

**PHYLOGENOMICS AND COMPARATIVE PLASTOME
ANALYSIS OF MYCOHETEROTROPHIC PLANTS**

by

Vivienne Ka Yee Lam

B.Sc., The University of British Columbia, 2005

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Botany)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

July 2016

© Vivienne Ka Yee Lam, 2016

Abstract

Plastid genomes (plastomes) of fully mycoheterotrophic plants (which obtain nutrition from fungi and have lost photosynthesis) may exhibit accelerated substitution rates, gene losses and structural rearrangements compared to their more stable photosynthetic relatives. Distantly related lineages provide independent data points to study plastome degradation. I used Sanger sequencing to assess the utility of three nonphotosynthetic plastid genes in phylogenetic inference of seven monocot families that include mycoheterotrophic taxa. I also assembled full plastome genomes for multiple mycoheterotrophic monocots, a heterotrophic conifer (*Parasitaxus*, Podocarpaceae) and autotrophic relatives for comparative analysis. Phylogenomic inferences are robust to different likelihood approaches and often extensive gene loss, are generally congruent with the few-gene analyses, and are insensitive to long branches, in contrast to parsimony. Patterns of gene loss and retention are largely in agreement with hypothesized trajectories, starting with plastid NAD(P)H dehydrogenase, followed by the loss of other photosynthesis-related genes, and ending in gradual loss of transcriptional apparatus and other non-photosynthesis related genes. I observed retention (delayed loss?) of genes encoding subunits of plastid-encoded RNA polymerase (*Parasitaxus* and some species in *Petrosavia*, Petrosaviaceae), plastid ATP synthase (*Petrosavia*, perhaps *Parasitaxus* in modified form) and Rubisco (*Petrosavia*), consistent with secondary non-photosynthetic functions of the latter two complexes. Some group IIA introns appear to be retained despite the loss of the plastid intron maturase, *matK*. Retained open reading frames are generally under strong purifying selection in *Sciaphila* (Triuridaceae). Genome contraction is the major mode of genome rearrangement, with severe reduction seen in some lineages (e.g., *Apteria* in Burmanniaceae is reduced to ~16 kb). Some mycoheterotrophs are nearly or completely colinear with autotrophic lineages (*Geosiris* in Iridaceae, at ~123 kb). Others have multiple minor or major rearrangements, which may be unrelated to the presence or absence of an inverted repeat (IR). Four independent IR losses were observed (in Burmanniaceae, Corsiaceae, Petrosaviaceae and Triuridaceae), an extra IR copy evolved in *Campylosiphon* (Burmanniaceae), and an entire IR re-evolved in *Parasitaxus*. Shifts in IR boundaries were also found in all mycoheterotrophs. Within-taxon comparisons (e.g., in

Corsiaceae and *Petrosavia*) also underline that idiosyncratic evolutionary changes may occur following each loss of photosynthesis.

Preface

A version of Chapter 2 has been published (**Lam, V.K.Y.**, Merckx, V.S.F.T., and Graham, S.W. 2016. A few-gene plastid phylogenetic framework for mycoheterotrophic monocots. *American Journal of Botany* 103: 692-708). I produced most of the sequence data and carried out all analyses. I also led the writing; the study was conceived and co-written with S.W. Graham. Vincent S.F.T Merckx provided plant material and helped with writing and editing.

A version of Chapter 3 has been published (**Lam, V.K.Y.**, Soto Gomez, M., and Graham, S.W. 2016. The highly reduced plastome of mycoheterotrophic *Sciaphila* (Triuridaceae) is colinear with its green relative and is under strong purifying selection. *Genome Biology and Evolution* 7: 2220-2236). I sequenced the new plastome data with Marybel Soto Gomez. I aligned all sequence data and conducted the phylogenetic analysis; Soto Gomez conducted the dN/dS tests. This chapter was co-led with Soto Gomez; we conceived and wrote the manuscript with the help of S.W. Graham.

For Chapter 4, I generated new plastome sequence data for 17 taxa and carried out the analyses. Greg Ross and M. Soto Gomez prepared some of the sequencing libraries. Plant material was provided by Vincent S.F.T. Merckx, Tomohisa Yukawa, Daniel McNair and Stephanie P. Lyon. I conceived and wrote the manuscript with S.W. Graham. Data from several taxa were also used in a publication that is not included here (Mennes, C. B., **Lam, V.K.Y.**, Rudall, P.J., Lyon, S. P., Graham, S.W., Smets E.F., and Merckx, V.S.F. 2015. Ancient Gondwana break-up explains the distribution of the mycoheterotrophic family Corsiaceae (Liliales). *Journal of Biogeography* 42: 1123-1136).

For Chapter 5, I generated new plastome data for five taxa and carried out the analyses. Dean Kelch provided several DNAs. David Tack conducted the C-Sibelia-based variation analysis. I led the writing; the study was conceived and co-written with S.W. Graham.

David Tack and Daisie Huang kindly provided filtering scripts used in Chapters 3-5.

Table of Contents

Abstract	ii
Preface	iv
Table of contents	v
List of tables	ix
List of figures	x
Acknowledgements	xiii
Dedication	xiv
Chapter 1: Introduction	1
1.1 Plastid genome evolution in heterotrophic plants.....	1
1.2 Mycoheterotrophy as an "alternative" plant lifestyle.....	3
1.3 Reconstructing evolutionary relationships of mycoheterotrophs	4
1.4 Overview of the thesis	6
Chapter 2: A few-gene plastid phylogenetic framework for mycoheterotrophic monocots	11
2.1 Summary	11
2.2 Introduction.....	12
2.3 Materials and methods	17
2.3.1 DNA extraction, primer design, amplification and sequencing.....	17
2.3.2 Alignment construction.....	18
2.3.3 Phylogenetic analysis.....	19
2.3.4 Constraint tests of monophyly	20
2.3.5 Characterizing rate elevation in heterotrophic lineages.....	21
2.4 Results	22
2.4.1 Green monocot phylogeny inferred from <i>accD</i> , <i>clpP</i> and <i>matK</i>	22
2.4.2 Phylogenetic placement of mycoheterotrophs in monocot phylogeny.....	22
2.4.3 Relationships in Burmanniaceae.....	25
2.4.4 Rate elevation in mycoheterotrophic monocots.....	25
2.5 Discussion	26
2.5.1 Using plastid genes to place mycoheterotrophic monocots.....	26
2.5.2 Utility and limits of the current approach.....	27

2.5.3	Are Burmanniaceae and Thismiaceae closely related?.....	30
2.5.4	Phylogenetic relationships within Burmanniaceae	31
2.5.5	Retention of plastid genes and genomes in monocot mycoheterotrophs	31
2.5.6	Dealing with contamination in heterotrophic samples.....	33
2.5.7	Suitability of the three genes as DNA barcoding markers.....	34
2.5.8	Conclusions	34
Chapter 3: The highly reduced plastome of mycoheterotrophic <i>Sciaphila</i> (Triuridaceae) is		
colinear with its green relatives and is under strong purifying selection.....48		
3.1	Summary	48
3.2	Introduction	48
3.3	Materials and methods	51
3.3.1	Taxon sampling.....	51
3.3.2	DNA isolation and library preparation	51
3.3.3	<i>De novo</i> contig assembly, plastid gene annotation and plastome reconstruction	52
3.3.4	Data matrix construction and sequence alignment	53
3.3.5	Phylogenetic inference.....	53
3.3.6	Model-based tests of selective regime in plastid genes	55
3.4	Results	56
3.4.1	Full circular plastomes.....	56
3.4.2	The phylogenetic position of <i>Sciaphila</i> (Triuridaceae)	58
3.4.3	Tests of selection.....	59
3.5	Discussion.....	60
3.5.1	Gene loss and retention in <i>Sciaphila</i> (Triuridaceae).....	60
3.5.2	General retention of colinearity despite genome reduction in <i>Sciaphila</i>	62
3.5.3	Model-based tests of selective regime in plastid genes	63
3.5.4	Resolution of the phylogenetic position of Triuridaceae in Pandanales.....	65
Chapter 4: Comparative phylogenomics of mycoheterotrophic monocots77		
4.1	Summary	77
4.2	Introduction	78
4.3	Materials and methods	81
4.3.1	Taxon sampling.....	81

4.3.2	DNA extraction and sequencing	81
4.3.3	<i>De novo</i> assembly, gene annotation and plastome reconstruction.....	81
4.3.4	Matrix assembly and phylogenetic analyses.....	82
4.3.5	Plastome comparison of mycoheterotrophic taxa and autotrophic relatives	83
4.4	Results	84
4.4.1	Plastomes of autotrophic lineages.....	84
4.4.2	Overview of surveyed fully mycoheterotrophic monocot plastomes	85
4.4.3	Gene loss and retention in mycoheterotrophic monocots.....	86
4.4.4	Assessments of colinearity in mycoheterotrophic plastomes	87
4.4.5	Inverted repeat (IR) evolution in mycoheterotrophic plastomes	88
4.4.6	Phylogenetic placement of mycoheterotrophic lineages.....	89
4.5	Discussion.....	90
4.5.1	Phylogenetic placement of mycoheterotrophic monocot lineages.....	90
4.5.2	Structural diversity.....	93
4.5.2.1	Overall colinearity	93
4.5.2.2	IR loss	94
4.5.2.3	IR boundaries and composition	94
4.5.3	Models of gene loss and retention in heterotrophic plants	95
4.5.3.1	Loss of <i>ndh</i> genes.....	96
4.5.3.2	Extensive loss of photosynthesis-related genes in heterotrophic lineages.....	97
4.5.3.3	Retention and loss of plastid-encoded RNA polymerase genes	98
4.5.3.4	Retention and loss of ATP synthase genes	98
4.5.3.5	Housekeeping genes involved in plastid translation.....	99
4.5.3.6	Retention and loss of other housekeeping genes	102
Chapter 5: Comparative plastome analysis of <i>Parasitaxus usta</i> (Podocarpaceae)		117
5.1	Summary	117
5.2	Introduction	117
5.3	Materials and methods	119
5.3.1	Plastome sequencing.....	119
5.3.2	Matrix assembly and phylogenetic analysis	119
5.3.3	Identification of colinear regions in conifer plastomes.....	121

5.3.4	Identification of sequence variation between <i>Parasitaxus</i> populations.....	121
5.4	Results	121
5.4.1	Novel plastomes of autotrophic taxa	121
5.4.2	Characterization of the <i>Parasitaxus</i> plastome	122
5.4.3	Mauve-based assessments of colinearity	123
5.4.4	Population-level plastome variation in <i>Parasitaxus</i>	123
5.4.5	Phylogenetic placement of <i>Parasitaxus</i> within Podocarpaceae.....	124
5.5	Discussion.....	124
5.5.1	Phylogenetic placement of <i>Parasitaxus</i>	124
5.5.2	Evolution of an inverted repeat and plastome evolution in <i>Parasitaxus</i>	126
5.5.3	Gene loss and retention in <i>Parasitaxus</i>	128
5.5.4	Possible retention of ATP synthase CF ₁ region suggests a novel function	129
5.5.5	Within-species plastome variation.....	130
Chapter 6:	Conclusion.....	142
6.1	Phylogenetics/phylogenomics of mycoheterotrophic taxa	142
6.2	Gene loss (and retention) trajectory in mycoheterotrophs.....	145
6.3	Patterns of plastome rearrangement in mycoheterotrophs.....	146
6.4	Future directions	147
Bibliography	149
Appendices	181
Appendix A :	Supplementary tables and figures for Chapter 2	181
Appendix B :	Supplementary tables and figures for Chapter 3	222
Appendix C :	Supplementary tables and figures for Chapter 4	248
Appendix D :	Supplementary tables and figures for Chapter 5	288

List of Tables

Table 1.1 GenBank accessions of heterotrophic plant plastome sequences	8
Table 3.1 Genes retained in <i>Sciaphila</i> relative to <i>Carludovica</i>	68
Table 3.2 Genes retained in published mycoheterotrophic plastomes.....	69
Table 4.1 Properties of mycoheterotrophic monocots and photosynthetic plastomes.....	105
Table 4.2 IR boundaries in fully mycoheterotrophic monocots and close green relatives.....	107
Table 5.1 Genes retained in <i>Parasitaxus</i> relative to green relatives.....	131
Table A.1 Specimen and herbarium information for Chapter 2	181
Table A.2 Data partitioning schemes for Chapter 2	190
Table B.1 Specimen and herbarium information for Chapter 3.....	222
Table B.2 Data partitioning schemes for Chapter 3.....	223
Table B.3 Log likelihoods of branch models for 18 genes retained in <i>Sciaphila</i>	227
Table B.4 Log likelihoods of branch-site models for 18 genes retained in <i>Sciaphila</i>	229
Table C.1 Specimen and herbarium information for Chapter 4.....	248
Table C.2 Data partitioning schemes for Chapter 4.....	251
Table C.3 Status of <i>matK</i> and group IIA introns in full mycoheterotrophs	253
Table D.1 Specimen and herbarium information for Chapter 5	288
Table D.2 Data partitioning schemes for Chapter 5	289
Table D.3 Overview of plastome sizes and characteristics of gymnosperms in Chapter 5	291
Table D.4 Indel variation between three accessions of <i>Parasitaxus usta</i>	292

List of Figures

Figure 1.1 Summary of monocot relationships.....	10
Figure 2.1 Three-gene phylogeny of photosynthetic monocots.....	36
Figure 2.2 Three-gene phylogeny of photosynthetic and mycoheterotrophic monocots.....	38
Figure 2.3 Local placements of monocot mycoheterotrophs based on partitioned ML analyses..	40
Figure 2.4 Placement of green Burmanniaceae in monocot-wide phylogeny	42
Figure 2.5 Placement of individual species of Thismiaceae.....	44
Figure 2.6 Relative substitution rates among green and fully mycoheterotrophic monocots.....	46
Figure 3.1 Circular plastome map of <i>Carludovica palmata</i>	70
Figure 3.2 Circular plastome map of <i>Sciaphila densiflora</i>	72
Figure 3.3 Comparison of linearized plastomes of <i>Carludovica</i> and <i>Sciaphila</i>	74
Figure 3.4 Phylogenetic relationships in Pandanales in overall monocot phylogeny.....	76
Figure 4.1 Linearized plastomes of mycoheterotroph monocots and green representatives	109
Figure 4.2 A “heat-map” showing gene loss and retention across selected land plants	111
Figure 4.3 Relationship between retained genes and plastome size of mycoheterotrophs.....	113
Figure 4.4 Colinearity between plastomes of mycoheterotrophic monocots and relatives	114
Figure 4.5 ML-based phylogenomic analysis of mycoheterotrophic monocots.....	116
Figure 5.1 Circular plastome map of <i>Manoao colensoi</i>	132
Figure 5.2 Circular plastome map of <i>Lepidothamnus laxifolius</i>	134
Figure 5.3 Circular plastome map of <i>Parasitaxus usta</i>	136
Figure 5.4 Mauve-base alignments of gymnosperm autotrophs and <i>Parasitaxus</i>	138
Figure 5.5 ML-based phylogenomic analysis of gymnosperms and <i>Parasitaxus</i>	140
Figure A.1 Primer maps for <i>accD</i> , <i>clpP</i> and <i>matK</i>	192
Figure A.2 Placement of putative contaminant <i>matK</i> sequence of <i>Thismia aseroe</i>	194
Figure A.3 Placement of putative contaminant <i>matK</i> sequence of <i>Geomitra clavigera</i>	196
Figure A.4 Parsimony analysis of photosynthetic and mycoheterotrophic monocots.....	198
Figure A.5 Unpartitioned ML analysis of photosynthetic and mycoheterotrophic monocots.....	200
Figure A.6 Phylogenetic placement of <i>Geosiris</i> based on partitioned ML analysis	202
Figure A.7 Partitioned ML analysis of mycoheterotrophic Orchidaceae	204
Figure A.8 Partitioned ML analysis of mycoheterotrophic Burmanniaceae	206

Figure A.9 Partitioned ML analysis of mycoheterotrophic Corsiaceae	208
Figure A.10 Partitioned ML analysis of mycoheterotrophic Petrosaviaceae	210
Figure A.11 Partitioned ML analysis of of mycoheterotrophic Triuridaceae	212
Figure A.12 Unpartitioned ML analysis of mycoheterotrophic Triuridaceae	214
Figure A.13 Partitioned ML analysis of mycoheterotrophic Thismiaceae	216
Figure A.14 Unpartitioned ML analysis of mycoheterotrophic Thismiaceae	218
Figure A.15 Relative substitution rates among green and mycoheterotrophic monocots	220
Figure B.1 Unpartitioned ML analysis of <i>Sciaphila</i>	230
Figure B.2 ML analysis of <i>Sciaphila</i> partitioned by ‘codon’ partitioning scheme	232
Figure B.3 ML analysis of <i>Sciaphila</i> partitioned by ‘gene x codon’	234
Figure B.4 Unpartitioned ML analysis of <i>Sciaphila</i> using an amino acid model.....	236
Figure B.5 ML analysis of <i>Sciaphila</i> using an amino acid model, partitioned by genes.....	238
Figure B.6 Unpartitioned ML analysis of <i>Sciaphila</i> using a codon-based model	240
Figure B.7 Parsimony analyses of <i>Sciaphila</i>	242
Figure B.8 Unpartitioned ML analysis of <i>Sciaphila</i> (re-aligned).....	244
Figure B.9 ML analysis of <i>Sciaphila</i> partitioned by ‘gene x codon’ (re-aligned).....	246
Figure C.1 Circular plastome map of <i>Campynema lineare</i>	256
Figure C.2 Circular plastome map of <i>Iris missouriensis</i>	258
Figure C.3 Circular plastome map of <i>Japonolirion osense</i>	260
Figure C.4 Circular plastome map of <i>Lilium superbum</i>	262
Figure C.5 Mauve-based alignment of <i>Campylosiphon congestus</i>	264
Figure C.6 Mauve-based alignment of <i>Burmannia itoana</i>	266
Figure C.7 Mauve-based alignment of <i>Gymnosiphon longistylus</i>	268
Figure C.8 Mauve-based alignment of <i>Apteria aphylla</i>	270
Figure C.9 Mauve-based alignment of <i>Corsia cf boridiensis</i>	272
Figure C.10 Mauve-based alignment of <i>Arachnitis uniflora</i>	274
Figure C.11 Mauve-based alignment of <i>Geosiris aphylla</i>	276
Figure C.12 Mauve-based alignment of <i>Petrosavia</i> spp.....	278
Figure C.13 Mauve-based alignment of <i>Burmannia capitata</i>	280
Figure C.14 Mauve-based alignment of <i>Sciaphila densiflora</i>	282
Figure C.15 Unpartitioned ML analysis of photosynthetic and mycoheterotrophic monocots...	284

Figure C.16 Parsimony analysis of photosynthetic and mycoheterotrophic monocots	286
Figure D.1 Mauve-based alignments of <i>Ginkgo</i> with Podocarpaceae	294
Figure D.2 BRIG-based diagram of nucleotide substitutions and indels in <i>Parasitaxus</i>	296
Figure D.3 Gymnosperm phylogeny inferred from unpartition ML analysis	298
Figure D.4 Gymnosperm phylogeny inferred from parsimony analysis	300

Acknowledgements

I thank Dr. Sean Graham for his leap of faith in taking on a biochemistry undergraduate who knew barely anything about plants. Under his supervision, he immensely broadened my understanding of genome evolution, and taught me, by example, the need and dedication required for clear and concise scientific writing. I would also like to thank my committee members Dr. Jeannette Whitton and Dr. Patrick Keeling. Their advice has been instrumental in the completion of my thesis, and their questions have challenged me to explore my findings in the context of other branches of biology.

