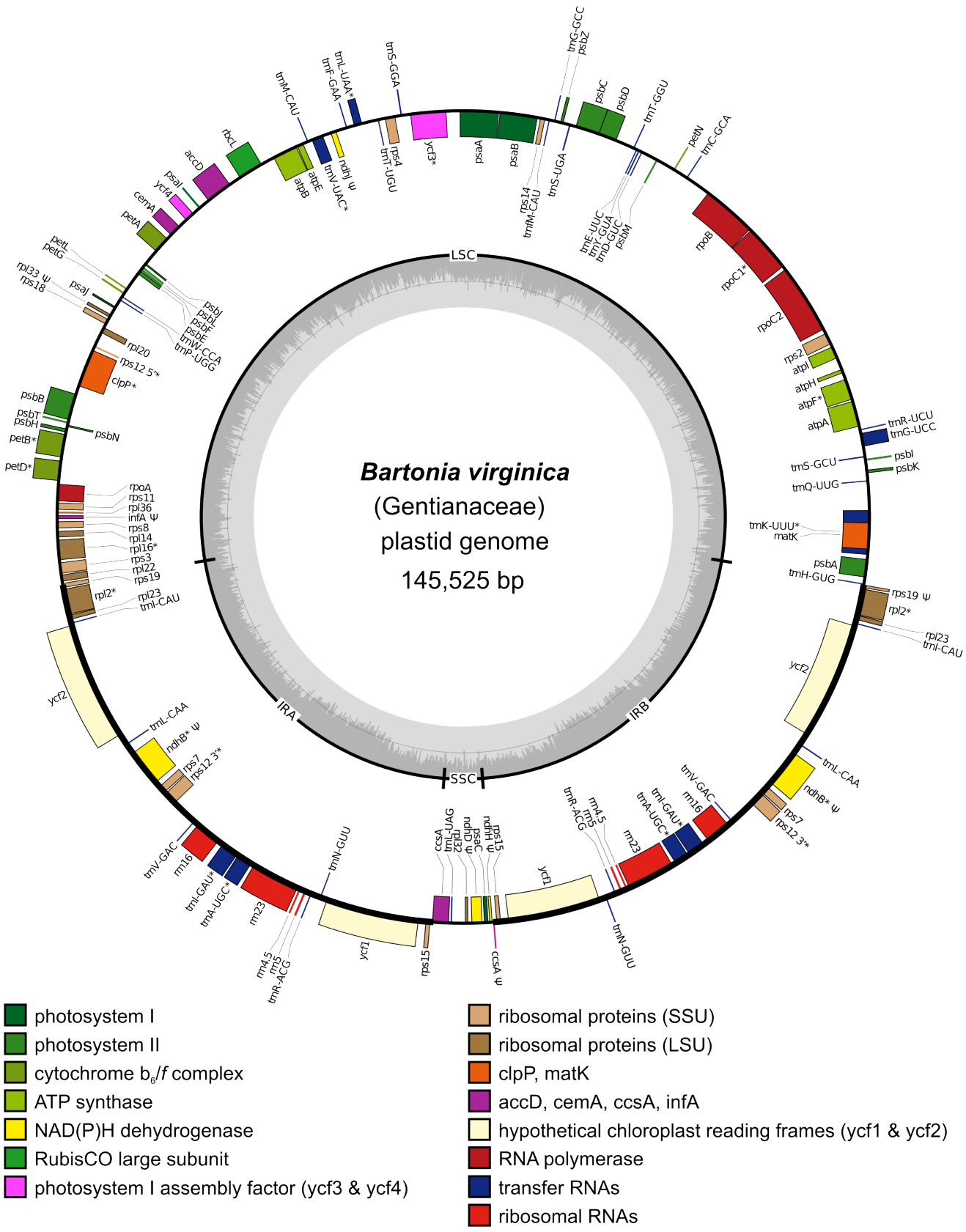

Bartonia virginica
(Gentianaceae)
plastid genome
145,525 bp
LSC
IRA
IRB
SSC
photosystem I
photosystem II
cytochrome b6/f complex
ATP synthase
NAD(P)H dehydrogenase
RubisCO large subunit
photosystem I assembly factor (ycf3 & ycf4)
ribosomal proteins (SSU)
ribosomal proteins (LSU)
clpP, matK
accD, cemA, ccsA, infA
hypothetical chloroplast reading frames (ycf1 & ycf2)
RNA polymerase
transfer RNAs
ribosomal RNAs

Figure S6. Circular plastome map of *Obolaria virginica* (Gentianaceae). Genes located inside the circle are transcribed clockwise, those outside are transcribed counterclockwise. The grey circle marks the GC content; the inner circle marks a 50% threshold. Thick black lines indicate inverted repeat (IR) copies. Genes with introns are indicated with asterisks (*). Pseudogenes are marked as 'Ψ'.