Time spent with fellow Graham lab members, past and present, have provided me with many happy memories. I thank Hardeep Rai and Ying Chang for their mentorship. I thank William Iles for his help with dating analyses and interesting conversations about evolution. I especially thank Marybel Soto Gomez, Gregory Ross, and Isabel Marques for making our foray into next generation sequencing ‘fun’ (late nights and all!).

I thank my officemate, David Tack, not only for his scripting prowess but also for his friendship and fun hours spent in the office. I also thank Jackie Dee for a wonderful friendship that extends beyond our love of science.

Many thanks to those who shared data and materials with me: Vincent Merckx, Tomohisa Yukawa, Ray Neyland, Daniel McNair, Adrien Wulff, Pete and Michelle Hollingsworth, Constantijn Mennes, Stephanie Lyon, the Gymnosperm AToL (especially Sarah Mathews and Linda Raubeson), and the Monocot AtoL (especially Craig Barrett, Tom Givnish, and Jim Leebens-Mack). I would also like to thank Tim Brodribb who provided guidance and advice, and Mark Chase for his help.

I appreciate the funding UBC has provided over the years in the form of scholarships and teaching assistantships.

Lastly, I am grateful for the love and support from my parents. I especially thank my mother, who always encouraged me to pursue my academic goals. My brother, Michael Lam, ensured that I survived my thesis by supplying a consistent source of nourishment. My husband, Vic Ying-udomrat, was always there to offer a listening ear, and for that I am always thankful.

To my parents, who encouraged me to be curious about the world around me,
and to read books of all kinds.

Chapter 1: Introduction

1.1 Plastid genome evolution in heterotrophic plants

The land plants (Embryophyta) are remarkable organisms capable of converting light energy into chemical energy via the autotrophic process known as photosynthesis. Photosynthesis occurs in specialized membrane-bound plastids called chloroplasts; other specialized plastids include leucoplasts (involved in starch storage) and chromoplasts (production and storage of pigments). Plastids originated in eukaryotes from a primary endosymbiosis event that involved the uptake of a free-living cyanobacterium, and they usually retain a streamlined bacterial genome (e.g., Margulis 1981; Martin and Hermann 1998; Martin et al. 1998, Timmis et al. 2004). The plastid genome (plastome) of photosynthetic land plants is a double-stranded circular DNA molecule that is usually highly conserved in terms of size, gene content and organization (e.g., Palmer 1985; Bock 2007; Wicke et al. 2011, see Bendich 2004 for a review on alternative linear and branched plastome structures). It is typically 120-160 kb in length, and codes for 100-120 unique genes in a quadripartite arrangement consisting of a large single copy (LSC) region and small single copy (SSC) region, separated by two inverted repeat (IR) regions (the latter comprise two sets of identical genes). The boundaries between these major regions are typically well conserved in related land-plant lineages, although they tend to drift as evolutionary distances increase among taxa (e.g., Goulding et al. 1996; Wang et al. 2008; Zhu et al. 2015). In general, colinearity (the relative arrangement of genes) along the plastid chromosome is also highly conserved across land-plant taxa (e.g., Palmer and Thompson 1982; Palmer and Stein 1986), with exceptions in a few green and non-green lineages (e.g., Wolfe et al. 1992; Funk et al. 2007; Cai et al. 2008; Delannoy et al. 2011; Guisinger et al. 2011).

Some plants have lost the ability to photosynthesize, and instead obtain energy-containing compounds directly from other plants (holoparasites) or from fungal partners (full mycoheterotrophs); other plants continue to photosynthesize, but obtain water and minerals from other plants (hemiparasites), or are partially heterotrophic for some or all of their life cycles (initial and partial mycoheterotrophs, respectively) (e.g., Kuijt 1969; Leake 1994). Because of the opportunities that they present for understanding plastid genome function in an altered nutritional mode, a disproportionately high number of full mycoheterotrophs have had their plastid genomes sequenced: 12 of ~840 land-plant plastomes sequenced to date (Table 1.1;

number of autotrophs based on NCBI, <http://www.ncbi.nlm.nih.gov/genome/browse/> -- accessed April 2016), representing a ~14-fold overrepresentation compared to the number of fully mycoheterotrophic vs. autotrophic land plants (~514 species of full mycoheterotrophs, Merckx et al. 2013, ~500,000 autotrophs, Ruhfel et al. 2014). All partial mycoheterotrophs sampled to date come from a single orchid genus, *Corallorhiza* (Table 1.1), which also includes full mycoheterotrophs. Most fully mycoheterotrophic species in this genus appear to derive from a homologous loss of photosynthesis (Barrett et al. 2014). Examining multiple mycoheterotrophs from the same lineage is useful for investigating divergent pathways of change following a single loss of photosynthesis. However, the plastid genomes of most mycoheterotrophic lineages that represent independent losses of photosynthesis (convergent transitions to full mycoheterotrophy), or other independent origins of partial mycoheterotrophy, have yet to be explored. The study of plastid genome evolution in multiple previously unsampled lineages of mycoheterotrophs is a major topic of my thesis.

There may be a consistent pattern of gene loss and retention, such that the amount of change is correlated with the degree of dependency on heterotrophy (e.g., Krause 2008, 2011, 2012; Krause and Scharff 2014). The switch to full mycoheterotrophy was hypothesized to eventually follow an irreversible trajectory by Barrett and Davis (2012) and Barrett et al. (2014), as photosynthesis genes are not expected to be regained, once lost. These authors hypothesized that plastid-encoded genes are pseudogenized and then lost in stages, beginning with NAD(P)H genes (likely before full photosynthesis loss), followed by other photosynthesis-related genes, then plastid-encoded RNA polymerase genes (PEP) and ATP synthase genes (their 2012 and 2014 papers differ in how and when genes involved in the latter two complexes are lost), and ending with gradual degradation of translational apparatus and other housekeeping genes, and perhaps eventually in plastome loss. However, the order of gene loss in the intermediate stages of heterotrophy is uncertain, as some taxa exhibit gene losses or retentions that are contrary to the proposed trajectory (e.g., *Petrosavia stellaris* has retained *rbcL* despite loss of other photosynthesis genes, Logacheva et al. 2011; several ‘holoparasitic’ species of *Cuscuta* have lost PEP genes despite retention of photosynthesis genes, Funk et al. 2007; McNeal et al. 2007). A high degree of plastome reduction may be the ultimate evolutionary fate of fully heterotrophic taxa (e.g., Wolfe et al. 1992; Delannoy et al. 2011; Wicke et al. 2013; Schelkunov et al. 2015; Bellot and Renner 2016; Gruzdev et al. 2016 and Naumann et al. 2016), which may eventually

experience complete plastome loss (*Rafflesia lagascae*, Molina et al. 2014). The smallest minimal plastomes encode genes in the last category (housekeeping genes, including those involved in the translation apparatus), and usually comprise core sets of essential non-photosynthesis genes that have a range of functions, including plastid-based protein synthesis (i.e., small and large ribosomal protein subunit, tRNA and rDNA genes), fatty-acid formation and tetrapyrrole synthesis (e.g., Bungard 2004; Barbrook et al. 2006; Wicke et al. 2011; Barrett et al. 2014; Schelkunov et al. 2015; Gruzdev et al. 2016; Naumann et al. 2016), although see Bellot and Renner (2016) for a case involving extreme reduction in endoparasitic plants, which appears to involve loss of most of these genes too. Comparative plastome analyses from mycoheterotrophic taxa can potentially provide additional data for clarifying existing models of gene loss and retention (e.g., Barrett and Davis 2012; Barrett et al. 2014), and improve our understanding of the mechanisms and consequences of plastome evolution following the transition to heterotrophy.

1.2 Mycoheterotrophy as an “alternative” plant lifestyle

Mycoheterotrophic plants usually exploit mycorrhizae, and can be thought of as epiparasites of the green-plant partners of mycorrhizal fungi (Bidartondo 2005). Approximately 80-90% of land plants rely on mycorrhizal symbioses (e.g., Wang and Qiu 2006; Merckx 2013), mutualistic associations in which photosynthetic plants provide sugars to fungal partners and in return receive increased intake of mineral nutrients and water. Mycoheterotrophic plants instead draw nutrients, water and carbon from mycorrhizal fungi (and thus indirectly parasitize associated autotrophic plants), or in a few cases from saprophytic fungi (Leake 1994). This unique plant trophic strategy is distinct from saprophytism (which mycoheterotrophs were assumed to be in some of the older literature, and are even mislabelled as such in some modern field guides and floras, e.g., Hitchcock and Cronquist 1973; Pojar et al. 1994; Douglas et al. 2001), because mycoheterotrophs cannot degrade organic material (Leake 2005). Mycoheterotrophy is also distinct from plant parasitism, in which modified roots (“haustoria”) of the parasitic plant form direct physiological connections with host plants (Heide-Jørgensen 2008). Instead, mycoheterotrophs directly obtain nutrition from fungal intermediates that they attract and house in modified roots (Imhof 2010), usually members of arbuscular mycorrhizal fungi (*Glomus* Group A, Glomeraceae; or less commonly in Acaulosporaceae or Gigasporaceae), or

ectomycorrhizal fungi (e.g., Bidartondo and Bruns 2001; Bidartondo et al. 2002; Franke et al. 2006; Merckx et al. 2012; Waterman et al. 2013). Full mycoheterotrophs are usually non-green/achlorophyllous (although chlorophyll is maintained in a few cases, e.g., *Corallorhiza* (Cummings and Welschmeyer 1998; Barrett et al. 2014). Partial mycoheterotrophs also retain chlorophyll and continue to photosynthesize. Dual-isotope profiling has confirmed the trophic status of some partial mycoheterotrophs (e.g., Gebauer and Meyer 2003; Trudell et al. 2003; Hynson et al. 2009; Merckx et al. 2010; Motomura et al. 2010; Bolin et al. 2015). The trophic status of many others is unknown, although non-green plants without plant-to-plant connections are typically assumed to be full mycoheterotrophs (see discussion in Merckx et al. 2006).

Monocots have particularly rich examples of mycoheterotrophy, and comprise 91% of extant mycoheterotrophic species. This richness may be attributed to their herbaceous habit, fibrous root systems, and expansive subterranean organs suitable for facilitating mycorrhizal colonization (Imhof 2010). Seven monocot families collectively comprise ~38 of the estimated 48 independent origins of mycoheterotrophy in land plants (in Burmanniaceae, Corsiaceae, Iridaceae, Orchidaceae, Petrosaviaceae, Thismiaceae and Triuridaceae; Merckx et al. 2013; Fig. 1.1), and are the subject of the majority of my thesis. Iridaceae have a single mycoheterotrophic species *Geosiris aphylla*, whereas several other families are entirely composed of full mycoheterotrophs (i.e., Corsiaceae, Thismiaceae and Triuridaceae). Petrosaviaceae, Burmanniaceae and Orchidaceae have both autotrophic and fully mycoheterotrophic members; the latter two families also include partially mycoheterotrophic members, and have experienced multiple independent origins of mycoheterotrophy (probably at least eight origins in Burmanniaceae and ~25 in Orchidaceae; Merckx et al. 2013). Mycoheterotrophy also evolved multiple times in the eudicots, representing approximately seven independent origins (Ericaceae, Gentianaceae, and Polygalaceae; e.g. Merckx et al. 2013). It evolved once each in the liverworts (in *Aneura mirabilis*, Aneuraceae; Wickett et al. 2008) and the conifers (in *Parasitaxus usta*, Podocarpaceae; Sinclair et al. 2002; Biffin et al. 2011). *Parasitaxus* is the subject of one thesis chapter here (Chapter 5; see below).

1.3 Reconstructing evolutionary relationships of mycoheterotrophs

The higher-order classification of mycoheterotrophs based on morphological data can be challenging, as mycoheterotrophy (heterotrophy in general) may lead to convergent reduction in

vegetative features and extreme modifications of reproductive structures. The latter likely evolved to conserve carbon sources for reproduction (e.g., Leake 1994). Unusual floral morphologies of some taxa (e.g., *Corsiopsis*, Corsiaceae, Zhang et al. 1999; Triuridaceae, Rudall and Bateman 2006; Rudall et al. 2016; Thismiaceae, Franke 2004; Mar and Saunders 2015) may facilitate reproduction in the understory (Leake 1994; Merckx et al. 2013). These unusual morphologies led to a distorted understanding of their taxonomic placement. For example, Cronquist (1981) placed Burmanniaceae, Corsiaceae, Geosiridaceae (*Geosiris*), and Orchidaceae in the order Orchidales, and Dahlgren et al. (1985), and more recently, Takhtajan (2009) placed Burmanniaceae, Corsiaceae and Thismiaceae in a single order, Burmanniales. These taxonomic arrangements are now understood to be unnatural (see below).

Modern plant classification still relies heavily on a few widely sampled plastid genes (such as *rbcL* and *atpB*, Chase et al. 1993; Soltis et al. 2000; APG 1998, 2003, 2009), often rendering the inclusion of full mycoheterotrophs in these well-established frameworks challenging, as photosynthetic genes are typically lost in them. In addition, retained genes were thought to be problematic because of elevated substitution rates, potentially leading to phylogenetic misinference due to long-branch attraction, (e.g., Felsenstein 1978; Henny and Penny 1989). However, this problem is not limited to the plastome, as heterotrophic plants can exhibit rate acceleration in all three plant genomes (Lemaire et al. 2011; Bronham et al. 2013). Some early placements of mycoheterotrophic taxa were based on analyses that included plastid genes, sometimes in combination with nuclear and mitochondrial genes (e.g., Davis et al. 2004; Petrosaviaceae, Tamura et al. 2004; Givnish et al. 2005; Graham et al. 2006, see also the review in Cameron et al. 2003; *Geosiris*, Reeves et al. 2001; Goldblatt et al. 2008), but others involving the most reduced mycoheterotrophs were based on nuclear and mitochondrial data (e.g. Burmanniaceae, Merckx et al. 2006, 2008; Corsiaceae, Neyland and Hennigan 2003; Triuridaceae, Mennes et al. 2013). Next-generation sequencing technologies facilitate relatively straightforward assembly of full plastome sequences (e.g., Table 1.1), suggesting that it would be worth exploring their utility in understanding phylogenetic placement of mycoheterotrophs, despite expectations of rate elevation and gene loss in them (e.g., Chase et al. 1993; dePamphilis et al. 1997). In addition, the phylogenetic placements of several mycoheterotrophic taxa have not previously been studied using plastid data. The increasing availability of mycoheterotroph plastomes, especially from families that experienced multiple independent origins of

mycoheterotrophy, should also help refine our understanding of whether there are regularities in gene loss and retention in the plastid genome, and the general effects of this transition on plastid genome structure.

1.4 Overview of the thesis

In Chapter 2, I revisit the utility of few-gene plastid surveys for inferring the phylogenetic placement of mycoheterotrophic monocots. I survey three plastid genes (*accD*, *clpP* and *matK*) that are typically retained in heterotrophic plants, surveying them across all major autotrophic monocot orders and all seven monocot families with mycoheterotrophic members. I construct a monocot-wide framework to infer the phylogenetic relationships of mycoheterotrophic taxa, and explore the sensitivity of phylogenetic analyses to including vs. excluding other mycoheterotrophs. I test the monophyly of Burmanniaceae *sensu lato* (APG 2003 and 2009) including Thismiaceae, and consider hypotheses of single vs. multiple origins of mycoheterotrophy in Burmanniaceae using plastid evidence. I also characterize differences in relative substitution rates test across autotrophic and mycoheterotrophic monocots.

In Chapter 3 I use plastid genome data to infer the local placement of a single mycoheterotrophic taxon *Sciaphila densiflora* (representing the fully mycoheterotrophic family Triuridaceae), resolving the uncertain local placement of Triuridaceae in Pandanales (Chase et al. 2000; Rudall and Bateman 2006; Mennes et al. 2013). Specifically, I incorporate newly produced plastome sequences for *Sciaphila* and several autotrophic members of Pandanales, updating a recent monocot-wide plastome matrix (Ross et al. 2016). For retained genes in *Sciaphila* that have uninterrupted reading frames (putatively functional proteins), I also conduct selection tests (dN/dS) to assess whether they are evolving under a different selective regime than autotrophic outgroup lineages.

In Chapter 4 I present phylogenetic inferences based on newly generated full plastome sequences for ten mycoheterotrophic taxa from five families (including Triuridaceae presented in more detail in Chapter 3) and five autotrophic relatives. I summarize patterns of gene loss and retention in the mycoheterotrophic species, in the context of hypotheses by Barrett and Davis (2012) and Barrett et al. (2014) for plastid genome degradation. I also characterize structural rearrangements in these taxa in comparison to the closest green relatives of each mycoheterotrophic lineage, including changes in plastid inverted repeat (IR) boundaries.

In Chapter 5 I generate three full plastome sequences from two populations of *Parasitaxus usta* (Podocarpaceae) a heterotrophic conifer, and two podocarp relatives, and use these to infer the local placement of *Parasitaxus* in Podocarpaceae, to quantify among-population variation in the *Parasitaxus* plastome, and to characterize plastid genome arrangements and models of gene loss and retention in this lineage.

Table 1.1 GenBank accessions of heterotrophic plant plastome sequences. Mycoheterotrophic taxa are highlighted in **bold**. ‘Status’ refers to nutritional mode. Abbreviations: MH=mycoheterotroph, P=parasite

Family	Species	Status	Study	Accession no.	
Aneuraceae	<i>Aneura mirabilis</i>		Full MH	Wickett et al. 2008	NC_010359.1
Apodanthaceae	<i>Pilostyles aethiopica</i>		Holo-P	Bellot and Renner 2016	KT981955.1
	<i>Pilostyles hamiltonii</i>		Holo-P	Bellot and Renner 2016	KT981956.1
Convolvulaceae	<i>Cuscuta exaltata</i>		Hemi-P	McNeal et al. 2007	NC_009963.1
	<i>Cuscuta gronovii</i>		Hemi-P	Funk et al. 2007	NC_009765.1
	<i>Cuscuta obtusiflora</i>		Hemi-P	McNeal et al. 2007	NC_009949.1
	<i>Cuscuta reflexa</i>		Hemi-P	Funk et al. 2007	NC_009766.1
Ericaceae	<i>Monotropa hypopitys</i>		Full MH	Gruzdev et al. 2016	KU640958.1
Hydnoraceae	<i>Hydnora visseri</i>		Holo-P	Naumann et al. 2016	KT970098
Orchidaceae	<i>Corallorhiza bulbosa</i>		Partial MH	Barrett et al. 2014	NC_025659.1
	<i>Corallorhiza macrantha</i>		Partial MH	Barrett et al. 2014	NC_025660.1
	<i>Corallorhiza mertensiana</i>		Full MH	Barrett et al. 2014	NC_025661.1
	<i>Corallorhiza maculata</i> var. <i>maculata</i>		Full MH	Barrett et al. 2014	KM390014
	<i>Corallorhiza maculata</i> var. <i>occidentalis</i>		Full MH	Barrett et al. 2014	KM390016.1
	<i>Corallorhiza maculata</i> var. <i>mexicana</i>		Partial MH	Barrett et al. 2014	KM390015.1
	<i>Corallorhiza odontorhiza</i>		Partial MH	Barrett et al. 2014	NC_025664.1
	<i>Corallorhiza striata</i>		Full MH	Barrett et al. 2014	JX087681.1
	<i>Corallorhiza trifida</i>		Partial MH	Barrett et al. 2014	NC_025662.1
	<i>Corallorhiza wisteriana</i>		Partial MH	Barrett et al. 2014	NC_025663.1
	<i>Epipogium aphyllum</i>		Full MH	Schelkunov et al. 2015	NC_026449.1
	<i>Epipogium roseum</i>		Full MH	Schelkunov et al. 2015	NC_026448.1
	<i>Neottia nidus-avis</i>		Full MH	Logacheva et al. 2011	NC_016471.1

Family	Species	Status	Study	Accession no.	
	<i>Rhizanthella gardneri</i>		Full MH	Delannoy et al. 2011	NC_014874.1
Orobanchaceae	<i>Boulardia latisquama</i>		Holo-P	Wicke et al. 2013	NC_025641.1
	<i>Cistanche diserticola</i>		Holo-P	Li et al. 2013	NC_021111.1
	<i>Cistanche phelypaea</i>		Holo-P	Wicke et al. 2013	NC_025642.1
	<i>Conopholis americana</i>		Holo-P	Wicke et al. 2013	NC_023131.1
	<i>Epifagus virginiana</i>		Holo-P	Wolfe et al. 1992	NC_001568.1
	<i>Lathraea squamaria</i>		Holo-P	Samigullin et al. 2016	NC_027838.1
	<i>Orobanche austrohispanica</i>		Holo-P	Cusimano and Wicke 2015	KT387721
	<i>Orobanche (Myzorrhiza) californica</i>		Holo-P	Wicke et al. 2013	NC_025651.1
	<i>Orobanche densiflora</i>		Holo-P	Cusimano and Wicke 2015	KT387723
	<i>Orobanche crenata</i>		Holo-P	Wicke et al. 2013	NC_024845.1
	<i>Orobanche cumana</i>		Holo-P	Cusimano and Wicke 2015	KT387722
	<i>Orobanche gracilis</i>		Holo-P	Wicke et al. 2013	NC_023464.1
	<i>Orobanche pancicii</i>		Holo-P	Cusimano and Wicke 2015	KT387724
	<i>Orobanche (Phelipanche) purpurea</i>		Holo-P	Wicke et al. 2013	NC_023132.1
	<i>Orobanche (Phelipanche) ramosa</i>		Holo-P	Wicke et al. 2013	NC_023465.1
	<i>Orobanche rapum-genistae</i>		Holo-P	Cusimano and Wicke 2015	KT387725
	<i>Schwalbea americana</i>		Hemi-P	Wicke et al. 2013	NC_023115.1
Petrosaviaceae	<i>Petrosavia stellaris</i>		Full MH	Logacheva et al. 2014	NC_023356.1
Santalaceae	<i>Osyris alba</i>		Hemi-P	Petersen et al. 2015	NC_027960.1
	<i>Viscum album</i>		Hemi-P	Petersen et al. 2015	NC_028012.1
	<i>Viscum crassulae</i>		Hemi-P	Petersen et al. 2015	NC_027959.1
	<i>Viscum minimum</i>		Hemi-P	Petersen et al. 2015	NC_027829.1
Triuridaceae	<i>Sciaphila densiflora</i>		Full MH	Lam et al. 2015 (Chapter 4)	NC_027659.1

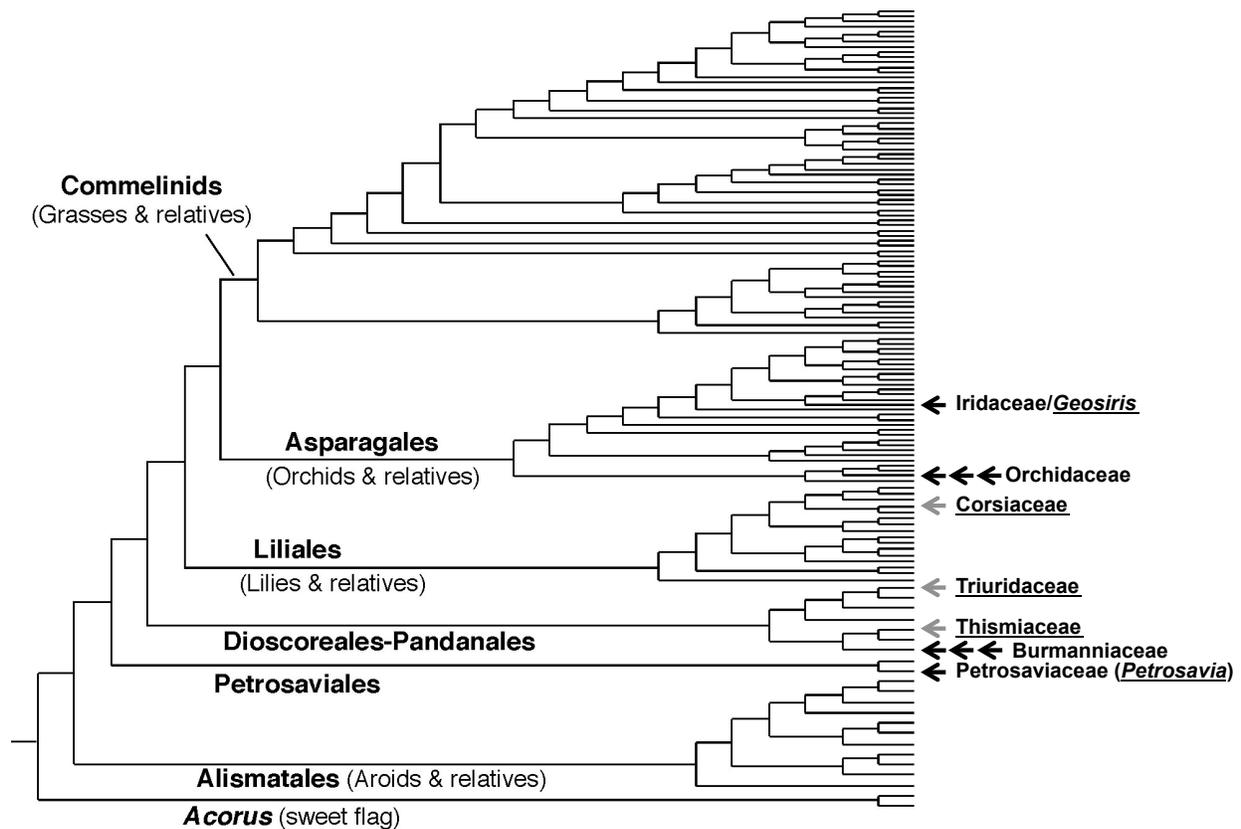


Figure 1.1 Summary of monocot relationships from parsimony analysis of 17 plastid genes, adapted from Graham et al. (2006). Possible placements of mycoheterotrophic taxa (e.g., Cameron et al. 2003; Cameron 2009; Freudenstein et al. 2004; Goldblatt et al. 2008; Mennes et al. 2013, 2015; Merckx et al. 2006, 2008; Rudall and Bateman 2006) belonging to seven monocot families are indicated with arrows; families with multiple independent origins of mycoheterotrophy are indicated with three arrows. Grey arrows indicate families with less certain placements at the ordinal level. Fully mycoheterotrophic families/taxa are underlined.

Chapter 2: A few-gene plastid phylogenetic framework for mycoheterotrophic monocots¹

2.1 Summary

Few-gene studies with broad taxon sampling have provided major insights into phylogeny and underpin plant classification. However, they have typically excluded heterotrophic plants because of loss, pseudogenization, or rapid evolution of plastid genes. Here I performed a phylogenetic survey of three commonly retained plastid genes to assess their utility in placing mycoheterotrophs. I surveyed *accD*, *clpP* and *matK* for 34 taxa from seven monocot families that include full mycoheterotrophs and from a broad sampling of photosynthetic taxa. After screening for weak contaminants, I conducted phylogenetic analyses and characterized among-lineage rate variation. Likelihood analyses strongly supported local placements of fully mycoheterotrophic taxa for Corsiaceae, Iridaceae, Orchidaceae and Petrosaviaceae, in positions consistent with other studies. Depression of likelihood bootstrap support values near mycoheterotrophic clades was alleviated when each mycoheterotrophic family was considered separately. Triuridaceae (*Sciaphila*) monophyly was recovered in a partitioned likelihood analysis, and the family then placed as sister to Cyclanthaceae-Pandanaceae. Burmanniaceae placed in Dioscoreales with weak to strong support depending on analysis details, and I inferred a plastid-based phylogeny for the family. Thismiaceae species may retain a plastid genome, based on *accD* retention. The inferred position of Thismiaceae is unstable, but was close to Taccaceae (Dioscoreales) in some analyses. Long branches/elevated substitution rates, missing genes, and occasional contaminants are challenges for plastid-based phylogenetic inference with full mycoheterotrophs. However, most mycoheterotrophs can be readily integrated into the broad picture of plant phylogeny using several plastid genes and broad taxonomic sampling.

¹This chapter has been published as ‘Lam, V.K.Y., Merckx, V.S.F.T and S.W. Graham. 2016. A few-gene plastid phylogenetic framework for mycoheterotrophic monocots. *American Journal of Botany* 103: 692-708.’

2.2 Introduction

Mycoheterotrophic plants acquire carbon, water, and essential nutrients from fungal partners, typically those involved in mycorrhizal symbioses with green plants (Leake 2004, 2005; Merckx et al. 2009a). As in holoparasitic plants, some mycoheterotrophs have become dependent on the host organism for all their nutritional needs and have lost the ability to photosynthesize: full mycoheterotrophs rely on fungal carbon for their entire life cycle (see Merckx 2013). This major nutritional transition has evolved at least 47 times in land-plant evolution (Merckx et al. 2013a). In angiosperms the loss of photosynthesis may be associated with substantial changes in plant morphology, including reduced stature and foliage, and novel floral forms (Leake 1994). Fully mycoheterotrophic plants are generally not green (although limited chlorophyll production is sometimes retained; Cummings and Welschmeyer 1998; Barrett et al. 2014). Heterotrophic plants also often have substantially modified plastid genomes due to elevated rates of substitution, gene losses, and rearrangements (e.g., Wickett et al. 2008; Delannoy et al. 2011; Logacheva et al. 2011, 2014; Barrett and Davis 2012; Lam et al. 2015; Mennes et al. 2015; Schelkunov et al. 2015; Bellot and Renner 2016). The ultimate fate of the plastid genome in heterotrophs may be loss, although this may have occurred in only one embryophyte group (putatively in the holoparasite *Rafflesia*; Molina et al. 2014). The usual persistence of plastid genomes in heterotrophs is thought to reflect nonphotosynthetic functions performed by some plastid-encoded gene products (e.g., Bungard 2004; Barbrook et al. 2006; Wicke et al. 2011). However, retained functional genes may also experience elevated rates of evolution (e.g., Logacheva et al. 2011; Barrett et al. 2014). Gene losses, pseudogenization, and rate elevation may all contribute to uncertainty in inferring higher-order placement of mycoheterotrophs in plant phylogeny (Merckx and Freudenstein 2010).

There are ~514 extant species of fully mycoheterotrophic plants (Merckx et al. 2013a). The majority of species (91%) and origins (~83%) are found in monocots, which seem to be particularly prone to this evolutionary transition (Imhof 2010; Merckx et al. 2013a). In total, seven monocot families include fully mycoheterotrophic taxa (i.e., Burmanniaceae, Corsiaceae, Iridaceae, Orchidaceae, Petrosaviaceae, Thismiaceae and

Triuridaceae; Leake 1994). Orchidaceae have at least 235 fully mycoheterotrophic species, which derive from an estimated 25 losses of photosynthesis (Merckx et al. 2013a, b), and the entire family is mycoheterotrophic during seedling establishment (e.g., Bernard 1909; Leake 1994; Rasmussen 1995; Merckx 2013), referred to as initial mycoheterotrophy. Burmanniaceae are a small family in Dioscoreales with approximately 13 genera and 130 species that include both autotrophic and fully mycoheterotrophic members, with multiple independent origins of full mycoheterotrophy (see Merckx et al., 2006 for a summary of the family's systematic history). Corsiaceae are a family of full mycoheterotrophs, comprising three small genera (*Arachnitis*, *Corsia* and *Corsiopsis*; the latter is likely extinct; Zhang et al. 1999; Mennes et al. 2015). Historically, the family also has had a highly uncertain placement (see Mennes et al. 2015), although it was recently resolved as the sister group of Campynemataceae (Liliales) based on nuclear, mitochondrial, and whole plastome data (Mennes et al. 2015; Bodin et al. 2016). *Geosiris* (comprising three species) is the only fully mycoheterotrophic member of Iridaceae (Fay et al. 2000; Reeves et al. 2001; Goldblatt et al. 2008; Merckx et al. 2013b).

Petrosaviaceae are a small family comprising the autotrophic *Japonolirion osense* and three species of fully mycoheterotrophic *Petrosavia* (Cameron et al., 2003). The family is the sole component of Petrosaviales, the sister group of all monocots except Alismatales and Acorales in most phylogenetic studies (e.g., Fuse and Tamura 2000; Davis et al. 2004; Chase et al. 2006; Graham et al. 2006). Thismiaceae are a fully mycoheterotrophic family with ~five genera and 50 species. The family may be closely related to Burmanniaceae (e.g., APG 2009) although recent molecular evidence suggests that it is instead most closely related to photosynthetic *Tacca* in Taccaceae (Merckx et al. 2009b, 2010). Finally, Triuridaceae are a fully mycoheterotrophic family with ~nine genera and 50 species. The family experienced an ancient loss of photosynthesis based on its inferred crown age (Mennes et al. 2013); late Cretaceous fossils are also known that may represent stem-lineage Triuridaceae (Gandolfo et al. 1998, 2002; Iles et al. 2015), but their trophic status is not known. The family's position in plant phylogeny has been highly uncertain until recently. Molecular evidence from mitochondrial and nuclear genes strongly supports its membership in Pandanales (Mennes et al. 2013), an arrangement

first reported by Chase et al. (2000) based on nuclear 18S rDNA. A recent plastome-based study placed Triuridaceae as the sister group of a clade comprising Cyclanthaceae and Pandanaceae with strong support across a broad variety of likelihood analyses (Lam et al. 2015).

Until recently, most of what is known about the broad picture of land-plant relationships came from pioneering phylogenetic surveys based on one to a few surveyed across a broad range of taxa. For example, many of the phylogenetic relationships underpinning Angiosperm Phylogeny Group (APG) classification schemes (APG, 1998, 2003, 2009) were based on two plastid genes (*atpB*, *rbcL*) and a single nuclear gene, 18S rDNA (e.g., Soltis et al. 2000). Individual genes can also have a strong phylogenetic signal concerning higher-order relationships; indeed, the broad outline of plant phylogeny is often apparent in single-gene analyses (e.g., Chase et al. 1993 for angiosperms using only *rbcL*; Givnish et al. 2005 for monocots using only *ndhF*). Analyses that included a few plastid genes have been used to infer the phylogenetic placement of several mycoheterotrophic lineages, including *Geosiris* (Fay et al. 2000) and *Petrosavia* (Petrosaviales; Fuse and Tamura, 2000; Cameron et al. 2003). However, the use of plastid markers in phylogenetic studies of heterotrophs has generally been regarded as problematic, reflecting expectations of elevated rates of evolution in retained genes (potentially problematic in phylogenetic inference if it leads to long-branch attraction, e.g., Felsenstein 1978; Hendy and Penny 1989) and the hypothesized loss of multiple photosynthetic or other genes or even entire plastid genomes (e.g., Cronquist 1988, p.467; Merckx et al. 2009b). Phylogenetic studies of mycoheterotrophs have therefore tended to focus on analyses on data sets comprising several mitochondrial and nuclear genes (e.g., Neyland and Hennigan 2003; Merckx et al. 2006; Mennes et al. 2013). Despite these concerns, recent studies have validated the use of whole-plastid genomes to infer the placement of mycoheterotrophs (e.g., Logacheva et al. 2014; Lam et al. 2015; Mennes et al. 2015) by providing results congruent with other studies, often with strong branch support.

These recent results suggest that it would be useful to revisit the utility of plastid genes in broadly sampled, few-gene studies. The two photosynthesis-related genes *atpB* and *rbcL* have been among the most widely used plastid genes in higher-order plant

phylogenetic studies. Both genes may have additional nonphotosynthetic roles (Bungard 2004; Wicke et al. 2011, 2013), but with few exceptions (e.g., *Petrosavia*, Cameron et al., 2003; see also Logacheva et al. 2014) they have been lost or pseudogenized in fully heterotrophic lineages. The widely employed plastid *trnL-trnF* spacer region is also often missing in heterotrophic plastomes, in contrast to *matK*, another widely used gene that is commonly retained in heterotrophic lineages (see summaries in Barrett et al. 2014; Lam et al. 2015) and is often used in phylogenetic studies (e.g., Hilu et al. 2003). A portion of *matK* is also widely used as one of the core plant DNA barcoding markers (e.g., Hollingsworth et al. 2009). It is thus one of the few widely used genes that may be suitable for inference of mycoheterotroph phylogeny. Obtaining plastid genes from heterotrophic plants can be difficult, and at least one study has suffered from the inclusion of erroneous plastid sequences from mycoheterotrophs (e.g., Kim et al. 2013; Kim, personal communication in Mennes et al. 2015): I suspect that PCR-based amplifications of plastid genes in heterotrophs may be prone to the recovery of contaminant sequences (cross-contaminants with other plant taxa) when authentic genes are not recoverable, for example, due to gene loss (also discussed later).

Here I assessed three plastid genes, *accD*, *clpP* and *matK*, for their potential as phylogenetic markers in large-scale analyses and, in particular, to address whether they can help place mycoheterotrophs in monocot phylogeny with moderate to strong bootstrap support. The gene *accD* codes for the β -carboxyltransferase subunit of the acetyl-CoA carboxylase (ACCase), which is involved in fatty acid metabolism, and also regulates ACCase activity (Bungard 2004); *clpP* codes for the catalytic subunit of plastid Clp protease (Shanklin et al. 1995; Wicke et al. 2011); *matK* encodes a maturase involved in splicing plastid group IIA plastid introns (Ems et al. 1995; Vogel et al. 1997; McNeal et al. 2009), although some group-IIA splicing may occur despite its loss (Delannoy et al. 2011). Thus, all three genes have functional roles that are not directly related to photosynthesis, and they are all typically retained in the plastid genomes of heterotrophic plants (holoparasites and full mycoheterotrophs; see Wicke et al. 2011; Barrett et al. 2014; Lam et al. 2015). They are therefore good candidate genes for plastid-based phylogenetic inferences that include these plants.