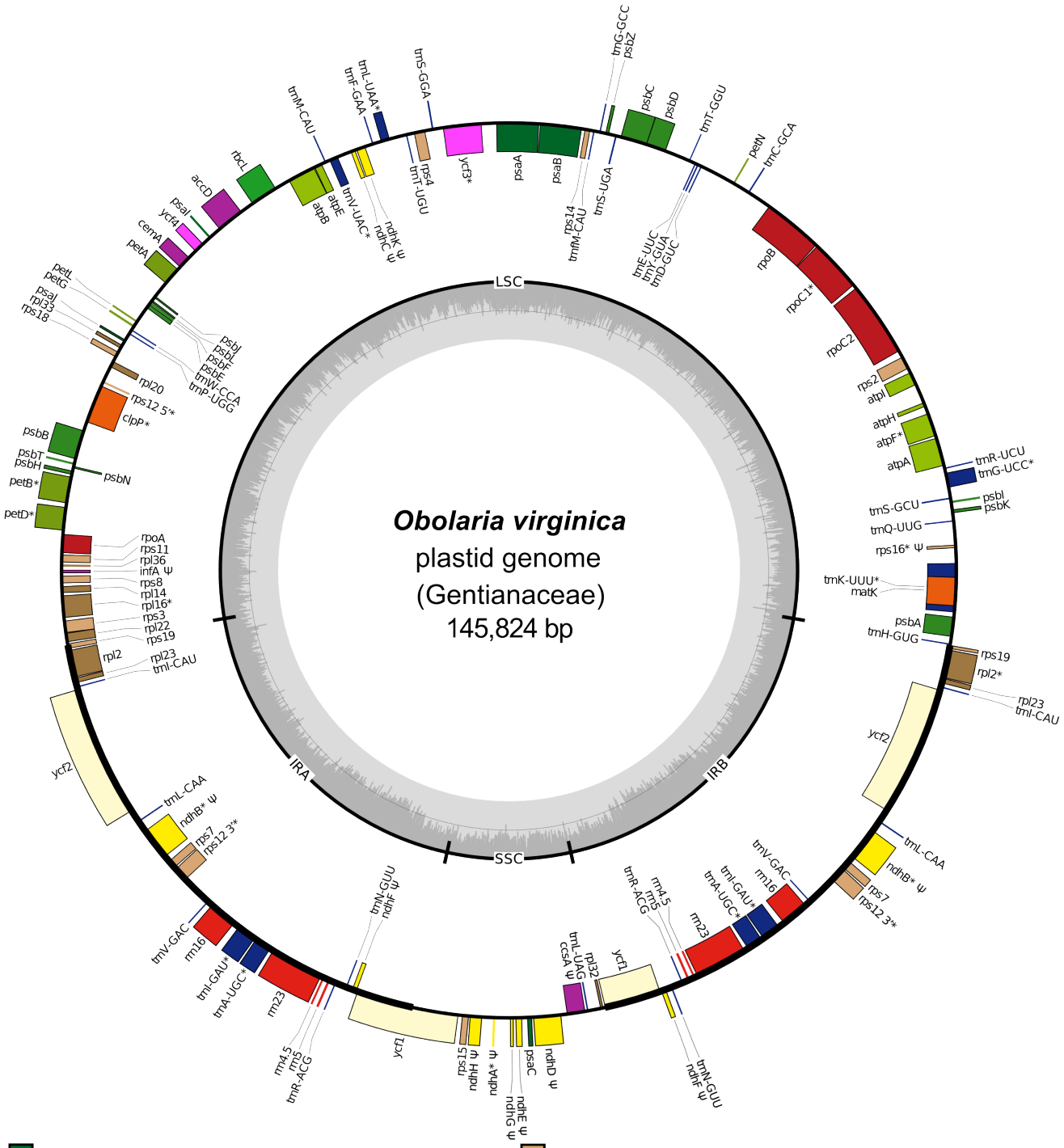

Obolaria virginica
plastid genome
(Gentianaceae)
145,824 bp
LSC
IRA
IRB
SSC
photosystem I
photosystem II
cytochrome b6/f complex
ATP synthase
NAD(P)H dehydrogenase
RubisCO large subunit
photosystem I assembly factor (ycf3 & ycf4)
ribosomal proteins (SSU)
ribosomal proteins (LSU)
clpP, matK
accD, cemA, ccsA, infA
hypothetical chloroplast reading frames (ycf1 & ycf2)
RNA polymerase
transfer RNAs
ribosomal RNAs

Figure S7. Circular plastome map of *Voyria clavata* (Gentianaceae). Genes located inside the circle are transcribed clockwise, those outside are transcribed counterclockwise. The grey circle marks the GC content; the inner circle marks a 50% threshold. Thick black lines indicate inverted repeat (IR) copies. Genes with introns are indicated with asterisks (*). Pseudogenes are marked as 'Ψ'.