To test the utility of these three genes in broad-scale phylogenetic inference, I analyzed a monocot-wide data set in a range of analyses that consider all families containing mycoheterotrophic taxa. Several factors contributed to the complexity of this study. First, I had to assemble a phylogenetic framework for these three genes that represented a broad array of photosynthetic taxa, as two of the three genes (*accD* and *clpP*) had not previously been widely sampled in phylogenetic studies. Thus, I had to amplify and sequence these genes for most green taxa, which I aligned with those available from previous studies and new sequences from mycoheterotrophs. I used this alignment for phylogenetic inferences that included or excluded various sets of mycoheterotrophs. Second, I had to screen mycoheterotroph sequences for possible cross-contaminant sequences, which I occasionally encountered and removed from analysis.. Third, genes recovered from individual species sometimes exhibited highly elevated substitution rates. I therefore based my phylogenetic inferences primarily on maximum-likelihood (ML) analyses, as likelihood methods are understood to be less prone to long-branch attraction problems than parsimony (e.g., Felsenstein 1988; Yang 1996; Huelsenbeck 1997, 1998; Swofford et al. 2001; Yang and Rannala 2012), and Bayesian methods may lead to inflated confidence in inferred phylogenetic relationships (e.g., Simmons et al. 2004; Yang and Rannala 2005; Kolaczowski and Thornton 2009). I considered both unpartitioned and partitioned DNA substitution models to assess the degree to which the use of simple vs. complex models affects phylogenetic inferences; the partitioning scheme used accommodates differences in DNA substitution models (including model parameters) among genes and codon positions. I also included a parsimony analysis for comparison. Fourth, not all of the genes could be obtained experimentally from all mycoheterotrophic taxa that I surveyed, which inevitably led to somewhat patchy concatenated alignments.

My analyses focused on non-orchid mycoheterotrophic monocots, those that interact with arbuscular mycorrhizal (AM) fungi (Leake 2004; Waterman et al. 2013), although I also included data from recently published plastomes from several fully mycoheterotrophic orchids that associate with ectomycorrhizal fungi (i.e., *Rhizanthella gardneri*, Delannoy et al. 2011; *Neottia nidus-avis*, Logacheva et al. 2011; *Corallorhiza striata*, Barrett and Davis 2012). My study addresses two major objectives: (1) Does this

few-gene data set allow placement of individual mycoheterotrophic taxa (species or families) with moderate to strong bootstrap support? (2) Are these placements congruent with recent results using other sources of molecular data, including nuclear and mitochondrial genes and recent whole-plastid genome studies? I also included a representative sampling of Burmanniaceae to compare to published phylogenetic inferences for this family based on nuclear and mitochondrial data (e.g., Caddick et al. 2002; Neyland 2002; Davis et al. 2004; Merckx et al. 2006, 2009b). Finally, as plastid genes have not been reliably recovered from Thismiaceae to date, I was particularly interested in assessing whether I could recover any of the three plastid genes from this family for use in phylogenetic inference.

2.3 Materials and methods

2.3.1 DNA extraction, primer design, amplification and sequencing

I obtained new DNA sequence data for 34 photosynthetic or fully mycoheterotrophic taxa from mycoheterotrophic lineages, and 61 additional photosynthetic taxa (59 of which are monocots; see Table A.1 for sources). Total genomic DNAs were extracted from silica-dried material using a modified CTAB protocol (Doyle and Doyle 1987; Rai et al. 2003) or were obtained from colleagues or DNA banks (Royal Botanic Gardens, Kew; Missouri Botanical Garden; SANBI, Kirstenbosch). I amplified and sequenced *accD*, *clpP* and *matK*. DNA amplification and Sanger sequencing used previously published and new primers (Fig. A.1). I designed additional novel primers by considering monocot-wide sequence alignments generated from available GenBank sequences, using visual inspection to identify conserved regions suitable for primer placement, and Oligo 7 (Rychlik 2007) and Amplify 3x (Engels 1993) to screen candidate primers based on their predicted success in DNA amplification (for criteria, see Graham and Olmstead, 2000). Exon boundaries for *clpP* were based on the complete plastid genome of *Dioscorea elephantipes* (GenBank accession NC_009601.1). DNA amplification and sequencing protocols generally followed Graham and Olmstead (2000), although in some cases I replaced *Taq* polymerase with Paq5000 (Agilent Technologies, Santa Clara, California, USA). I purified amplification products using QIAquick PCR purification columns

(QIAGEN Inc., Valencia, California, USA) or ExoSAP-IT reagent (USB Corporation, Cleveland, OH, USA), following manufacturer instructions, and used 1/26 reactions of BigDye Terminator v. 3.1 (Applied Biosystems, Foster City, CA, USA) to perform dideoxy sequencing reactions. I sequenced all regions at least twice, generally twice each in both forward and reverse directions. For some taxa, I was not able to sequence through the entire second intron in *clpP*, and so I represented these sequences in the alignment as two fragments. I obtained one or more genes for several taxa using assemblies of circular plastid genomes retrieved from genome skims using Illumina data, following Mennes et al. (2015); these genomes will be presented in Chapter 4.

2.3.2 Alignment construction

I performed base-calling and contig assembly using the program Sequencher 4.2.2. (Gene Codes Corp., Ann Arbor, Michigan, USA), and aligned finalized contigs using Se-AL 2.0a11 (Rambaut 2002) and the criteria outlined by Graham et al. (2000), Kelchner (2000) and Simmons and Ochoterena (2000). I added new sequences for the three genes to those retrieved from published plastid genomes of 53 additional taxa (see Table A.1).

I identified several cases of possible contaminant sequences among individual genes recovered from mycoheterotrophs (see below), initially by using BLAST (Altschul et al. 1990). I suspect that these derive from residual non-target DNA contamination that becomes apparent when the main amplification fails due to gene loss or rapid gene evolution away from the targeted priming sequences. I identified probable contaminants by examining maximum-likelihood (ML) trees inferred for each locus individually, to check for terminal taxa that are distantly related to congeneric or confamilial taxa. Typically these had short connecting branches to the nearest green taxon. One example is a likely contaminant *matK* sequence obtained from a *Thismia aseroe* extract (Fig. A.2) that shows 97% similarity to *Moraea riparia* (Iridaceae; GenBank accession JX903631.1). The behaviour of this sequence contrasts with what I believe to be genuine Thismiaceae *accD* sequences (259-314 bp sequenced portions, compared to 511 bp portions sequenced for other taxa), found on much longer branches (these also often placed in unusual positions, see below), that have BLAST scores of 87-88% to other monocots, eudicots and other angiosperms. Other putative contaminants that placed

outside monocots were not so straightforward to place phylogenetically because of limited taxon sampling in other angiosperms here. An example of this is a *matK* sequence retrieved from a *Geomitra clavigera* extract, which I identified as a probable contaminant based on close BLAST-based similarity to a eudicot (99% to *Capparis spinosa*, Capparaceae; GenBank accession AY491650.1; in a phylogenetic analysis, this sequence grouped with two eudicots included as outgroups, Fig. A.3). I excluded all candidate contaminant sequences identified in this way, but tentatively included other genes from the same sample that passed my phylogenetic screen (see Table A.1; note that none of the other sequences included in final alignments from these samples showed obvious double peaks, so contaminants are likely present at low concentration in the original DNA extracts).

2.3.3 Phylogenetic analysis

All subsequent analyses considered the final concatenated three-gene DNA sequence alignment, which is publicly available at figshare.com (doi: 10.6084/m9.figshare.2062158).

I conducted initial analyses using parsimony and likelihood with all green and mycoheterotrophic taxa included. For the parsimony search I used the program PAUP* v.4.0a134 (Swofford 2003), with tree-bisection-reconnection branch swapping and 1,000 random stepwise addition replicates, holding 100 trees at each step, and otherwise used default settings in the search for shortest trees. For maximum likelihood (ML) analysis I used the program RAxML v. 7.4.2 (Stamatakis 2006) with a graphical interface (Silvestro and Michalak 2012), considering partitioned and unpartitioned versions of the data, see below. For these and all subsequent likelihood analyses, I ran 20 independent heuristic searches using different starting points.

I then performed individual likelihood analyses considering each mycoheterotrophic family (or taxon) separately to try to minimize the potential for attraction between distantly related taxa, using the search method outlined above. For Burmanniaceae, I ran an additional analysis that included only green species for the family (i.e., likely photosynthetic; Merckx et al. 2006), and for Thismiaceae, I also analyzed each species (three in total) separately, as I only retrieved a portion of *accD* for

these taxa. I also ran an analysis including only photosynthetic angiosperms (considering photosynthetic mycoheterotrophs, green orchids were included and green Burmanniaceae excluded) to investigate the general utility of the three genes in inference of monocot higher-order relationships.

I ran two variant ML analyses in all cases, one considering the data unpartitioned, and another with the data partitioned by gene and codon position. For the latter I defined 14 initial data partitions based on the three codon positions for coding regions (considering the two exons of *clpP* separately) and two introns (in *clpP*). I then used the program PartitionFinder v. 1.1.1 (Lanfear et al. 2012) to assess which of the initial partitions had significantly different DNA substitution models or model parameters in each case, using the Bayesian information criterion (BIC; Schwarz 1978; Sullivan and Joyce 2005) and the strict hierarchical clustering algorithm. I repeated these tests for each distinct likelihood analysis, and used the final partitioning schemes for each ML analysis (summarized in Table A.2). I applied the GTR+G model to all data partitions, as this model or close variants were found for all unpartitioned data sets, and for most individual data partitions in partitioned analyses (Table A.2; the GTR+G+I model was found in a few cases, but the ‘I’ parameter for invariant sites may be accommodated by the gamma parameter ‘G,’ see Yang 2006).

I assessed branch support using bootstrap analysis (Felsenstein 1985). For the likelihood analyses, I ran 500 bootstrap replicates (using the “thorough bootstrap” option in RAxML), considering the same partitioning schemes and models of DNA substitution described above. For the parsimony bootstrap analysis I used 1,000 bootstrap replicates, with 100 random addition replicates per bootstrap replicate, and otherwise used default settings. I considered 90% and above to be “strongly supported” (or “well-supported”), 70-89% to be “moderately supported” and <70% to be “weakly supported” (Zgurski et al. 2008).

2.3.4 Constraint tests of monophyly

I tested for the monophyly of two groups of interest that were each not recovered as monophyletic in any shortest trees (see below): (1) a putative clade comprising Burmanniaceae and Thismiaceae corresponding to the circumscription of Burmanniaceae

s.l. in recent versions of the Angiosperm Phylogeny Group classification system (APG 2003, 2009); (2) a clade comprising non-green Burmanniaceae, as these species are not expected to be grouped together in a clade, given multiple independent losses of photosynthesis are predicted in the family (Merckx et al. 2006). In both cases, I was interested in determining whether I had sufficient power to reject an hypothesis that placed the constrained taxa together. Each constraint analysis excluded other mycoheterotrophs that were not relevant to the hypothesis. I found the shortest likelihood trees that satisfied monophyly constraints that I set up for RAxML (using unpartitioned ML analysis). I then compared the resulting tree sets using the approximately-unbiased (AU) (Shimodaira 2002) and Shimodaira-Hasegawa (SH) tests in the program CONSEL (Shimodaira and Hasegawa 2001), using site-likelihoods from unpartitioned ML analysis, to ask whether the constrained trees were significantly worse explanations of the data than the best (unconstrained) tree.

2.3.5 Characterizing rate elevation in heterotrophic lineages

I characterized relative differences in overall substitution rates in mycoheterotrophic and green lineages in a Bayesian framework, using the program BEAST v. 1.8.2 (Drummond et al. 2012). My focus was on characterizing relative rate variation across lineages, and so I fixed the input tree topology based on the best likelihood tree recovered from the partitioned analysis that included all taxa (see Fig. 2.2), with the exception that I constrained Triuridaceae, represented by *Sciaphila*, to be monophyletic (see below). I specified a single GTR+G nucleotide substitution model across sites, a random local-clock model with a fixed mean rate of 1.0 substitution per site, and a Yule speciation model, and otherwise used default settings. I ran BEAST for a combined 800 million generations across 20 separate analyses, sampling trees every 1000 generations, assessing convergence using the program Tracer v. 1.6 (Rambaut et al. 2014). All parameters had final effective sample size (ESS) values of at least 200. I used the programs LogCombiner v. 1.8.2. and TreeAnnotator v.1.6.2 (Drummond et al. 2012) to combine trees from individual analyses, discarding the first 25% of sampled trees as burn-in and re-sampling at a lower frequency to yield 10,000 final trees. The combined tree was visualized using the program FigTree v. 1.31 (Rambaut 2006).

2.4 Results

2.4.1 Green monocot phylogeny inferred from *accD*, *clpP* and *matK*

The placements of autotrophic monocot families, orders and higher-level taxa in the ML analysis of autotrophic taxa (Fig. 2.1) are congruent with previous monocot-wide studies using more genes (e.g., Graham et al. 2006; Hertweck et al. 2015). Most deep branches in the monocots were also strongly supported, with only minor exceptions (e.g., the placement of Dasypogonaceae, a sister-group relationship between Liliales and Asparagales-commelinids, both with <50% bootstrap support). There were no substantial differences in phylogenetic relationships of autotrophic monocots and their support values between the partitioned analysis and the unpartitioned analysis for this taxon set (data not shown).

2.4.2 Phylogenetic placement of mycoheterotrophs in monocot phylogeny

i. Placements with all mycoheterotrophs included

Initial analyses (parsimony and likelihood) included all green and mycoheterotrophic lineages together. Parsimony trees (most parsimonious trees) grouped all or most of the fast-evolving lineages in a single large clade that may result from long-branch attraction, marked with an arrow in Fig. A.4 (all shortest trees either recovered this clade or had a slightly smaller version lacking Thismiaceae; see below on rates of evolution in mycoheterotrophs). This clade was poorly supported and had poorly supported and variable internal structure across shortest trees (note that black dots in this figure indicate branches that collapse in a strict consensus of the most-parsimonious trees). In contrast, full mycoheterotrophs that terminated shorter branches (i.e., *Geosiris*, orchids, *Petrosavia*, *Corsia*) placed with moderate to strong support in the parsimony analysis in Iridaceae, Orchidaceae, Petrosaviaceae, and Liliales, respectively. Corsiaceae were not recovered as monophyletic in this analysis, as *Arachnitis* grouped within the fast clade (Fig. A.4).

In the likelihood analyses that included all mycoheterotrophs, fully mycoheterotrophic lineages on shorter branches again placed with moderate to strong likelihood bootstrap support in the positions observed in the parsimony analysis (see

Figs. 2.2, A.4, A.5 for *Geosiris*, orchids, *Petrosavia*). However, multiple rapidly evolving lineages (see below) that grouped together in the parsimony analyses were instead inferred to be in dispersed positions in the likelihood analyses. Specifically, Burmanniaceae (the monophyly of which was poorly supported) were poorly supported as the sister group of a clade that included Taccaceae, Trichopodaceae, Thismiaceae and Dioscoreaceae, with Thismiaceae then the sister group of *Tacca* (Taccaceae). The latter arrangement was poorly supported, but the monophyly of Thismiaceae was well supported. *Sciaphila* (Triuridaceae) was divided between two locations, with one pair of taxa as the sister group of Cyclanthaceae and Pandanaceae in Pandanales, and the other nested in Nartheciaceae (Dioscoreales) as the sister group of *Lophiola*. This split placement was seen in both partitioned and unpartitioned ML analyses (Figs. 2.2, A.5). The monophyly of Corsiaceae was well supported (i.e., *Arachnitis* no longer placed in the fast clade), and the entire family was supported as the sister group of Campynemataceae with weak to moderate support in both likelihood analyses (Figs. 2.2, A.5). The inclusion of full mycoheterotrophs tended to depress likelihood bootstrap support for branches neighboring all of the mycoheterotrophic families, compared to analyses that included only green taxa. Neighbouring branches often experienced at least 10% worse bootstrap support, and often had substantially larger drops in support (e.g., for the monophyly of Asparagales and Dioscoreales, cf. Figs. 2.1 and 2.2).

ii. Placements with mycoheterotrophic taxa considered individually

In general, the likelihood analyses that examined each mycoheterotroph family separately recovered placements of mycoheterotrophic taxa consistent with the analyses that included them all simultaneously (cf. Figs. 2.2 and 2.3, A.6-A.10), with exceptions outlined below for Thismiaceae and Triuridaceae. These local placements were also generally well supported (Fig. 2.3), a contrast with the likelihood analyses that included all mycoheterotrophic taxa simultaneously (Figs. 2.2, A.5), and there was less reduction in bootstrap support values for branches neighboring all of the mycoheterotrophic families, when the mycoheterotrophic families were added individually. For example, the monophyly of both Asparagales and Dioscoreales was again well supported (cf. Figs. 2.1-2.3). There was again strong support for the monophyly of Corsiaceae, Iridaceae,

Orchidaceae and Petrosaviaceae (Fig. 2.3). Support for the monophyly of Burmanniaceae when other mycoheterotrophic families were excluded was also higher than when they were included, but was still weak (61-69% support, compared to <50% support; see Figs. 2.2, 2.3b, A.5). However, when only photosynthetic Burmanniaceae were included, the bootstrap support for the family's monophyly improved substantially (to 100% in partitioned and unpartitioned ML analyses, see inset subtree in Fig. 2.4). The membership of Burmanniaceae in Dioscoreales was also well supported, although its local position in the order still lacked strong support (Fig. 2.4). I could not reject the existence of a clade comprising Burmanniaceae and Thismiaceae (AU and SH tests comparing trees from best unconstrained vs. constrained ML analyses that included both families: $P = 0.104$ and 0.125 , respectively, respectively; the unconstrained tree placed Burmanniaceae as the sister group of other Dioscoreales, and Thismiaceae as the sister group of *Tacca*, data not shown).

The behavior of Triuridaceae and Thismiaceae (in likelihood analyses, with each family included individually) was sensitive to the details of the analysis performed. In the partitioned likelihood analysis (that included only *Sciaphila*, my representative of Triuridaceae, and excluded other mycoheterotrophs), the family was inferred to be monophyletic in the partitioned likelihood analysis, and it placed as the sister group of Cyclanthaceae-Pandanaceae (in Pandanales), all with weak support (Figs. 2.3e, A.11). The unpartitioned version of this analysis recovered two isolated lineages of *Sciaphila* in Pandanales and Dioscoreales (Fig. A.12) similar to the likelihood analyses that included all mycoheterotrophs (Figs. 2.2, A.5), and again with weak support (Fig. 2.3e). Thismiaceae were unexpectedly inferred to be the sister group of *Japonolirion* in the partitioned analysis (Figs. 2.3c, A.13), but not in the unpartitioned analysis, where the family placed with Taccaceae and Trichopodaceae (Fig. A.14); note that Thismiaceae species are represented here only by *accD* (all analyses presented are concatenated three-gene analyses). Separate analyses that included individual species from the family placed the three surveyed species in divergent places (Fig. 2.5), but generally close to *Tacca* (Taccaceae) and/or *Trichopus* (Trichopodaceae) (i.e., *Geomitra clavigera* as the sister group of *Tacca* and *Trichopus* in both partitioned and unpartitioned likelihood analyses, *Thismia aseroe* as the sister group of *Tacca* for partitioned and unpartitioned analysis;

Thismia sp. as the sister group of Maundiaceae in Alismatales in unpartitioned likelihood analysis, but as the sister group of *Tacca* in partitioned analysis; Fig. 2.5).

2.4.3 Relationships in Burmanniaceae

The relative positions of the photosynthetic members of Burmanniaceae were largely consistent among likelihood analyses. Various fully mycoheterotrophic lineages were recovered as nested among the photosynthetic taxa, but for the most part with poor support for the relative arrangements of green and non-green taxa (Fig. 2.3b). Several branches had different arrangements between analyses with other mycoheterotrophic monocots excluded or included (e.g., Figs. 2.2, 2.3b), but these all involved poorly supported relationships. Well-supported relationships in Burmanniaceae include a small clade comprising *B. bicolor*, *B. biflora* and *B. stuebelii*, a sister-group relationship between *Burmannia itoana* and *B. wallichii*, between *B. coelestis* and *B. disticha*, between these two taxa and *B. longifolia*, and between *Gymnosiphon aphyllus* and *G. longistylus* (Fig. 2.3b). The intermingling of fully mycoheterotrophic members of Burmanniaceae with photosynthetic members of the family implies multiple evolutionary losses of photosynthesis (Figs. 2.3b, A.8). Consistent with a hypothesis of multiple losses of photosynthesis, the AU and SH tests rejected the existence of a constrained clade that comprises only non-photosynthetic taxa of Burmanniaceae ($P < 0.02$ in both cases). When only green Burmanniaceae were included, the relative placement of *B. capitata* was different (but not well supported; cf. Figs. 2.3b, 2.4), and the sister-group relationship of *B. madagascariensis* and the clade comprising *B. bicolor*, *B. biflora* and *B. stuebelii* was recovered with improved support (98-99%; Fig. 2.4).

2.4.4 Rate elevation in mycoheterotrophic monocots

Relative differences in the overall substitution rate are summarized in Fig. 2.6 (see Fig. A.15 for more detail). Three major rate bands are shown, with the highest rates (in red) at least twice as fast as the fastest slow ones (thin black lines); the thick black line is an intermediate rate. Many fully mycoheterotrophic lineages have intermediate to fast rates, although Petrosaviaceae, *Geosiris* (Iridaceae), *Corallorhiza striata* (Orchidaceae), and several fully mycoheterotrophic Burmanniaceae are exceptions (terminal photosynthetic

lineages of Burmanniaceae are also in the lowest rate band; Fig. A.15). Note that the rate band cut-offs used here (chosen to emphasize the most rapidly evolving lineages) include a broad range of rates (e.g., the lowest rate band includes a nearly nine-fold range of rates, see Fig. A.15 for more details). The most elevated rates (thick lines in Fig. 2.6) include several fully mycoheterotrophic lineages (*Arachnitis* in Corsiaceae, several lineages of Burmanniaceae, Orchidaceae, Thismiaceae and Triuridaceae). Several photosynthetic lineages in Alismatales, Poales and relatives are also rapidly evolving.

2.5 Discussion

2.5.1 Using plastid genes to place mycoheterotrophic monocots

Compared to photosynthetic taxa, plastid genomes retrieved from heterotrophic lineages can be both rapidly evolving and reduced in terms of size and gene content, sometimes exceptionally so (e.g., Lam et al. 2015; Mennes et al. 2015; Bellot and Renner 2016). My few-gene approach thus provides general insights into the suitability of highly reduced, patchily sampled, rapidly evolving genomes for inferring the phylogenetic placement of heterotrophic plant lineages. Several key problems had to be addressed to do so. First, I lacked a phylogenetic framework (large-scale alignment of commonly retained genes) for photosynthetic taxa to help place the heterotrophic lineages. I addressed this by constructing a large-scale alignment of three commonly retained genes in heterotrophs (*accD*, *clpP*, *matK*) from photosynthetic monocots, which I sampled most heavily in the clades (orders) thought to have given rise to heterotrophic taxa. The higher-order relationships of photosynthetic monocots inferred using the three-gene data set are congruent with recent phylogenetic studies that employed a wide variety of taxonomic and gene samplings (e.g., Chase et al. 2006; Graham et al. 2006; Givnish et al. 2010; Soltis et al. 2011), and generally well supported (Fig. 2.1). This congruence supports the general utility of using these genes for making inferences about high-order monocot relationships, at least for photosynthetic taxa.

The often highly elevated rates of evolution of retained plastid genes in heterotrophic lineages may be problematic for phylogenetic inference. Although our analytical understanding of long-branch effects is still quite limited (Parks and Goldman

2014), model-based methods like maximum likelihood are thought to be less sensitive than parsimony to problematic long branches than parsimony (see also Felsenstein 1978, 1988; Yang 1996; Huelsenbeck 1997, 1998, Swofford et al. 2001; Yang and Rannala 2012). This lower sensitivity appears to be the case here, as my parsimony analysis depicted a collection of rapidly evolving mycoheterotrophic taxa as part of a poorly supported ‘fast’ clade (Fig. A.4) that is taxonomically highly heterogeneous (e.g., APG 2009). This result is a strong contrast with the likelihood analyses (cf. Figs. 2.2 and A.5). Despite substantial rate elevation, likelihood analyses placed mycoheterotrophic lineages in phylogenetic positions consistent with studies using nuclear and mitochondrial genes (e.g., Merckx et al. 2006, 2009b; Mennes et al., 2013, 2015). Thorough taxon sampling can help avoid long-branch artefacts (e.g., Pollock et al. 2002; Soltis and Soltis 2004; Hedtke et al. 2006; Heath et al. 2008); however, adding taxa can introduce additional problems if the additional taxa are also on long branches (e.g., Kim 1996; Hillis 1998). The latter effect may also explain why including all mycoheterotrophic taxa simultaneously leads to substantially larger reductions in branch support for the branches neighboring heterotrophic lineages (those that were well supported in the analysis of green taxa only), than occurred when I included each mycoheterotrophic family individually (cf. Figs. 2.1-2.3).

Finally, missing data can present a challenge for phylogenetic inference, and may be problematic in some cases here, as one or two of the three genes I considered were sometimes lost from the genome or were not readily recoverable for several taxa. Incompletely sampled taxa may still tend to improve the accuracy of phylogenetic inference when taxon sampling is otherwise limited (e.g., Wiens 2006; Wiens and Tiu 2012). However, the combination of missing data and long branches may nonetheless prove to be especially severe (e.g., Wiens 2005, 2006), which is potentially problematic here for several patchily sampled taxa with elevated substitution rates (e.g., Figs. 2.2, 2.5; Table A.1).

2.5.2 Utility and limits of the current approach

How well do my phylogenetic markers perform in likelihood analyses? In general, I did not see major differences in levels of bootstrap support values between unpartitioned and

partitioned likelihood analyses for these taxa across the different datasets that I analyzed (Figs. 2.2 vs. A.5; Fig. 2.3), supporting the idea that my results are not sensitive to whether the data are partitioned or not (partitioned likelihood analyses attempt to account for DNA substitution model or model parameter differences among different regions). This finding is also consistent with the findings in Chapter 3 in my whole plastid-genome study of Triuridaceae. Because long branches in different mycoheterotroph families likely interfere with well-supported placement of individual taxa when considered simultaneously (Fig. 2.2), I focus the remaining discussion on the separate likelihood analyses of mycoheterotrophs (i.e., each family considered individually), summarized in Figs. 2.3-2.5.

I placed the majority of mycoheterotrophic taxa in monocot phylogeny with strong support, at least regarding their family-level placements (Fig. 2.3a, d, f). My results are also consistent with current understanding of their phylogenetic placements based on other data sets (i.e., *Geosiris*, Fay et al. 2000; mycoheterotrophic Orchidaceae, Ruhfel et al. 2014; Corsiaceae, Mennes et al. 2015, Bodin et al. 2015; Petrosaviaceae, Cameron et al. 2003, Logacheva et al. 2014). The position of Corsiaceae has been unclear until very recently. The family was strongly supported here as the sister group of Campynemataceae in Liliales, which is consistent with recent whole plastome data and other data for the family (Bodin et al. 2015; Mennes et al. 2015). *Arachnitis* (Corsiaceae) has a particularly long branch due to highly elevated substitution rate (Figs. 2.2, 2.6, A.9). However, the other genus in the family, *Corsia*, has a moderately elevated rate of evolution (Fig. 2.6) and has retained all three surveyed genes, which may account for the family's clear and well-supported placement here in likelihood analyses (Figs. 2.3d, Table A.1). My results do not support the findings of Neyland and Hennigan (2003) that Corsiaceae are polyphyletic, recovered in their analysis of a nuclear 26S rDNA data set (a probable artifact due to limited taxon sampling and their use of parsimony, see Mennes et al. 2015).

Three families were less confidently placed in monocot phylogeny by my three-gene data set: Burmanniaceae, Thismiaceae, and Triuridaceae (Figs. 2.3-2.5). With regards to the first, Burmanniaceae, my best likelihood trees are generally consistent with other recent studies based on nuclear and mitochondrial data (Merckx et al. 2006, 2008).

I found strong support for Burmanniaceae being part of the order Dioscoreales when only green members of the family were included (Fig. 2.4), and moderate support for this arrangement when fully mycoheterotrophic taxa were included (Fig. 2.3b), although the family's position within the order was still ambiguous in both cases. The green taxa of Burmanniaceae may be partial mycoheterotrophs based on vegetative reduction and reduction in chlorophyll (see Merckx et al. 2006), supported by isotopic evidence in *B. coelestis* (Bolin et al. in press). The green taxa often have lower substitution rates than full mycoheterotrophs in the family (Figs. 2.6, A.15, but the family as a whole appears to have an elevated rate of evolution when only green taxa are considered (long branch subtending the green clade in Fig. 2.4, see also Figs. 2.6, A.15).

A diversity of phylogenetic studies based on molecular and morphological data have placed Triuridaceae in Pandanales (e.g., Chase et al. 2000; Rudall and Bateman 2006; Mennes et al. 2013). The family was recently resolved as the sister group of Cyclanthaceae and Pandanaceae with strong support, based on a whole-plastome analysis (Lam et al. 2015). Here it was recovered as monophyletic only in the partitioned ML analysis (see shortest tree in Fig. A.11), but with poor support (Fig. 2.3e). The two samples of *S. densiflora* included here (vouchers Pillon et al. 88, and Duangjai 029 in Table A.1) were distantly related to each other (Figs. A.11, A.12), perhaps pointing to within-genus misidentification of one of them, although I did not confirm this possibility here. The lack of monophyly of *Sciaphila* in some likelihood analyses is presumably a long-branch artifact. Triuridaceae were weakly supported as a member of Pandanales in the partitioned ML analysis (Fig. 2.3e). The family has some of the most elevated rates of evolution and some of the longest branches in my study (e.g., Fig. 2.2; Figs. 2.6, A.15).

Particular caution seems warranted concerning my results for Thismiaceae. The data reported here for this family are especially intriguing because they may represent the first genuine plastid data to be recovered for it (i.e., for *accD*), from each of three species in the family (*Geomitra clavigera*, *Thismia aseroe* and *Thismia* sp.; Table A.1). Rates of molecular evolution are elevated for these taxa compared to photosynthetic taxa (Figs. 2.6, A.15; note that these are three-gene analyses that include only *accD* from this family). The placement of Thismiaceae in analyses that include all three taxa was unstable (sister to Taccaceae or Petrosaviaceae; Figs. A.13, A.14) and very poorly

supported (e.g., Fig. 2.3c). Analyses that included each species individually placed them in various positions with weak support that were also easily perturbable between analyses (e.g., by using partitioned vs. unpartitioned likelihood; Fig. 2.5, a-c). The *Geomitra clavigera* and *Thismia* sp. samples placed close to or within Taccaceae and/or Trichopodaceae in some analyses (Fig. 2.5), consistent with previous studies by Merckx et al. (2006, 2009b), who presented mitochondrial and nuclear data that supported a placement of Thismiaceae in Dioscoreales, close to Taccaceae. The precise position of Thismiaceae in the order was strongly supported in their earlier study (Merckx et al. 2006), and moderately supported in a subsequent one (Merckx et al. 2009b), and they also provided evidence that *Afrothismia* (not included here) represents a lineage distinct from other Thismiaceae (Merckx et al. 2009b). The successful retrieval of *accD* sequences for *Geomitra*, *Thismia aseroe* and *Thismia* sp. support a retention of plastid genomes in at least some species in Thismiaceae. As such, I may expect to find small, cryptic plastomes in Thismiaceae that include, at the very least, the *accD* locus.

2.5.3 Are Burmanniaceae and Thismiaceae closely related?

The proposed taxonomic circumscription of Burmanniaceae to include Thismiaceae (APG 2003, 2009) was based primarily on studies that surveyed several photosynthetic plastid genes (*atpB*, *rbcL*) from these families (Caddick et al. 2000, 2002) that may include contaminants (M. Chase, Royal Botanical Gardens, Kew; pers. comm.). Consistent with this, these genes have not been recovered in full circular genomes of any fully heterotrophic taxa in Burmanniaceae (see Chapter 4). My inability here to reject a close relationship between these families in the AU and SH tests may simply reflect a lack of power to reject such hypotheses from the genes I surveyed (in particular, only *accD* was recoverable from Thismiaceae). I therefore propose that the broad treatment of Burmanniaceae to include members of Thismiaceae should be abandoned in future APG treatments until further evidence (ideally including plastid data and other sources of evidence) is obtained for where Thismiaceae fits in monocot phylogeny. Nonetheless, at least some of the plastid-based likelihood trees inferred here (Figs. 2.2, 2.5, A.14) are consistent with current evidence from mitochondrial and nuclear data, which support a

position of Thismiaceae distinct from Burmanniaceae and closer to Taccaceae (Caddick et al. 2002; Merckx et al. 2006, 2009b).

2.5.4 Phylogenetic relationships within Burmanniaceae

Relationships within Burmanniaceae are generally poorly supported here. Not surprisingly, fully mycoheterotrophic taxa in this family often had highly elevated rates of evolution (Figs. 2.6, A.15). Several taxa have one or two of the three genes that were not retrieved (Table A.1), confirmed to be missing genes in several cases (see Chapter 4). Family relationships within Burmanniaceae were mostly strongly supported when only green Burmanniaceae are included in analysis (Fig. 2.4), as was family monophyly. Merckx et al. (2006) suggested that there were at least six independent losses of photosynthesis in Burmanniaceae, based on an analyses of the mitochondrial *nad1 b-c* intron and nuclear 18S rDNA sequences. Although my sampling of Burmanniaceae has fewer taxa than theirs, the relationships inferred here also point to multiple independent losses of autotrophy among taxa, as there are three green lineages here (*B. capitata* and two small clades with three-four taxa) that are deeply nested among non-green lineages (Figs. 2.3b, A.8). Although these “backbone” relationships were not well supported here (see Fig. 2.3b; note that my sampling of the family is more limited than Merckx et al. 2006 and Merckx et al. 2008, and I also used their scorings of photosynthetic vs. fully mycoheterotrophic taxa), the AU and SH tests indirectly support scenarios with more than one loss of photosynthesis, as a tree constraining all non-green taxa as a clade is significantly longer than the best tree. My analyses recovered several intra-familial relationships that were congruent with those from Merckx et al. (2006) including a clade comprising achlorophyllous taxa *B. oblonga* and *B. lutescens*, another comprising the achlorophyllous *B. itoana* and *B. wallichii*, and a clade comprising three autotrophic species (*Burmannia bicolor*, *B. stuebelii* and *B. flora*).

2.5.5 Retention of plastid genes and genomes in monocot mycoheterotrophs

Although some non-photosynthetic plants are hypothesized to have lost their plastid genomes entirely (e.g., the eudicot holoparasite *Rafflesia*; Molina et al. 2014), which is more clearly demonstrated in multiple lineages of secondarily heterotrophic unicellular

eukaryotes (Abrahamsen et al. 2004; Smith and Lee 2014; Janouškovec et al. 2015), all other heterotrophic plant lineages surveyed to date have retained their plastid genomes (e.g., Wolfe et al. 1992; McNeal et al. 2007; Wickett et al. 2008; Delannoy et al. 2011; Logacheva et al. 2014; Lam et al. 2015; Mennes et al. 2015; Bellot and Renner 2016). A small set of genes (including *accD*, *clpP*, *trnE-UUC*, I-CAU and *fM-CAU*, the four rDNAs, some ribosomal proteins) are commonly retained across many heterotrophic lineages that have highly reduced plastomes (e.g., Wicke et al. 2013; Barrett et al. 2014; Lam et al. 2015, but see Bellot and Renner 2016), suggesting that they are essential genes and may not be readily replaced by nuclear counterparts (either functionally transferred genes or replacement by analogous systems; Barbrook et al. 2006). I chose the three plastid genes (*accD*, *clpP* and *matK*) for my phylogenetic survey because they have non-photosynthetic roles and are frequently retained when photosynthetic genes are lost. Both *accD* and *clpP* have been lost occasionally in flowering plants, but this appears to be unrelated to the loss of photosynthesis (e.g., Jansen et al. 2007; Straub et al. 2011); *matK* loss may eventually occur when a sufficient number of plastid genes with group IIA introns have been lost, which may only occur in the later stages of plastid genome degradation in heterotrophic lineages (McNeal et al. 2009).

Some of the genes recovered here by amplification could conceivably represent nuclear or mitochondrial inserts of plastid genes. However, successful functional gene transfer from the plastid to nuclear genome is generally rare in land plants (e.g., Martin et al. 1998) and has not been confirmed in any plants for these three genes (although a potentially functional copy of *accD* may have been transferred to the nucleus in *Trachelium*; Rousseau-Gueutin et al. 2013). Nonfunctional copies would also be expected to degrade rapidly in the nuclear genome (e.g., Matsuo et al. 2005; but see Cusimano and Wicke 2015). Mitochondrial inserts should also degrade (functional transfer of plastid genes to mitochondria is unknown, e.g., Hao and Palmer 2009) and the genes would generally be expected to evolve slowly after insertion, consistent with the lower mutation rate of this genome (e.g., Wolfe et al. 1987; Palmer and Herbon 1988). The phylogenetically widespread retention here in mycoheterotrophs of rapidly evolving plastid genes that retain open reading frames are likely hallmarks of their retention in the plastid genome. In a few cases, I have also obtained plastid genomes from the same

species (*Apteria aphylla*, *Arachnitis uniflora*, *Burmannia bicolor*, *Burmannia capitata*, *Geosiris aphylla*, *Petrosavia sakurii*, *Petrosavia* aff. *sakurii* and *Sciaphila densiflora*; Table A.1), confirming the retention of genes in plastid genomes that I also retrieved using PCR amplification.