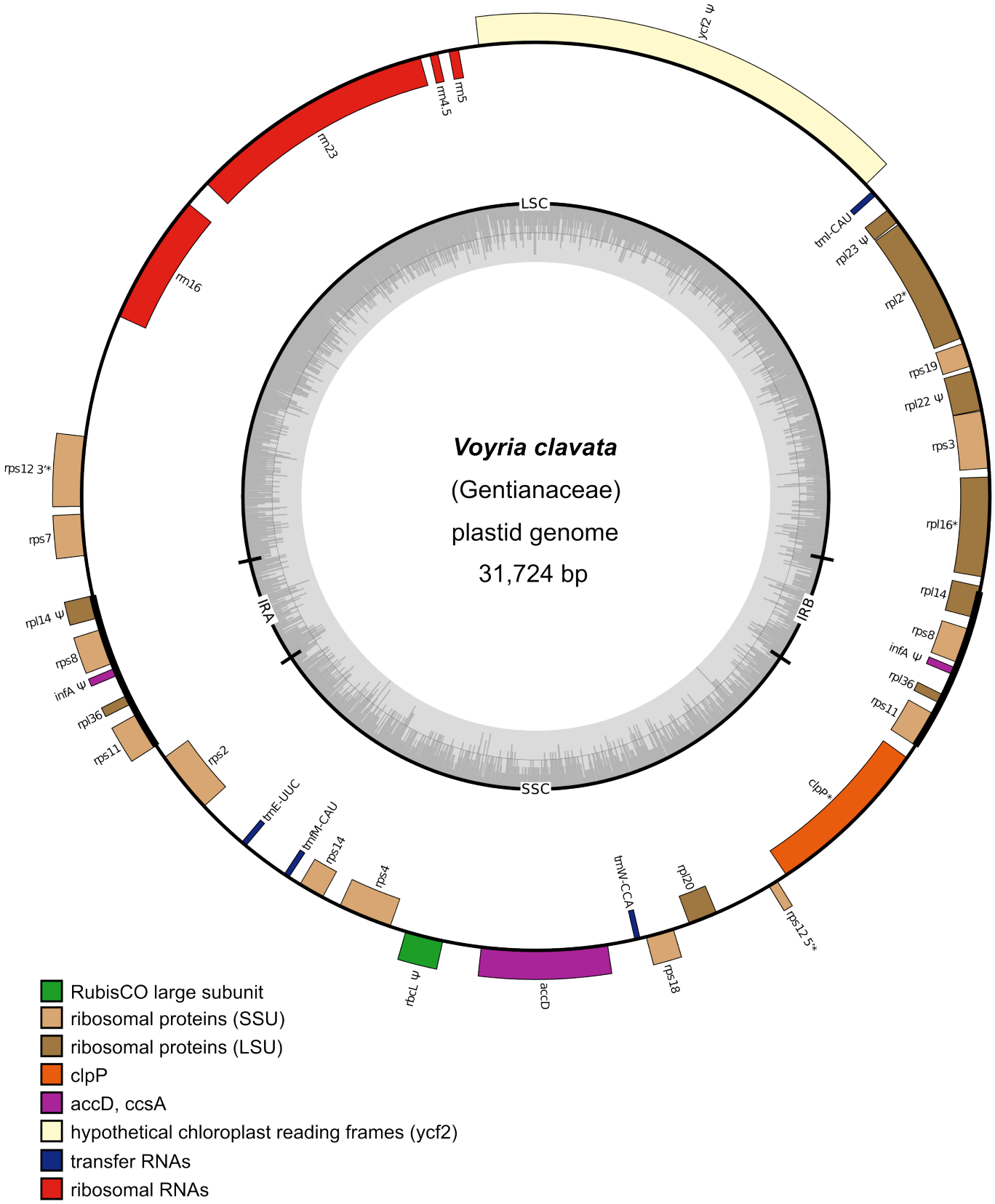

Voyria clavata
(Gentianaceae)
plastid genome
31,724 bp
LSC
IRA
IRB
SSC
ycf2 ψ
trnI-CAU
rpl23 ψ
rpl2*
rps19
rpl22 ψ
rps3
rpl16*
rpl14
rps8
infA ψ
rpl36
rps11
clpP*
rps12 5'*
rps18
rpl20
trnV-CCA
accD
rbcL ψ
rps4
rps14
trnfM-CAU
trnE-UUC
rps2
rps11
rpl36
infA ψ
rps8
rpl14 ψ
rps7
rps12 3'*
rrn16
rrn23
rrn4.5
rrn5
RubisCO large subunit
ribosomal proteins (SSU)
ribosomal proteins (LSU)
clpP
accD, ccsA
hypothetical chloroplast reading frames (ycf2)
transfer RNAs
ribosomal RNAs

Figure S8. Linearized plastome map of the draft partial assembly of *Epirixanthes elongata* (Polygalaceae). Black lines below the map indicate the Sanger-connected Illumina contigs in the assembly, and the relative read depth is indicated (1X = ~200X read depth; 8X = ~eight times coverage; 4X = ~four times coverage; 2X = ~two times coverage; see main text). Arrows indicate regions of assembly where gaps and contig overlaps were respectively connected or confirmed using Sanger sequencing (not to scale; thin dashed lines are sequenced regions not represented in *de novo* contigs). Pseudogenes are indicated in red. Genes with introns are indicated with an asterisk (*). A ~1300 bp contig is not shown here (it has an uninterrupted copy of the 3'-*rps*12 and *rps*7; note that I did not recover the 5'-*rps*12 in the gene set), as its relative connection to the main assembly has not been confirmed. Scale is in kb.

Epirixanthes elongata

(Polygalaceae)

Plastid genome partial assembly

15,293 bp

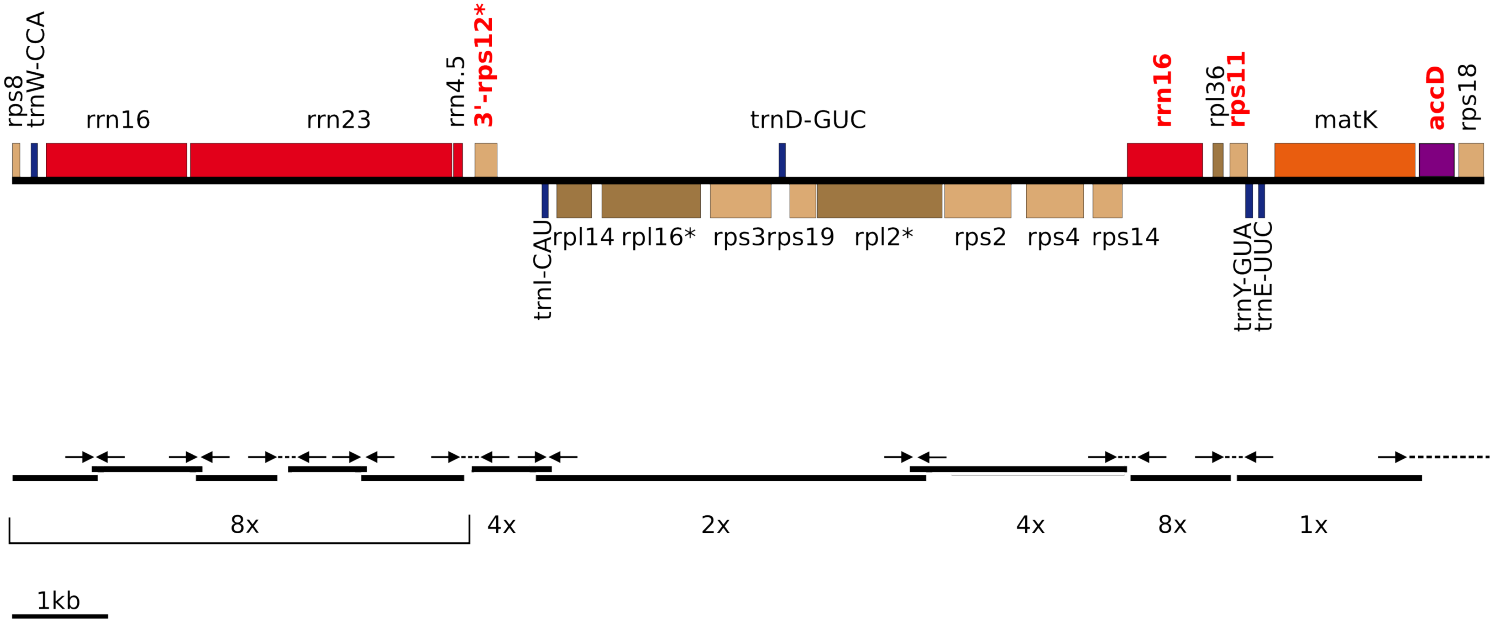


Figure S9. Angiosperm phylogeny inferred in an unpartitioned likelihood analysis of 82 plastid genes (ORF-only; see text and Table 3). Log likelihood score of best tree: -1,527,008.266. Bootstrap support values are indicated beside branches. Thick lines indicate 100% bootstrap support; '--' indicates <50% bootstrap support. Eudicot families where mycoheterotrophy has evolved are indicated in blue. The scale bar indicates estimated substitutions per site.