2.5.6 Dealing with contamination in heterotrophic samples

Recovering contaminant sequences may be a frequent problem with samples from heterotrophic plants. I detected several possible instances here (seven of 34 taxa). I suspect these represent very weak contaminants that would not normally be evident, but which are apparent here when the target gene is no longer present, or has evolved rapidly away from amplification or sequencing primer sites. It is possible that these represent instances of horizontal gene transfer, as has been demonstrated for parasitic plants (e.g. host to parasite, Davis and Wurdack 2004; Yoshida et al. 2010, and parasite to host; Mower et al. 2004). However, this possibility may be unlikely, because mycoheterotrophs do not have plant-to-plant connections, and any such transfers would have to involve a fungal intermediate. If not detected, the inclusion of contaminant genes would be highly problematic for phylogenetic inference (or in DNA barcoding studies, see below). While I propose that my approach for identifying contaminants is conservative, it may sometimes result in false positives, as sampling of outgroup sequences for *clpP* and *accD* is relatively sparse here and on GenBank, limiting my ability to distinguish the green taxa that contaminant sequences are most closely related to using tree-based methods. As a consequence, it was not always possible to confirm the probable identity of putative contaminating sequences. However, I have unpublished whole-plastid genome data in hand for several species that I also surveyed by in the three-gene analysis here, which allowed me to confirm authentic gene loss from the plastid genomes for several taxa where I obtained contaminant amplifications from conspecific samples (i.e., *clpP* and *matK* for *Apteria aphylla* in Burmanniaceae, and *matK* for *Arachnitis uniflora* in Corsiaceae; Table A.1). These genomic data confirm the validity of including non-contaminant genes from samples that had residual contamination, at least for these two species, because the genes obtained by PCR

amplification and Sanger sequencing here, which I included in analyses were identical or very similar to those obtained using next-generation methods (Figs. A.8, A.9).

2.5.7 Suitability of the three genes as DNA barcoding markers

I did not specifically address the utility of the three plastid markers here as DNA barcoding markers, although two of them have been tested for their suitability in previous DNA barcoding studies (*accD* and *matK*, see below). In green plants, portions of *matK* and *rbcL* have been used as core barcoding markers, often supplemented with the *psbA-trnH* intergenic spacer, and the nuclear ITS (internal transcribed spacer) region (e.g., Fazekas et al. 2008; Hollingsworth et al. 2009, 2011; Li et al 2011). However, *rbcL* and *psbA-trnH* are not suitable as barcoding markers in fully heterotrophic plants because they either are, or involve, photosynthetic genes. The *matK* region used here is the DNA barcoding region for this gene (Fig. A1; Hollingsworth et al. 2009), and my study demonstrates that it can be recovered from fully mycoheterotrophic monocots with relative ease. However, I was unable to recover *matK* or *clpP* in multiple cases, and my amplification strategy sometimes recovered contaminants from these two genes that only become apparent after careful checking (Table A.1). In contrast, I was able to recover the *accD* locus from all taxa sampled here, without any contamination issues (note that *accD* is pseudogenized in some eudicot holoparasites; Wicke et al. 2013). The *accD* locus has only rarely been considered for DNA barcoding studies (Newmaster et al. 2008) and it would be useful to explore further its utility in DNA barcoding surveys that include mycoheterotrophs or holoparasites. It may be a useful supplementary barcoding region to consider including when the focus is not exclusively on green plants.

2.5.8 Conclusions

Mycoheterotrophic lineages may be placed in overall monocot phylogeny using my three-gene data set, generally with moderate to strong support for their placement, despite some issues with contamination, rate elevation and missing genes. My approach to inferring phylogenetic placement is a cost-effective alternative to next-generation sequencing and potentially allows phylogenetic surveys of many heterotrophic taxa. It is also a potentially useful screen in selecting DNAs that will be successful for next-

generation library preparation. At least two of the markers considered here (*accD*, *matK*) may also be useful targets for including heterotrophic plants in DNA barcoding surveys. The underlying alignment I used could be improved upon further, and may be particularly useful to do for other parasitic or mycoheterotrophic lineages, such as Orchidaceae, where I included only a few previously published sequences: this family includes numerous additional origins of full mycoheterotrophy (e.g. Freudenstein and Senyo 2008; Merckx et al. 2013b; Barrett et al. 2014).

Figure 2.1 Three-gene phylogeny of photosynthetic monocots based on a partitioned three-gene maximum likelihood (ML) analysis (for *accD*, *clpP* and *matK*). The analysis includes green orchids but excludes green members of Burmanniaceae (see text for details). Branches with 100% bootstrap support are shown as thick lines; other bootstrap values are indicated beside branches (< 50% support indicated with a dash, '-'). The scale bar indicates estimated number of substitutions per site.

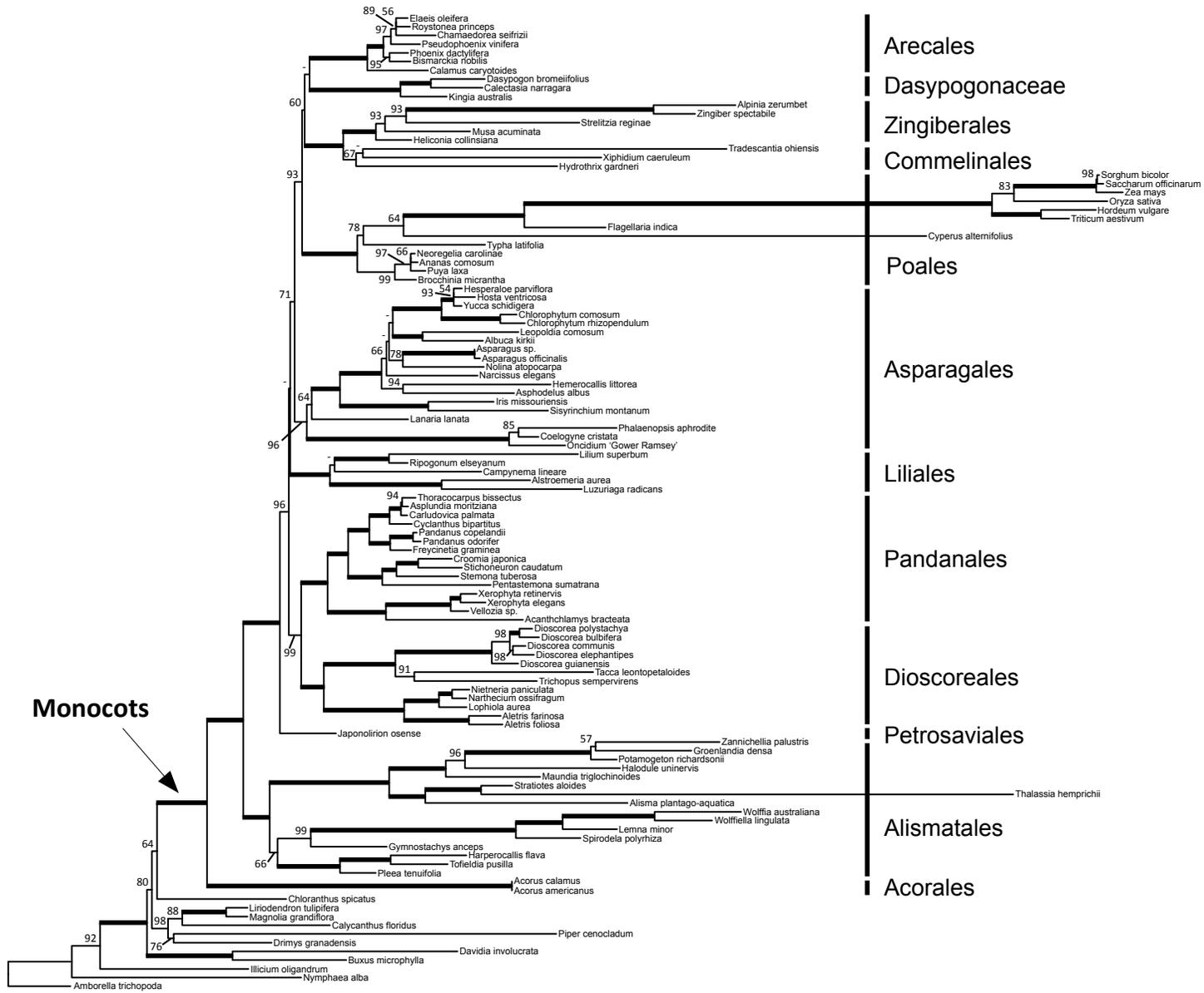


Figure 2.2 Three-gene phylogeny of photosynthetic and fully mycoheterotrophic monocots (all lineages considered simultaneously) based on a partitioned three-gene maximum likelihood analysis (for *accD*, *clpP* and *matK*). Lineages with mycoheterotrophs are indicated in red (asterisks indicate photosynthetic taxa in Burmanniaceae and Orchidaceae, the remainder are full mycoheterotrophs). Branches with 100% bootstrap support are shown as thick lines; other bootstrap values are indicated beside branches (<50% support indicated with a dash, '-'). The scale bar indicates estimated number of substitutions per site.

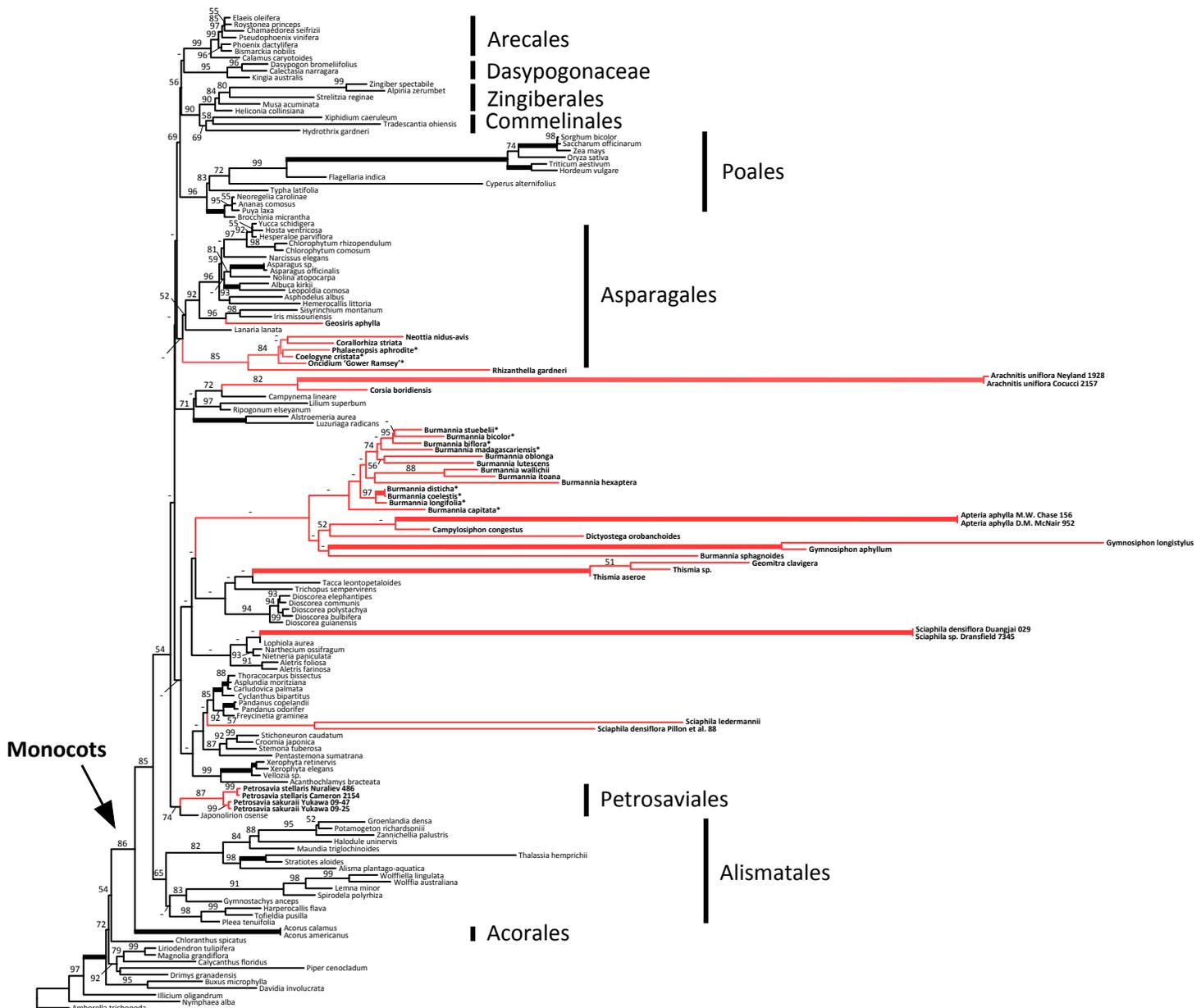
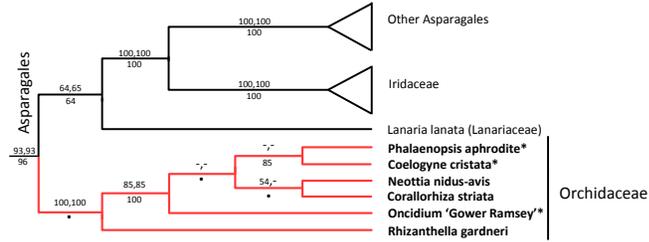
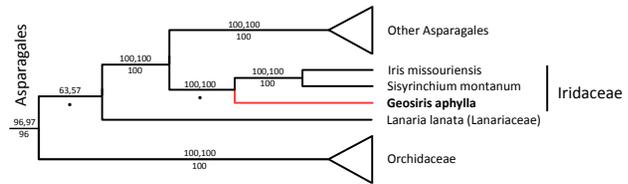
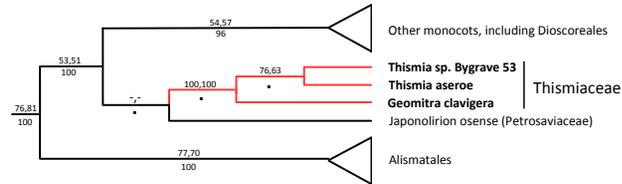


Figure 2.3 Summary of bootstrap support for local placements of monocot mycoheterotrophs in separate analysis of each family, based on partitioned maximum likelihood (ML) analysis of *accD*, *clpP* and *matK* (see Fig. S6-S13 for branch lengths). (a) Iridaceae and Orchidaceae, (b) Burmanniaceae *S.S.*, (c) Thismiaceae, (d) Corsiaceae, (e) Triuridaceae, (f) Petrosaviaceae. The ordinal placement of individual families according to APG (2009) is noted beside each figure (phylogenetic analyses for Iridaceae and Orchidaceae were conducted separately). Major clades collapsed for simplicity. Bootstrap support values noted beside branches: values above branches are with the mycoheterotrophic taxon included (left-hand = partitioned ML analysis, right-hand = unpartitioned ML analysis); those below branches are with it excluded. Bootstrap support values < 50% are indicated with a dash, '-'; a dot indicates that the support value is not applicable (reflecting fewer taxa in analyses with photosynthetic taxa only). Lineages with mycoheterotrophs are indicated in red (asterisks indicate photosynthetic taxa in Burmanniaceae and Orchidaceae, the remainder are full mycoheterotrophs). Voucher names are included for species identified only to genus, or where two samples were included for a species (see also Table A.1).

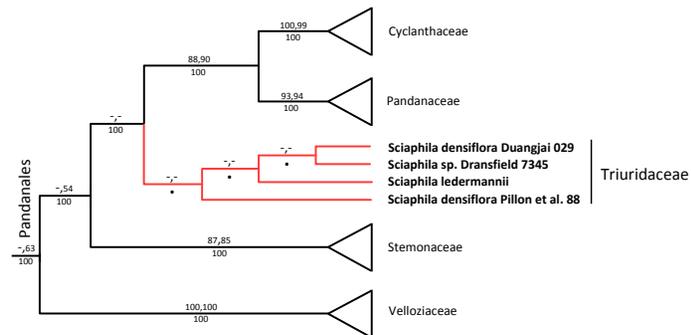
a) Asparagales: Iridaceae & Orchidaceae



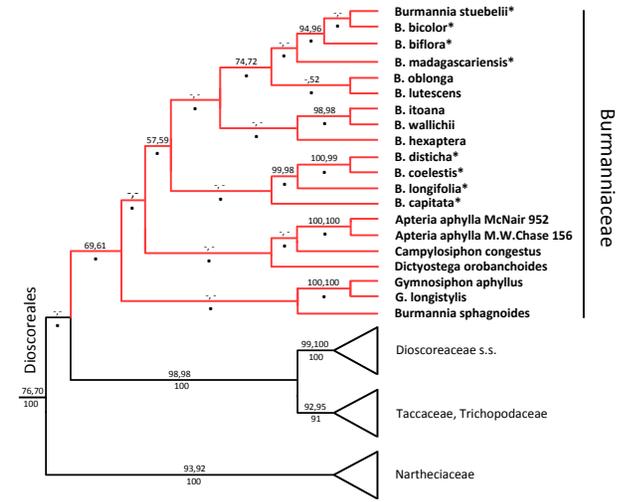
c) Dioscoreales (?): Thismiaceae



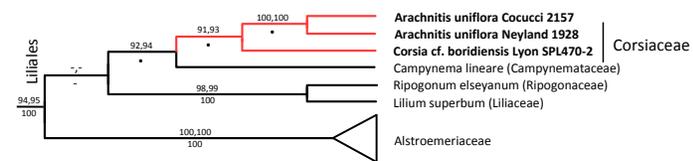
e) Pandanales: Triuridaceae



b) Dioscoreales: Burmanniaceae



d) Liliales: Corsiaceae



f) Petrosaviales: Petrosaviaceae

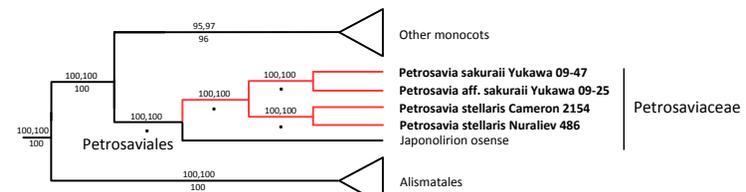


Figure 2.4 Phylogenetic placement of green Burmanniaceae in monocot-wide phylogeny, based on partitioned three-gene maximum likelihood (ML) analysis. The phylogram indicates the placement of green Burmanniaceae (following Merckx et al. 2006; see text for details). Families within Dioscoreales are indicated. The inset figure summarizes bootstrap support values within Dioscoreales (above branches: left-hand = partitioned ML analysis, right-hand = unpartitioned ML analysis; below branches: Burmanniaceae excluded from analysis). Bootstrap support values < 50% are indicated with a dash, '-'; a dot indicates that the support value is not applicable (reflecting fewer taxa in analyses with Burmanniaceae excluded). The scale shows the estimated number of substitutions per site.

Figure 2.5 Phylogenetic placements inferred for individual species of Thismiaceae based on maximum likelihood (ML) analysis (three-gene analyses that include only *accD* for this family, the only plastid gene recovered for it). Summary of results for: (a) *Geomitra clavigera*; (b) *Thismia aseroe*; (c) *Thismia* sp. Left-hand panel: phylograms of shortest trees for partitioned ML analysis (scale bars indicate estimated number of substitutions per site; mycoheterotrophic lineages indicated with blue branches). Right-hand panel: summary of bootstrap support for placements of individual species based on: (i) partitioned ML analysis; (ii) unpartitioned ML analysis; values above branches are with Thismiaceae sequences included; those below are with them excluded. Bootstrap support values < 50% are indicated with a dash, '-'; a dot indicates that the support value is not applicable (reflecting fewer taxa in analyses with photosynthetic taxa only). The scale bar (a-c) indicates the estimated number of substitutions per site.

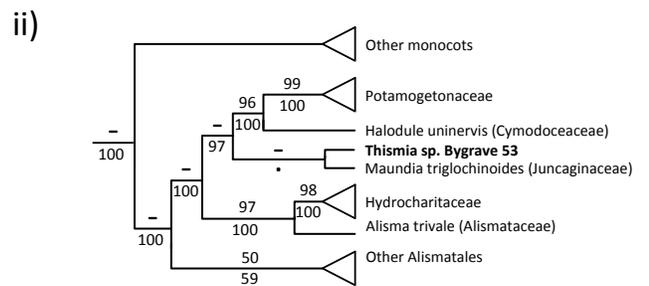
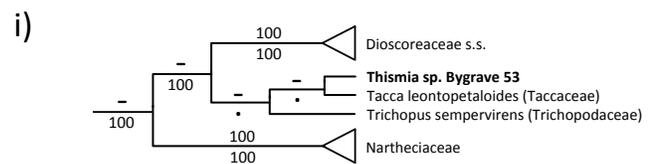
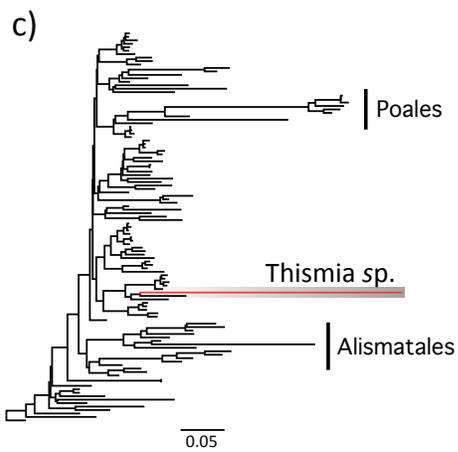
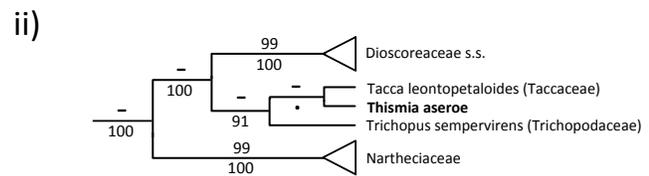
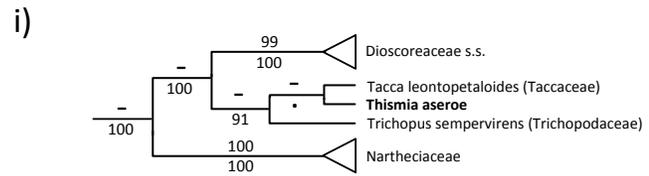
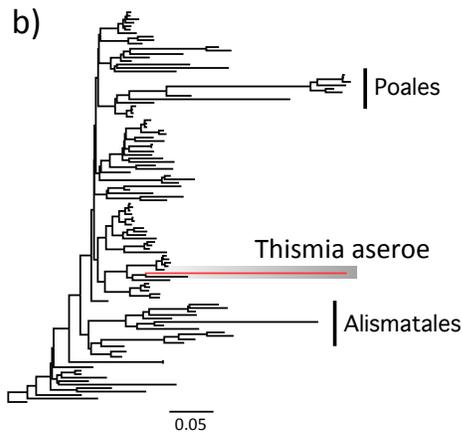
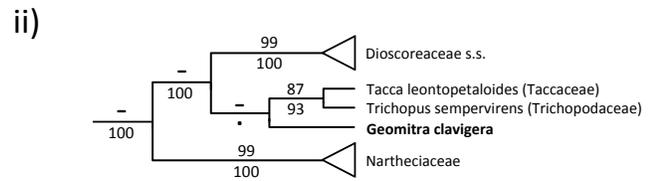
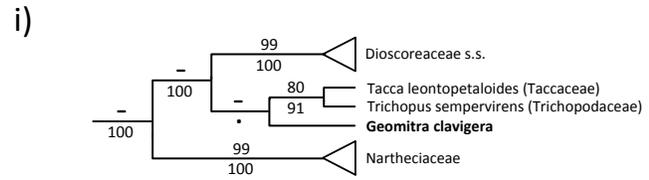
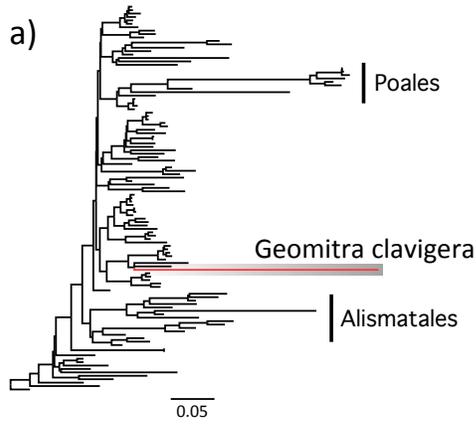
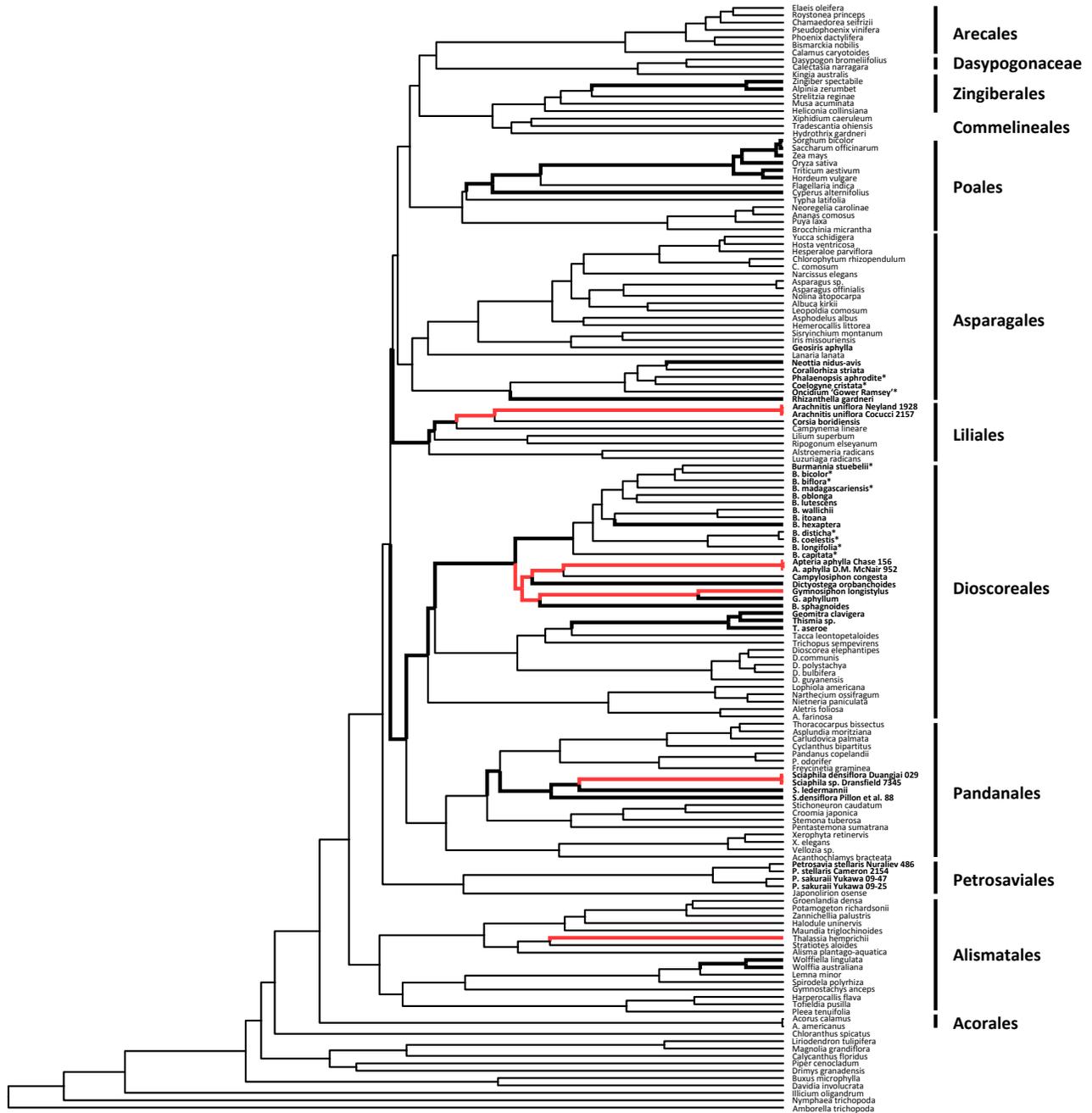


Figure 2.6 Relative substitution rates among green and fully mycoheterotrophic monocot lineages based on Bayesian analyses of a three-gene plastid (*accD*, *clpP* and *matK*) dataset, with a random-local-clock model and a constrained topology (see text for details). Thin branches indicates relative rates of <2.0 substitutions per site and below, thicker branches indicate intermediate relative rates of 2.0-<4.0, and thick red branches indicates those 4.0 and higher (see Fig. S15 for full details). Green species in Burmanniaceae and Orchidaceae are indicated with an asterisk (*), the others in these families are full mycoheterotrophs.



Chapter 3: The highly reduced plastome of mycoheterotrophic *Sciaphila* (Triuridaceae) is colinear with its green relatives and is undergoing strong purifying selection¹

3.1 Summary

The enigmatic monocot family Triuridaceae provides a potentially useful model system for studying the effects of an ancient loss of photosynthesis on the plant plastid genome, as all of its members are mycoheterotrophic and achlorophyllous. However, few studies have placed the family in a comparative context, and its phylogenetic placement is only partly resolved. It was also unclear whether any taxa in this family have retained a plastid genome. Here I used genome survey sequencing to retrieve plastid genome data for *Sciaphila densiflora* (Triuridaceae) and ten autotrophic relatives in the orders Dioscoreales and Pandanales. I recovered a highly reduced plastome for *Sciaphila* that is nearly colinear with *Carludovica palmata*, a photosynthetic relative that belongs to its sister group in Pandanales, Cyclanthaceae-Pandanaceae. This phylogenetic placement is well supported and robust to a broad range of analytical assumptions in maximum likelihood inference, and is congruent with recent findings based on nuclear and mitochondrial evidence. The 28 genes retained in the *Sciaphila densiflora* plastid genome are involved in translation and other non-photosynthetic functions, and I demonstrate that nearly all of the 18 protein-coding genes are under strong purifying selection. My study confirms the utility of whole plastid genome data in phylogenetic studies of highly modified heterotrophic plants, even when they have substantially elevated rates of substitution.

3.2 Introduction

Mycoheterotrophic plants obtain some or all of their nutrients from soil fungi, typically those involved in mycorrhizal interactions with other plants (e.g., Merckx 2013). Merckx and

¹This chapter has been published as ‘Lam, V.K.Y., Soto Gomez, M., 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.’

Freudenstein (2010) counted at least 50 independent origins of full mycoheterotrophy, in which plants have lost the ability to photosynthesize and rely completely on fungal associates. Most of the 400 or so species of full mycoheterotrophs are monocots, a major clade that appears to be particularly prone to this evolutionary transition (Imhof 2010; Merckx and Freudenstein 2010; Merckx et al. 2013a). Of these, about 50 species belong to a single monocot family, Triuridaceae, which is exclusively mycoheterotrophic. The family comprises nine extant genera, and has a pantropical distribution (Maas-van de Kamer and Weustenfeld 1998). Triuridaceae are small achlorophyllous, perennial herbs with tiny flowers and reduced scale-like leaves, found mostly in damp and deep-shaded forest habitats (Furness et al. 2002; Merckx et al. 2013b). There are relatively few collections of this ephemeral and inconspicuous lineage, and their general biology, genomics and evolutionary history remain poorly understood (e.g., Rudall 2003). It seems likely that Triuridaceae experienced an ancient loss of photosynthesis, as a molecular dating analysis indicates that the (non-photosynthetic) crown clade of the family arose around the Cretaceous or Lower Paleocene (Mennes et al. 2013).

Genome survey sequencing techniques (e.g., Cronn et al. 2008) now allow relatively straightforward retrieval of whole plastid genomes (plastomes) of green and heterotrophic plants (mycoheterotrophs and parasitic plants) for use in comparative analysis. Several studies of whole plastid genomes of heterotrophs have recently investigated their molecular evolution and characterized the structural rearrangements and losses that often occur following the loss of photosynthesis (Krause 2008; Barrett and Davis 2012; Wicke et al. 2013, 2014). Genes encoded in the plastome are also expected to show evidence of degradation due to relaxation or release of purifying selection for photosynthesis-related genes (e.g., Barrett et al. 2014). Seven full circular plastomes of mycoheterotrophic taxa have been published to date, representing the five orchids *Corallorhiza striata*, *Epipogium roseum*, *Epipogium aphyllum*, *Neottia nidus-avis* and *Rhizanthella gardneri* (Barrett and Davis 2012; Delannoy et al. 2011; Logacheva et al. 2011; Schelkunov et al. 2015), the monocot *Petrosavia stellaris* (Logacheva et al. 2014), and the liverwort *Aneura mirabilis* (Wickett et al. 2008). These plastomes exhibit variation in patterns of gene loss and retention, gene order and plastome structure. For example, the plastome of *Corallorhiza striata* (Barrett and Davis 2012) is in a relatively early stage of genome degradation, and has retained a gene order consistent with its green relatives, whereas the plastomes of other mycoheterotrophs (e.g., *Petrosavia stellaris*, *Neottia nidus-avis*, *R. gardneri*)

show more complex rearrangements, including substantial reductions in plastome size associated with considerable gene loss.

Based on patterns of gene loss and retention in heterotrophic plants, Barrett and Davis (2012) proposed an ordered trajectory of gene loss in mycoheterotrophs. They hypothesized an initial loss of genes encoding plastid subunits of the NAD(P)H complex, which appears to be involved in responding to photooxidative stress (Martin and Sabater 2010), followed by correlated losses of genes encoding photosynthesis-related protein complexes. Housekeeping genes involved in translation and other non-photosynthetic functions tend to be retained the longest. Genes retained as open reading frames are expected to be under purifying selection, if functional. For example, Barrett et al. (2014) found that housekeeping genes retained in fully mycoheterotrophic *Corallorhiza* are under the same selective regime (i.e., purifying selection) as homologous genes in photosynthetic relatives, consistent with their continued functionality in the plastid, despite the loss of photosynthesis.

Whole plastid genomes retrieved from mycoheterotrophs have also recently been used to determine the phylogenetic placement of several fully mycoheterotrophic lineages with uncertain placement among their photosynthetic relatives. For example, Logacheva et al. (2014) used the 37 protein-coding genes retained in the plastid genome of *Petrosavia stellaris* (Petrosaviaceae; Petrosaviales) to confirm its placement as the sister group of a photosynthetic taxon, *Japonolirion osense*. More recently, Mennes et al. (2015) recovered multiple genes from the plastid genomes of two of the three genera of Corsiaceae (16 and 23 protein-coding genes for *Arachnitis uniflora* and *Corsia* cf. *boridiensis*, respectively, and four rDNA genes from both genera; several tRNA genes were also recovered). These plastid genes placed Corsiaceae as the sister group of Campynemataceae in Liliales and supported the family's monophyly, consistent with nuclear and mitochondrial evidence in the same study. Mennes et al. (2015) showed that all three plant genomes produced a congruent and well-supported picture of phylogenetic relationships of Corsiaceae, which in turn supports the idea that plastid genomes of heterotrophic plants are suitable for large-scale phylogenetic inference, despite extensive rate elevation and gene loss. Mennes et al. (2013) examined the phylogenetic position of Triuridaceae using mitochondrial and nuclear data, and demonstrated that it belongs in the monocot order Pandanales, confirming earlier results based on nuclear 18S rDNA data, mitochondrial *atpA* data and morphology (Chase et al. 2000; Davis et al. 2004; Rudall and Bateman 2006). However, the

precise relationships of the families within the order are still unclear; they were poorly supported in the analyses of Mennes et al. (2013), for example.

Here I report on full plastid genomes and plastid gene sets recovered from *Sciaphila densiflora* (Triuridaceae) and ten related green taxa in Pandanales and Dioscoreales (comprising complete plastid genomes for *Sciaphila* and a green relative, *Carludovica palmata*, and plastid gene sets for nine additional relatives). The data from *Sciaphila* represent the first plastid genome sequences from Triuridaceae. I used these data: (1) to characterize major changes in the plastid genome following the loss of photosynthesis, including gene losses and retentions, and structural rearrangements; (2) to assess whether genes retained in the plastid genome of *Sciaphila* as open reading frames are evolving under purifying selection or some other selective regime; (3) to confirm the placement of Triuridaceae in Pandanales using plastid evidence and to pinpoint its local placement among the four photosynthetic families in the order (Cyclanthaceae, Pandanaceae, Stemonaceae, Velloziaceae), while exploring the effect of different likelihood approaches on phylogenetic inference.

3.3 Materials and methods

3.3.1 Taxon sampling

I generated new plastid genome sequences for ten species in Pandanales and one in Dioscoreales (Table B.1), and added these to published angiosperm plastome data retrieved from GenBank and from monocot-focused matrices presented in Givnish et al. (2010), Barrett et al. (2013) and Mennes et al. (2015) (Table B.1). The 71-taxon matrix included at least one taxon from each of the five families of Pandanales (10 taxa in total), representatives of all major monocot lineages (50 taxa), in addition to representatives of the eudicots, magnoliids and the orders Amborellales, Nymphaeales and Austrobaileyales (i.e., ANA-grade taxa) as outgroups (11 taxa in total).