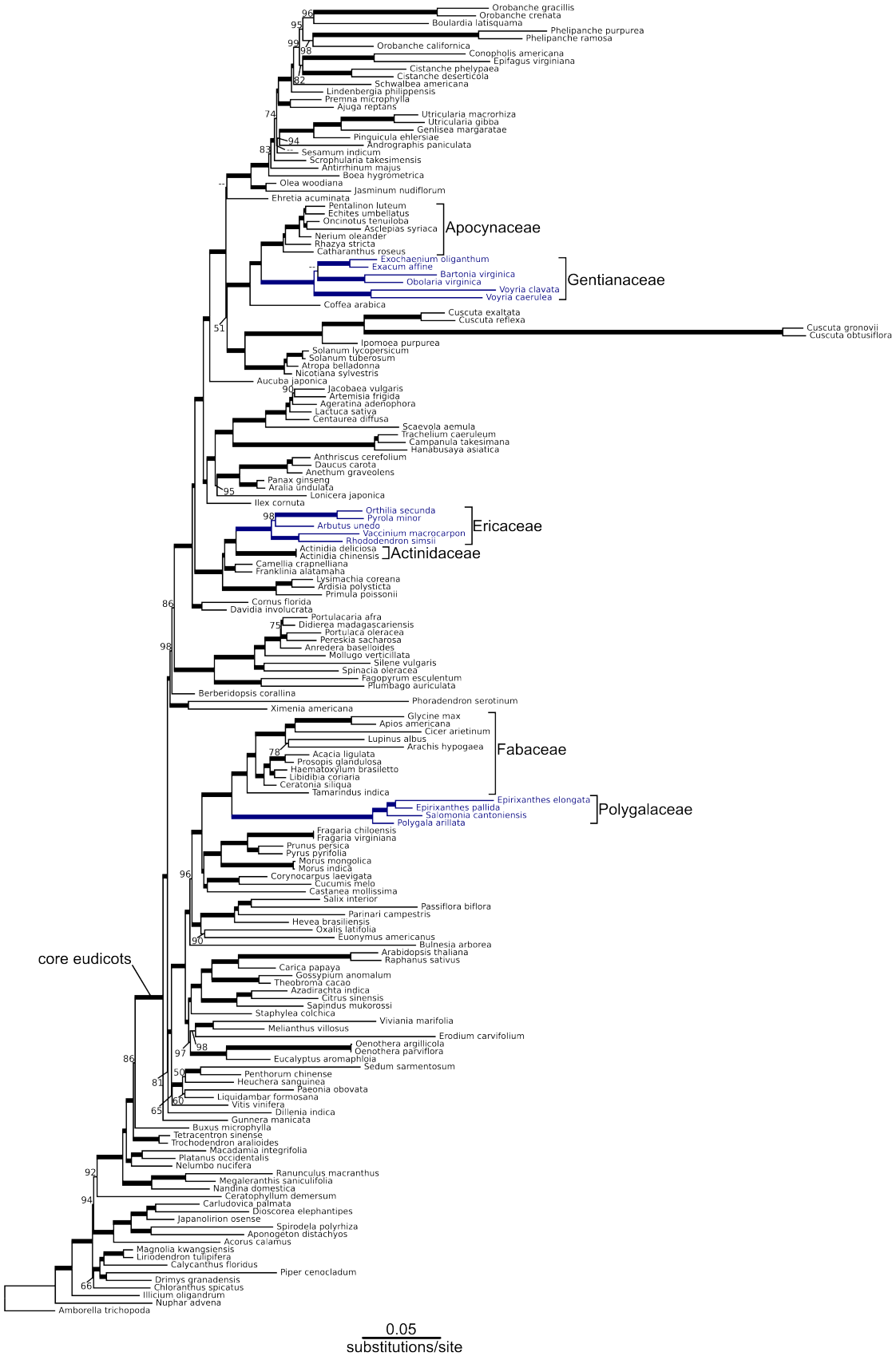

Orobanche gracillis
Orobanche crenata
Boulardia latisquama
Phelipanche purpurea
Phelipanche ramosa
Orobanche californica
Conopholis americana
Epifagus virginiana
Cistanche phelypaea
Cistanche deserticola
Schwalbea americana
Lindenbergia philippensis
Premna microphylla
Ajuga reptans
Utricularia macrorhiza
Utricularia gibba
Genlisea margaratae
Pinguicula ehlersiae
Andrographis paniculata
Sesamum indicum
Scrophularia takesimensis
Antirrhinum majus
Boea hygrometrica
Olea woodiana
Jasminum nudiflorum
Ehretia acuminata
Pentalinon luteum
Echites umbellatus
Oncinotus tenuiloba
Asclepias syriaca
Nerium oleander
Rhazya stricta
Catharanthus roseus
Apocynaceae
Exochaenium oliganthum
Exacum affine
Bartonia virginica
Obolaria virginica
Voyria clavata
Voyria caerulea
Gentianaceae
Coffea arabica
Cuscuta exaltata
Cuscuta reflexa
Cuscuta gronovii
Cuscuta obtusiflora
Ipomoea purpurea
Solanum lycopersicum
Solanum tuberosum
Atropa belladonna
Nicotiana sylvestris
Aucuba japonica
Jacobaea vulgaris
Artemisia frigida
Ageratina adenophora
Lactuca sativa
Centaurea diffusa
Scaevola aemula
Trachelium caeruleum
Campanula takesimana
Hanabusaya asiatica
Anthriscus cerefolium
Daucus carota
Anethum graveolens
Panax ginseng
Aralia undulata
Lonicera japonica
Ilex cornuta
Orthilia secunda
Pyrola minor
Arbutus unedo
Vaccinium macrocarpon
Rhododendron simsii
Ericaceae
Actinidia deliciosa
Actinidia chinensis
Actinidaceae
Camellia crapnelliana
Franklinia alatamaha
Lysimachia coreana
Ardisia polysticta
Primula poissonii
Cornus florida
Davidia involucrata
Portulacaria afra
Didierea madagascariensis
Portulaca oleracea
Pereskia sacharosa
Anredera baselloides
Mollugo verticillata
Silene vulgaris
Spinacia oleracea
Fagopyrum esculentum
Plumbago auriculata
Berberidopsis corallina
Phoradendron serotinum
Ximenia americana
Glycine max
Apios americana
Cicer arietinum
Lupinus albus
Arachis hypogaea
Acacia ligulata
Prosopis glandulosa
Haematoxylum brasiletto
Libidibia coriaria
Ceratonia siliqua
Tamarindus indica
Fabaceae
Epirixanthes elongata
Epirixanthes pallida
Salomonia cantoniensis
Polygala arillata
Polygalaceae
Fragaria chiloensis
Fragaria virginiana
Prunus persica
Pyrus pyrifolia
Morus mongolica
Morus indica
Corynocarpus laevigata
Cucumis melo
Castanea mollissima
Salix interior
Passiflora biflora
Parinari campestris
Hevea brasiliensis
Oxalis latifolia
Euonymus americanus
Bulnesia arborea
Arabidopsis thaliana
Raphanus sativus
Carica papaya
Gossypium anomalum
Theobroma cacao
Azadirachta indica
Citrus sinensis
Sapindus mukorossi
Staphylea colchica
Viviania marifolia
Melianthus villosus
Erodium carvifolium
Oenothera argillicola
Oenothera parviflora
Eucalyptus aromaphloia
Sedum sarmentosum
Penthorum chinense
Heuchera sanguinea
Paeonia obovata
Liquidambar formosana
Vitis vinifera
Dillenia indica
Gunnera manicata
Buxus microphylla
Tetracentron sinense
Trochodendron aralioides
Macadamia integrifolia
Platanus occidentalis
Nelumbo nucifera
Ranunculus macranthus
Megaleranthis saniculifolia
Nandina domestica
Ceratophyllum demersum
Cariudovica palmata
Dioscorea elephantipes
Japanolirion osense
Spirodela polyrhiza
Aponogeton distachyos
Acorus calamus
Magnolia kwangsiensis
Liriodendron tulipifera
Calycanthus floridus
Piper cenocladum
Drimys granadensis
Chloranthus spicatus
Illicium oligandrum
Nuphar advena
Amborella trichopoda
core eudicots
0.05
substitutions/site

Figure S10. Angiosperm phylogeny inferred in a likelihood analysis of 78 translated plastid genes (ORF-only) using the gene partitioning scheme (see text and Table S3 for details). Log likelihood score of best tree: -708,394.300. Bootstrap support values are indicated beside branches. Thick lines indicate 100% bootstrap support; '--' indicates <50% bootstrap support. Eudicot families where mycoheterotrophy has evolved are indicated in blue. The scale bar indicates estimated substitutions per residue.

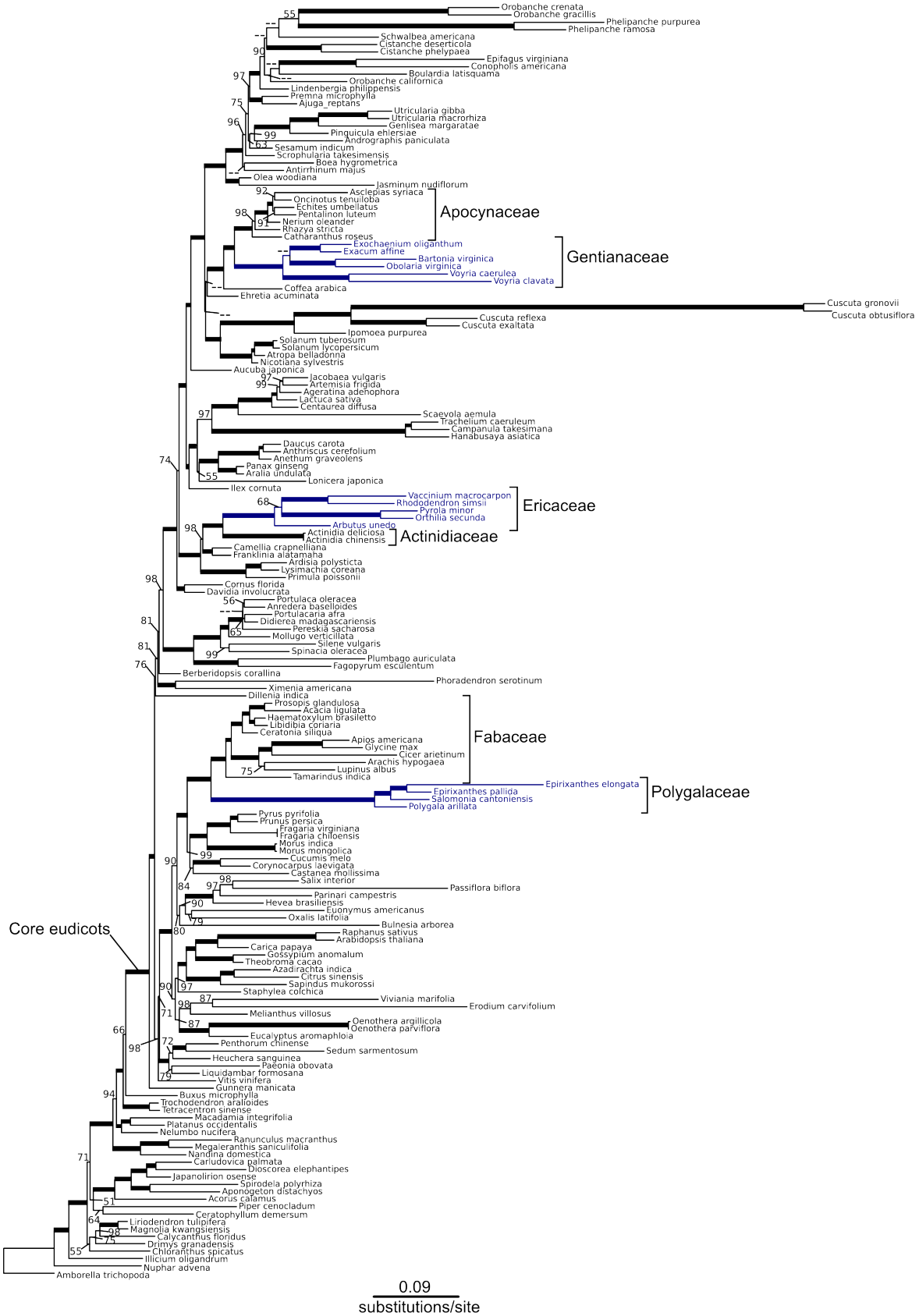

Orobanche crenata
Orobanche gracillis
Phelipanche purpurea
Phelipanche ramosa
Schwalbea americana
Cistanche deserticola
Cistanche phelypaea
Epifagus virginiana
Conopholis americana
Boulardia latisquama
Orobanche californica
Lindenbergia philippensis
Premna microphylla
Ajuga_reptans
Utricularia gibba
Utricularia macrorhiza
Genlisea margaratae
Pinguicula ehlersiae
Andrographis paniculata
Sesamum indicum
Scrophularia takesimensis
Boea hygrometrica
Antirrhinum majus
Olea woodiana
Jasminum nudiflorum
Asclepias syriaca
Oncinotus tenuiloba
Echites umbellatus
Pentalinon luteum
Nerium oleander
Rhazya stricta
Catharanthus roseus
Apocynaceae
Exochaenium oliganthum
Exacum affine
Bartonia virginica
Obolaria virginica
Voyria caerulea
Voyria clavata
Gentianaceae
Coffea arabica
Ehretia acuminata
Cuscuta gronovii
Cuscuta obtusiflora
Cuscuta reflexa
Cuscuta exaltata
Ipomoea purpurea
Solanum tuberosum
Solanum lycopersicum
Atropa belladonna
Nicotiana sylvestris
Aucuba japonica
Jacobaea vulgaris
Artemisia frigida
Ageratina adenophora
Lactuca sativa
Centaurea diffusa
Scaevola aemula
Trachelium caeruleum
Campanula takesimana
Hanabusaya asiatica
Daucus carota
Anthriscus cerefolium
Anethum graveolens
Panax ginseng
Aralia undulata
Lonicera japonica
Ilex cornuta
Vaccinium macrocarpon
Rhododendron simsii
Pyrola minor
Orthilia secunda
Ericaceae
Arbutus unedo
Actinidia deliciosa
Actinidia chinensis
Actinidiaceae
Camellia crapnelliana
Franklinia alatamaha
Ardisia polysticta
Lysimachia coreana
Primula poissonii
Cornus florida
Davidia involucrata
Portulaca oleracea
Anredera baselloides
Portulacaria afra
Didierea madagascariensis
Pereskia sacharosa
Mollugo verticillata
Silene vulgaris
Spinacia oleracea
Plumbago auriculata
Fagopyrum esculentum
Berberidopsis corallina
Phoradendron serotinum
Ximenia americana
Dillenia indica
Prosopis glandulosa
Acacia ligulata
Haematoxylum brasiletto
Libidibia coriaria
Ceratonia siliqua
Apios americana
Glycine max
Cicer arietinum
Arachis hypogaea
Lupinus albus
Tamarindus indica
Fabaceae
Epirixanthes elongata
Epirixanthes pallida
Salomonia cantoniensis
Polygala arillata
Polygalaceae
Pyrus pyrifolia
Prunus persica
Fragaria virginiana
Fragaria chiloensis
Morus indica
Morus mongolica
Cucumis melo
Corynocarpus laevigata
Castanea mollissima
Salix interior
Passiflora biflora
Parinari campestris
Hevea brasiliensis
Euonymus americanus
Oxalis latifolia
Bulnesia arborea
Raphanus sativus
Arabidopsis thaliana
Carica papaya
Gossypium anomalum
Theobroma cacao
Azadirachta indica
Citrus sinensis
Sapindus mukorossi
Staphylea colchica
Viviania marifolia
Erodium carvifolium
Melianthus villosus
Oenothera argillicola
Oenothera parviflora
Eucalyptus aromaphloia
Penthorum chinense
Sedum sarmentosum
Heuchera sanguinea
Paeonia obovata
Liquidambar formosana
Vitis vinifera
Gunnera manicata
Buxus microphylla
Trochodendron aralioides
Tetracentron sinense
Macadamia integrifolia
Platanus occidentalis
Nelumbo nucifera
Ranunculus macranthus
Megaleranthis saniculifolia
Nandina domestica
Cariudovica palmata
Dioscorea elephantipes
Japanolirion osense
Spirodela polyrhiza
Aponogeton distachyos
Acorus calamus
Piper cenocladum
Ceratophyllum demersum
Liriodendron tulipifera
Magnolia kwangsiensis
Calycanthus floridus
Drimys granadensis
Chloranthus spicatus
Illicium oligandrum
Nuphar advena
Amborella trichopoda
Core eudicots
0.09
substitutions/site

Figure S11. Angiosperm phylogeny inferred in a parsimony analysis of 82 plastid coding regions (ORF-only; see text and Table S3). This is one of the two shortest trees: length = 287,405 steps. Branches that collapse in the strict consensus are indicated with arrows. Bootstrap support values are indicated besides branches. Thick lines indicate 100% bootstrap support; '--' indicates <50% bootstrap support. Eudicot families where mycoheterotrophy has evolved are indicated in blue. The scale bar indicates the inferred number of changes.

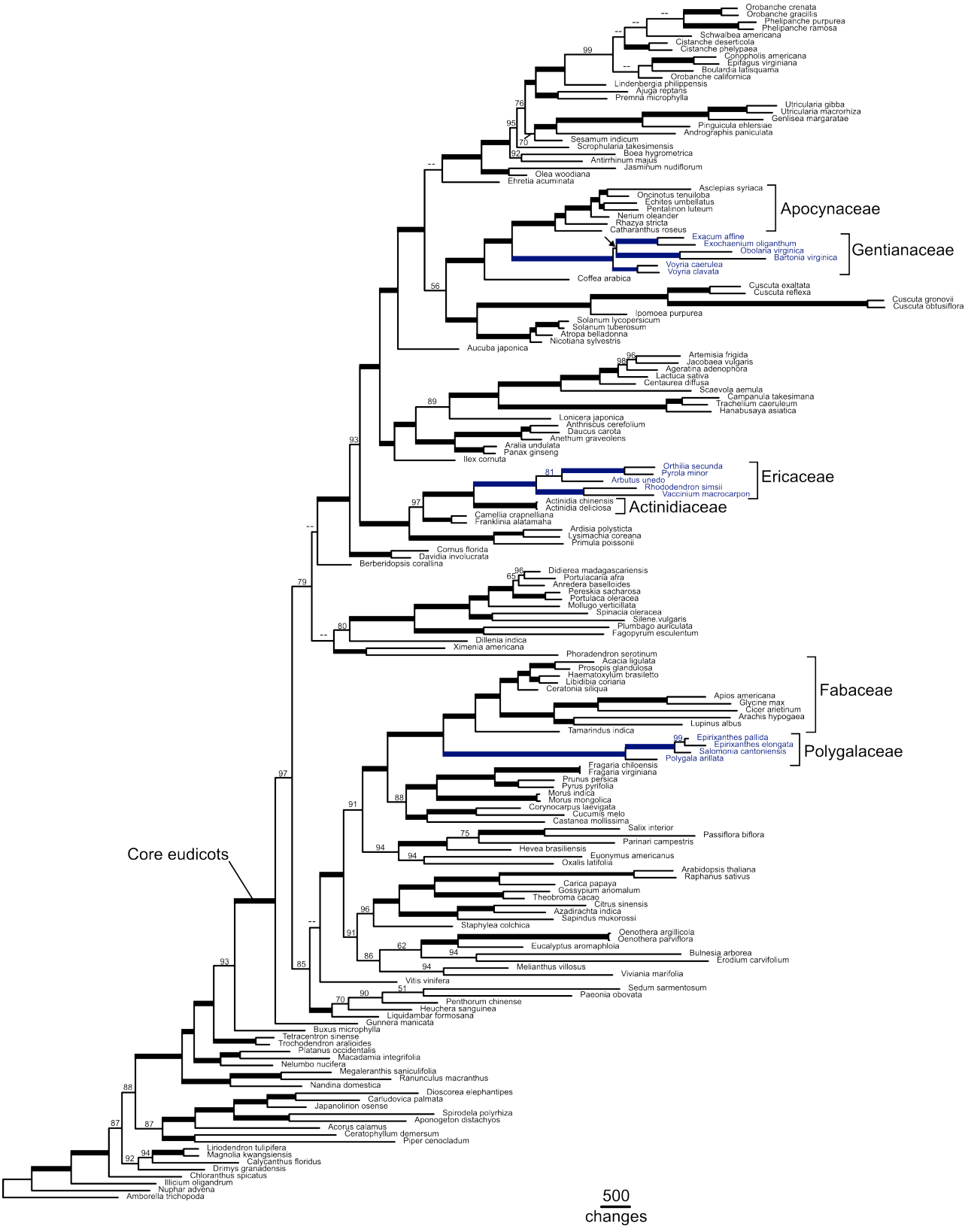
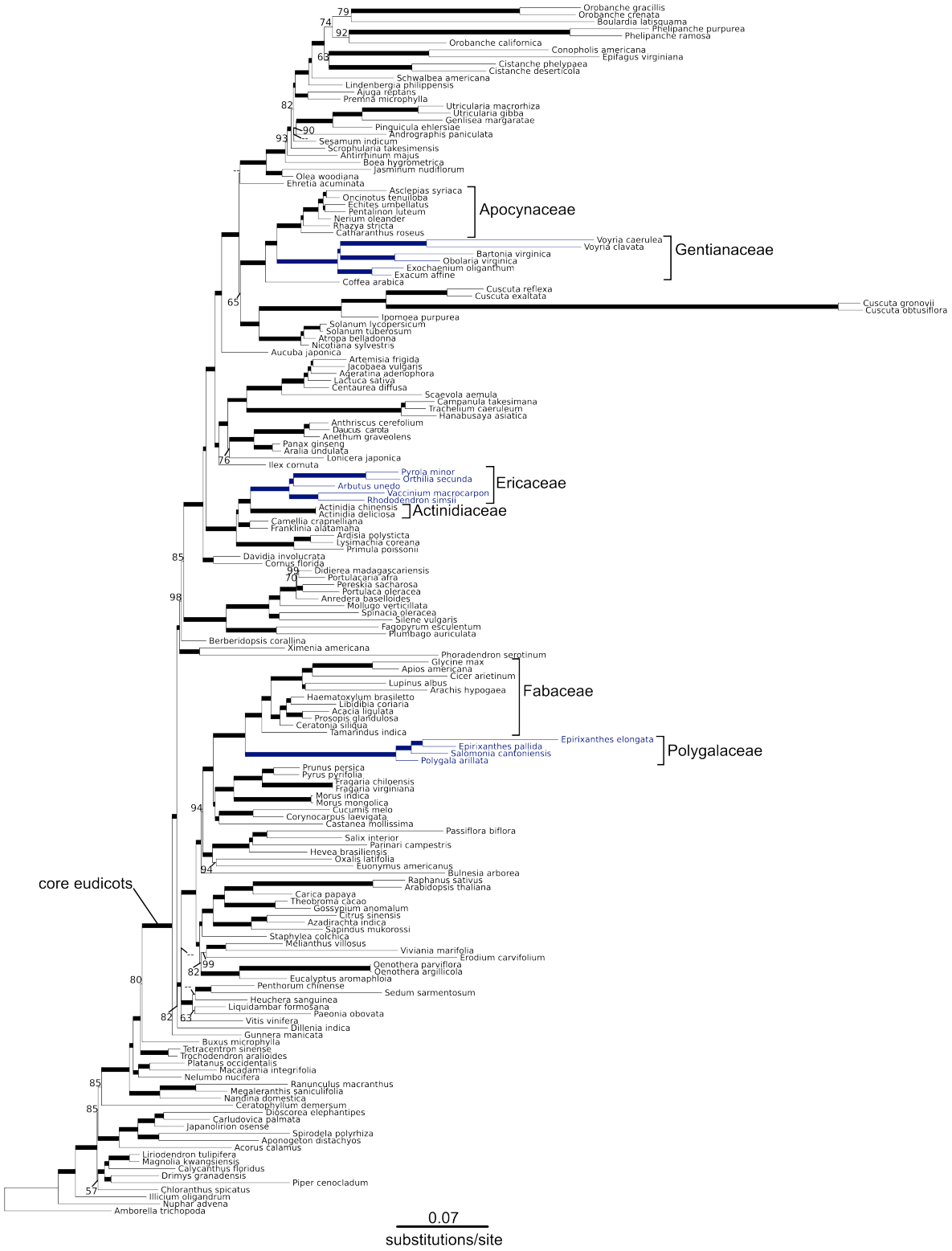

Orobanche crenata
Orobanche gracilis
Phelipanche purpurea
Phelipanche ramosa
Schwalbea americana
Cistanche deserticola
Cistanche phelypaea
Conopholis americana
Epifagus virginiana
Boulardia latisquama
Orobanche californica
Lindenbergia philippensis
Ajuga reptans
Premna microphylla
Utricularia gibba
Utricularia macrorhiza
Genlisea margaratae
Pinguicula ehlersiae
Andrographis paniculata
Sesamum indicum
Scrophularia takesimensis
Boea hygrometrica
Antirrhinum majus
Jasminum nudiflorum
Olea woodiana
Ehretia acuminata
Asclepias syriaca
Oncinotus tenuiloba
Echites umbellatus
Pentalinon luteum
Nerium oleander
Rhazya stricta
Catharanthus roseus
Apocynaceae
Exacum affine
Exochaenium oliganthum
Obolaria virginica
Bartonia virginica
Voyria caerulea
Voyria clavata
Gentianaceae
Coffea arabica
Cuscuta exaltata
Cuscuta reflexa
Cuscuta gronovii
Cuscuta obtusiflora
Ipomoea purpurea
Solanum lycopersicum
Solanum tuberosum
Atropa belladonna
Nicotiana sylvestris
Aucuba japonica
Artemisia frigida
Jacobaea vulgaris
Ageratina adenophora
Lactuca sativa
Centaurea diffusa
Scaevola aemula
Campanula takesimana
Trachelium caeruleum
Hanabusaya asiatica
Lonicera japonica
Anthriscus cerefolium
Daucus carota
Anethum graveolens
Aralia undulata
Panax ginseng
Ilex cornuta
Orthilia secunda
Pyrola minor
Arbutus unedo
Rhododendron simsii
Vaccinium macrocarpon
Ericaceae
Actinidia chinensis
Actinidia deliciosa
Actinidiaceae
Camellia crapnelliana
Franklinia alatamaha
Ardisia polysticta
Lysimachia coreana
Primula poissonii
Cornus florida
Davidia involucrata
Berberidopsis corallina
Didierea madagascariensis
Portulacaria afra
Anredera baselloides
Pereskia sacharosa
Portulaca oleracea
Mollugo verticillata
Spinacia oleracea
Silene vulgaris
Plumbago auriculata
Fagopyrum esculentum
Dillenia indica
Ximenia americana
Phoradendron serotinum
Acacia ligulata
Prosopis glandulosa
Haematoxylum brasiletto
Libidibia coriaria
Ceratonia siliqua
Apios americana
Glycine max
Cicer arietinum
Arachis hypogaea
Lupinus albus
Tamarindus indica
Fabaceae
Epirixanthes pallida
Epirixanthes elongata
Salomonia cantoniensis
Polygala arillata
Polygalaceae
Fragaria chiloensis
Fragaria virginiana
Prunus persica
Pyrus pyrifolia
Morus indica
Morus mongolica
Corynocarpus laevigata
Cucumis melo
Castanea mollissima
Salix interior
Passiflora biflora
Parinari campestris
Hevea brasiliensis
Euonymus americanus
Oxalis latifolia
Arabidopsis thaliana
Raphanus sativus
Carica papaya
Gossypium anomalum
Theobroma cacao
Citrus sinensis
Azadirachta indica
Sapindus mukorossi
Staphylea colchica
Oenothera argillicola
Oenothera parviflora
Eucalyptus aromaphloia
Bulnesia arborea
Erodium carvifolium
Melianthus villosus
Viviania marifolia
Vitis vinifera
Sedum sarmentosum
Paeonia obovata
Penthorum chinense
Heuchera sanguinea
Liquidambar formosana
Gunnera manicata
Buxus microphylla
Tetracentron sinense
Trochodendron aralioides
Platanus occidentalis
Macadamia integrifolia
Nelumbo nucifera
Megaleranthis saniculifolia
Ranunculus macranthus
Nandina domestica
Dioscorea elephantipes
Cariudovica palmata
Japanolirion osense
Spirodela polyrhiza
Aponogeton distachyos
Acorus calamus
Ceratophyllum demersum
Piper cenocladum
Liriodendron tulipifera
Magnolia kwangsiensis
Calycanthus floridus
Drimys granadensis
Chloranthus spicatus
Illicium oligandrum
Nuphar advena
Amborella trichopoda
Core eudicots
500
changes

Figure S12. Angiosperm phylogeny inferred in a likelihood analysis of 82 plastid genes that includes putative pseudogenes (see text and Table S3). Log likelihood score of best tree: -1,565,108.435. Bootstrap support values are indicated beside branches. Thick lines indicate 100% bootstrap support; '--' indicates <50% bootstrap support. Eudicot families where mycoheterotrophy has evolved are indicated in blue. The scale bar indicates estimated substitutions per site.


Orobanche gracillis
Orobanche crenata
Boulardia latisquama
Phelipanche purpurea
Phelipanche ramosa
Orobanche californica
Conopholis americana
Epifagus virginiana
Cistanche phelypaea
Cistanche deserticola
Schwalbea americana
Lindenbergia philippensis
Ajuga reptans
Premna microphylla
Utricularia macrorhiza
Utricularia gibba
Genlisea margaretae
Pinguicula ehlersiae
Andrographis paniculata
Sesamum indicum
Scrophularia takesimensis
Antirrhinum majus
Boea hygrometrica
Jasminum nudiflorum
Olea woodiana
Ehretia acuminata
Asclepias syriaca
Oncinotus tenuiloba
Echites umbellatus
Pentalinon luteum
Nerium oleander
Rhazya stricta
Catharanthus roseus
Apocynaceae
Voyria caerulea
Voyria clavata
Bartonia virginica
Obolaria virginica
Exochaenium oliganthum
Exacum affine
Gentianaceae
Coffea arabica
Cuscuta reflexa
Cuscuta exaltata
Cuscuta gronovii
Cuscuta obtusiflora
Ipomoea purpurea
Solanum lycopersicum
Solanum tuberosum
Atropa belladonna
Nicotiana sylvestris
Aucuba japonica
Artemisia frigida
Jacobaea vulgaris
Ageratina adenophora
Lactuca sativa
Centaurea diffusa
Scaevola aemula
Campanula takesimana
Trachelium caeruleum
Hanabusaya asiatica
Anthriscus cerefolium
Daucus carota
Anethum graveolens
Panax ginseng
Aralia undulata
Lonicera japonica
Ilex cornuta
Pyrola minor
Orthilia secunda
Arbutus unedo
Vaccinium macrocarpon
Rhododendron simsii
Ericaceae
Actinidia chinensis
Actinidia deliciosa
Actinidiaceae
Camellia crapnelliana
Franklinia alatamaha
Ardisia polysticta
Lysimachia coreana
Primula poissonii
Davidia involucrata
Cornus florida
Didierea madagascariensis
Portulacaria afra
Pereskia sacharosa
Portulaca oleracea
Anredera baselloides
Mollugo verticillata
Spinacia oleracea
Silene vulgaris
Fagopyrum esculentum
Plumbago auriculata
Berberidopsis corallina
Ximenia americana
Phoradendron serotinum
Glycine max
Apios americana
Cicer arietinum
Lupinus albus
Arachis hypogaea
Haematoxylum brasiletto
Libidibia coriaria
Acacia ligulata
Prosopis glandulosa
Ceratonia siliqua
Tamarindus indica
Fabaceae
Epirixanthes elongata
Epirixanthes pallida
Salomonia cantoniensis
Polygala arillata
Polygalaceae
Prunus persica
Pyrus pyrifolia
Fragaria chiloensis
Fragaria virginiana
Morus indica
Morus mongolica
Cucumis melo
Corynocarpus laevigata
Castanea mollissima
Passiflora biflora
Salix interior
Parinari campestris
Hevea brasiliensis
Oxalis latifolia
Euonymus americanus
Bulnesia arborea
Raphanus sativus
Arabidopsis thaliana
Carica papaya
Theobroma cacao
Gossypium anomalum
Citrus sinensis
Azadirachta indica
Sapindus mukorossi
Staphylea colchica
Melianthus villosus
Viviania marifolia
Erodium carvifolium
Oenothera parviflora
Oenothera argillicola
Eucalyptus aromaphloia
Penthorum chinense
Sedum sarmentosum
Heuchera sanguinea
Liquidambar formosana
Paeonia obovata
Vitis vinifera
Dillenia indica
Gunnera manicata
Buxus microphyla
Tetracentron sinense
Trochodendron aralioides
Platanus occidentalis
Macadamia integrifolia
Nelumbo nucifera
Ranunculus macranthus
Megaleranthis saniculifolia
Nandina domestica
Ceratophyllum demersum
Dioscorea elephantipes
Carludovica palmata
Japanolirion osense
Spirodela polyrhiza
Aponogeton distachyos
Acorus calamus
Liriodendron tulipifera
Magnolia kwangsiensis
Calycanthus floridus
Drimys granadensis
Piper cenocladum
Chloranthus spicatus
Illicium oligandrum
Nuphar advena
Amborella trichopoda
core eudicots
0.07
substitutions/site