3.3.2 DNA isolation and library preparation

I isolated DNA using a modified CTAB protocol (Doyle and Doyle 1987; Rai et al. 2003), and prepared whole genome shotgun sequencing libraries using several library preparation kits. I used Bioo Nextflex DNA sequencing kit (Bioo Scientific Corp., Austin, USA) and KAPA LTP Library Preparation kit (KAPA Biosystems, Boston, USA) when >10 ng of starting DNA was

available (I used the Bioo kit for *Sciaphila*). For lower amounts of initial DNA, I used NuGEN Ovation Ultralow Library System (NuGEN Technologies Inc., San Carlos, USA). I sheared DNAs to 400 bp fragments on a Covaris S220 sonicator (Covaris, Inc., Woburn, USA) for library preparation with all three kits, and size-selected all libraries (550-650 bp fragments). For the Bioo kit I size-selected using a 2% agarose gel, purifying the resulting DNA using a Zymoclean gel recovery kit (Zymo Research, Irving, USA). For the Kapa and NuGEN kits I used magnetic bead size selection (Agencourt AMPure XP magnetic beads, Beckman Coulter Genomics, Brea, USA). For quality control, I quantified all libraries by Qubit (Qubit fluorometer, ThermoFisher Scientific, Waltham, USA) to ensure a minimum DNA concentration of 0.5 ng/ul. Library fragment sizes were verified by Bioanalyzer (Agilent Technologies, Santa Clara, USA), and concentrations were measured by qPCR on an iQ5 real-time system (Illumina DNA standard kit, KAPA Biosystems, Boston, USA; Bio-Rad Laboratories, Inc., Hercules, USA). Individual libraries were multiplexed (Cronn et al. 2008) in several lanes on an Illumina HiSeq 2000 (Illumina, Inc., San Diego, USA) and sequenced as 100 bp paired-end reads.

3.3.3 *De novo* contig assembly, plastid gene annotation and plastome reconstruction

Illumina reads were processed with CASAVA 1.8.2. (Illumina, Inc., San Diego, US) to sort the multiplexed data by taxon. To obtain contigs, I performed *de novo* assemblies for each individual taxon using CLC Genomics Workbench v.6.5.1 (CLC bio, Aarhus, DK) with default settings. I selected all contigs >500 bp in length with >20X coverage, and used a custom Perl script (Daisie Huang, University of British Columbia) to BLAST contigs against a local database (Altschul et al. 1990) of plastid genes from *Dioscorea elephantipes* (GenBank accession NC_009601.1) in order to remove mitochondrial and nuclear contigs. For *Cyclanthus*, *Freycinetia*, *Sararanga*, *Croomia*, *Pentastemona*, *Stemona*, *Stichoneuron*, *Xerophyta* and *Lophiola* I annotated plastid genes using DOGMA (Wyman et al. 2004), manually inspecting gene and exon boundaries in Sequencher 4.2.2. (Gene Codes Corporation, Ann Arbor, US) using *Phoenix dactylifera* (NC_013991) and *Dioscorea elephantipes* to annotate start/stop codons and introns for each protein-coding gene. I exported final gene sets (coding regions) as individual FASTA files for each taxon. For *Carludovica* and *Sciaphila*, I assembled full circular plastomes, designing primers using Primer3 (Untergasser et al. 2007; Koressaar and Remm 2007) to bridge gaps between contigs or to verify contig overlap. Amplification of these regions was performed using

Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific, USA), followed by sequencing using BigDye Terminator v3.1 sequencing chemistry (Applied Biosystems, Inc., Foster City, USA) on an Applied Biosystems 3730S 48-capillary DNA analyzer (Applied Biosystems, Inc., Foster City, USA). I used Sequencher to produce a consensus plastome sequence by assembling contigs produced in CLC together with the Sanger-derived sequences, and annotated the consensus sequences in DOGMA, as discussed above. For *Sciaphila*, I additionally searched all intergenic spacer regions for potential pseudogenes. I used OGDRAW (Lohse et al. 2013) to generate the two plastome maps.

3.3.4 Data matrix construction and sequence alignment

I added data for ten newly sequenced species in Pandanales and one species in Dioscoreales to published data (Table B.1) for 82 plastid genes (78 protein coding genes and four ribosomal DNA genes, with 71 taxon terminals per file; missing genes were represented as blanks). I aligned each gene file in Se-AL v.2.0a11 (Rambaut 2002) using criteria laid out in Graham et al. (2000), staggering gene regions that were difficult to align (e.g., Saarela and Graham 2010). I verified that alignments for protein-coding genes were maintained as open reading frames, and concatenated all individual gene alignments into a single 102,897 bp matrix (derived from 67,506 bp of unaligned plastid sequence data in *Carludovica palmata*, for reference), including the inverted repeated regions only once. To check for compilation errors in the final matrix, I exported the concatenated gene sequences for each taxon, and used Sequencher to compare them to the original individual taxon files (none were found). I retrieved plastid gene *ycf1* for most taxa (the gene is absent in *Sciaphila*, see below), but did not include it in the final matrix due to difficulties in alignment.

3.3.5 Phylogenetic inference

I analyzed the data using parsimony and maximum likelihood (ML) methods. For the parsimony analysis, I ran a heuristic parsimony search for shortest trees in PAUP* v4.0a134 (Swofford 2003) using tree-bisection-reconnection branch swapping (TBR) and 1000 random stepwise addition replicates, holding 100 trees at each step, and otherwise using default settings. I estimated branch support with a bootstrap analysis (Felsenstein 1985), using 1000 replicates, with 100 random addition replicates per bootstrap replicate for the parsimony analysis (for the

ML analyses, see below). For all bootstrap analyses performed here, I considered well-supported branches to have at least 95% bootstrap support, and poorly supported branches to have less than 70% support, following Zgurski et al. (2008).

For the ML analyses, I first conducted heuristic searches of the DNA sequence data using nucleotide substitution models with RAxML v.7.4.2 (Stamatakis 2006), using a graphical interface for it (Silvestro and Michalak 2012). I ran three variant analyses, one with all the data unpartitioned, a second with the data partitioned by codon position (a “codon” partitioning scheme, with rDNA genes considered as additional data partitions), and a third with the data partitioned by both gene and codon positions (“GxC”, or gene by codon partitioning); see below for how the final partitioning schemes were set up. I analyzed translated protein-coding genes with amino-acid substitution models in RAxML with unpartitioned data, and with the data also partitioned by gene (described below). Finally, I analyzed the unpartitioned DNA sequence data using a codon-based substitution model implemented in Garli 2.0 (Zwickl 2006). For all ML analyses I conducted 20 independent searches for the best tree, and estimated branch support using 500 bootstrap replicates, using GTRGAMMA or PROTGAMMA approximations for the analyses based on nucleotide/codon vs. amino-acid substitution models, respectively (I used a subset of taxa for the analysis using the codon-based substitution model, because of computational limitations, see below). Each bootstrap analysis used the same substitution models as the searches for best trees.

For the unpartitioned ML analysis of the DNA sequence data, I used jModeltest 2.1.3 (Darriba et al. 2012; Guindon and Gascuel 2003) to find the optimal DNA substitution model using the Bayesian Information Criterion, BIC (Schwarz 1978). This chose GTR+G as the best model. For the various partitioned analyses I used PartitionFinder v.1.1.1 and PartitionFinderProtein 1.1.0 (Lanfear et al. 2012) to combine partitions that did not have significantly different DNA or amino-acid substitution models, using the hierarchical clustering algorithm and the BIC criterion, and used the final data partitioning schemes for phylogenetic inference. For the codon partitioning scheme, I allocated nucleotides in protein-coding genes according to whether they belong to the first, second or third codon position, and assigned four additional initial partitions for the plastid rDNA genes (for a total of seven initial partitions). PartitionFinder retained four partitions (one for each codon position, and one for all four rDNA genes), with GTR+G identified as the best model in each case (Table B.2a). For the GxC (gene

by codon) partitioning scheme for the DNA sequence data, I first partitioned the matrix both by gene (treating the trans-spliced exons of 5'-*rps12* and 3'-*rps12* as two genes, operationally) and by codon position (first, second and third position for the protein coding genes, leaving the rDNA genes as distinct partitions), for a total of 241 initial partitions. PartitionFinder retained 12 final partitions, with GTR+ G or GTR+ G +I selected as the best DNA substitution model in all cases (Table B.2b). I used the GTR + G model for individual partitions in subsequent phylogenetic analysis, as the I parameter (invariant sites) may be adequately accommodated by the gamma parameter (Yang 2006). For the amino-acid data, PartitionFinderProtein retained 71 partitions from the original 80 partitions (partitioned by gene, again considering 5'-*rps12* and 3'-*rps12* as two genes), and inferred a range of optimal amino-acid models that we used in subsequent phylogenetic inference (Table B.2c).

I also analyzed the nucleotide sequence dataset using an unpartitioned codon-based substitution model. For this analysis I applied the 6-rate (GTR) codon model with F3x4 codon frequencies and one dN/dS parameter, using Garli 2.0 (Zwickl 2006) on the CIPRES Portal (Miller et al. 2010). Because of severe computational constraints for the latter method, I estimated the bootstrap support for two subsets of this matrix, one including only taxa in Pandanales and Dioscoreales (12 taxa), and a second with additional representatives chosen from most major monocot lineages (26 taxa, see below).

3.3.6 Model-based tests of selective regime in plastid genes

I used the CodeML module in PAML4.8 (Yang 2007) to assess changes in selective regime in 18 protein-coding genes retained as open reading frames in the *Sciaphila* plastome (Table 3.1). The objective was to test hypotheses of different ω -values (ω is the ratio of nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site) for *Sciaphila* (indicated below as “MHT,” an abbreviation for “mycoheterotroph”), compared to photosynthetic (“green”) outgroups. I built two codon-based “branch” models, which can detect differences in selection regimes in particular lineages (Yang 2007). In the simplest model (M0, one ratio), all branches evolve under one ω -ratio (i.e., $\omega_{\text{MHT}} = \omega_{\text{green}}$; see Table B.3). In the alternative model (M1, two ratios), *Sciaphila* was allowed to evolve under a different ω -ratio than the green taxa (i.e., two ratios allowed, ω_{MHT} and ω_{green}). I also compared “branch-site” models to survey for positive selection that may affect only a few sites in a prespecified lineage

(Yang 2007). For this test I specified *Sciaphila* as the foreground lineage and all green taxa as background lineages. For the null model (H0), the foreground branch (ω_2) was fixed to $\omega_2=1$, allowing codons on this branch to evolve neutrally. In the alternative model (H1), $\omega_2 > 1$ was estimated, allowing positive selection in the foreground lineage.

To implement both the branch and branch-site models, I removed taxa in alignments lacking sequence data and regions with indels that resulted in missing data for 90% or more of the taxa. I used the 26-taxon best tree inferred from the codon-based substitution model ML analysis (see Fig. B.6) as a constraint tree, but with branch lengths generated in PAML, and pruned any taxa missing for individual genes. I ran all models on individual genes using the F3x4 codon frequency model. I used the likelihood ratio test (LRT) statistic $-2(\ln L_{M0/H0} - \ln L_{M1/H1})$ to compare the fit of M0 vs. M1 (branch models) or H0 vs. H1 (branch-site models), and calculated *P*-values based on a χ^2 test with one degree of freedom. I used a Bonferroni correction to account for multiple tests conducted on the same data (Anisimova and Yang 2007), and considered tests significant if the *P*-value was $< \alpha / m$, where *m* is the number of branches being tested using the same data ($m=2$ for both models). I identified any sites undergoing positive selection in the branch-site model using the Bayes empirical Bayes (BEB) test included in the CodeML package.

3.4 Results

3.4.1 Full circular plastomes

I assembled the plastid genome of *Carludovica palmata* (GenBank accession NC026786.1, Fig. 3.1) as a circular sequence of 158,545 bp, with an average of $\sim 734.5X$ coverage from ~ 21.24 million paired-end reads. The *Carludovica* plastome is comparable to those of other angiosperms in size and organization. It has the typical quadripartite structure of plant plastid genomes, with a 87,041 bp large single copy (LSC) region, a 18,366 bp small single copy (SSC) region, two inverted repeat (IR) regions of 26,569 bp, and has the same gene order as *D. elephantipes* (Hansen et al. 2007). I assembled *Sciaphila densiflora* as a circular sequence with a predicted minimum length of 21,485 bp, and an average of $\sim 50X$ coverage (GenBank accession KR902497.1, Fig. 3.2) from ~ 17.03 million paired-end reads. The DNA extractions for

Carludovica and *Sciaphila* were both done using fresh plant material (Edith Kapinos, Royal Botanic Gardens, Kew, pers. comm.), and so the order of magnitude lower coverage for the plastome in the mycoheterotroph compared to the autotroph may be consistent with substantially fewer plastid genomes per plant cell for it. Two neighboring sectors of the assembled *Sciaphila* plastome had substantially higher coverage than the remainder (Fig. 3.3), consistent with them being repeated regions. One sector with ~4X the average read depth (214X coverage) includes *rrn4.5*, *rrn5* and part of *rrn23*; the other includes the remainder of *rrn23* and had only ~twice the average read depth (93X coverage). It is possible that they represent a short series of tandem repeats, but they could also incorporate a reduced and cryptic inverted repeat. I was not able to confirm the number or arrangement of these putative repeats because I had a limited amount of DNA for experimental confirmation. No genes from the small single copy region that is typical of other angiosperm plastomes (e.g., Fig. 3.1) were recovered. The gene order depicted in Fig. 3.3 likely represents the correct order at the ends of any repeated regions, as it is consistent with my ability to connect contigs using direct sequencing (Fig. 3.3; the higher read-depth sectors are also indicated in the *Sciaphila* genome map). The possibility that high-depth regions instead represent inserts elsewhere (e.g., in the mitochondrial genome) cannot be excluded, although I did not observe obvious sequence variation in these genes that might be indicative of divergent copies in other genomes. The stoichiometry of the coverage levels relative to the remaining plastid contigs is suggestive of replication within the plastid genome rather than elsewhere (i.e., if they are located elsewhere, the coverage depth would not necessarily be near-integer multiples of the rest of the plastome). Also, if there are repeats, I doubt that I missed additional intervening genes (or pseudogenes), as BLAST-based attempts to recover genes missing from the *Sciaphila* genome were not successful. Finally, the high degree of colinearity demonstrated here between *Sciaphila densiflora* and its close photosynthetic relative, *Carludovica palmata* also supports the idea that I recovered the full complement of retained genes in the mycoheterotrophic species (Fig. 3.3).

My current model of the *Sciaphila* plastid genome provides a minimum size estimate for it (ignoring the possibility that some regions are duplicated), and is ~13.6% of the size of *Carludovica* (discounting duplications in the IR of the latter), with 28 plastid genes retained in total (Table 3.1). Considering unique sequences only, the coding sequences (proteins, rDNA and tRNA genes) account for 68.7% of the *Sciaphila* plastome, whereas 58.0% of the *Carludovica*

plastome is composed of coding sequence. The average GC content is also marginally higher in *Sciaphila* (39.9%) than in *Carludovica* (36.7%). Most of the 28 retained genes in *Sciaphila* are involved in protein synthesis; 10 code for small ribosomal proteins and five for large ribosomal proteins, all four rDNA loci are retained, along with six tRNA loci (Table 3.1). The remaining loci are *accD* (which codes for a subunit of acetyl-coA carboxylase or ACCase), *clpP* (which codes for a proteolytic subunit of the enzyme Clp-protease), and *matK* (the maturase gene for group-IIA plastid introns); see Wicke et al. (2011) for further details on gene function. All genes except *trnC*-GCA and *trnW*-CCA are transcribed on one strand (Fig. 3.2). I did not recover any pseudogenes (or at least all of the genes in Table 3.1 were open reading frames). Gene order in *Sciaphila* is nearly colinear with that in *Carludovica*, although I infer an inversion of a block comprising *rps18*, *trnW*-CAA and *accD* in the LSC of *Sciaphila* (Fig. 3.3) and a block comprising 3'-*rps12* and *rps7* in what was the IR, assuming deletion of intervening sequences in the original IR copies (Fig. 3.3). Genes inferred to be lost from *Sciaphila* include those coding for photosynthesis-related protein subunits (photosystems II and I, cytochrome *b₆f* complex and ATP synthase), all plastid-encoded RNA polymerase (PEP) loci, the majority of the tRNA loci, several genes involved in protein synthesis (ribosomal proteins *rps15*, *rps16*, *rpl22*, *rpl23*, *rpl32*, *rpl33* and *infA*), and two genes with uncertain function (*ycf1* and *ycf2*).

3.4.2 The phylogenetic position of *Sciaphila* (Triuridaceae)

I inferred *Sciaphila* to be a member of Pandanales in all analyses here, with strong support (Figs. 3.4, B.1-B.7). The monophyly of the order and its sister-group relationship to Dioscoreales were also confirmed with strong support in all likelihood analyses. The position of *Sciaphila* within Pandanales was completely consistent across all six likelihood analyses, and was also generally strongly supported: a clade comprising Triuridaceae and Cyclanthaceae-Pandanaceae was recovered with 95-99% bootstrap support in the DNA-based ML analyses that used nucleotide substitution models (Figs. 3.4, B.1-B.3), with 88-91% bootstrap support in the amino-acid analyses (Figs. B.4, B.5) and with 86-87% bootstrap support in the analyses that used a codon-based substitution model (Fig. B.6). The clade comprising Cyclanthaceae and Pandanaceae had 88-100% bootstrap support across likelihood analyses (Figs. 3.4, B.1-B.6). Stemonaceae were recovered as the sister group of this clade, and Velloziaceae (represented by *Xerophyta*) were supported as the sister group of all other Pandanales. The latter relationships all had strong

support in all likelihood analyses (97-100%; Figs. 3.4, B.1-B.6; note that one of the analyses shown in Fig. B.6 considered only taxa in Dioscoreales and Pandanales).

The sole analysis that yielded a different topology concerning the placement of *Sciaphila* was the parsimony analysis, which recovered it as sister to *Xerophyta*, but with poor bootstrap support (65%, Fig. B.7). The long branch typical of *Sciaphila* in the likelihood analyses was notably not evident in the parsimony analysis (cf. Figs. B.1-B.6; Fig. B.7). This analysis also supported the monophyly of Pandanales, but with only moderate support (71%), suggesting that *Sciaphila* destabilizes the support for relationships when included in parsimony analysis. I tested this by excluding *Sciaphila* and re-running the parsimony analysis; the underlying relationships were not affected, but support values within Dioscoreales and Pandanales improved dramatically (Fig. B.7). There were no major differences in monocot relationships across the various likelihood and parsimony analyses.

3.4.3 Tests of selection

A ω value greater than one is interpreted as evidence for positive selection, $\omega < 1$ suggests purifying (negative) selection, and $\omega \approx 1$ indicates neutral evolution (Zhang et al. 2005). Under the branch models (comparing the one-ratio model M0, and the two-ratio model M1), the M0 model fit the data better for 15 of the 18 genes tested (trans-spliced *rps12* was treated as two genes, operationally; these portions are listed separately in Table B.3, but are lumped as one gene in the discussion below) indicating that these retained genes in *Sciaphila* are evolving at the same ω rate as in the green outgroups (Table B.3). These 15 genes appear to be highly conserved and under purifying selection in the analyzed taxa ($0.096 < \omega < 0.368$). The two-ratio model (M1) was a better fit for *clpP*, *rpl14*, and *rps7* after Bonferroni correction, suggesting that these genes are under a significantly different selective regime in *Sciaphila* than in the green taxa. The *rps7* locus of *Sciaphila* ($\omega_{\text{MHT}} = 0.611$) approached the expectations for neutral evolution ($\omega \approx 1$), compared to evidence of strong purifying selection ($\omega = 0.203$) in green taxa. Although I detected significant differences in ω rates for *clpP* and *rpl14*, these two genes are still predicted to be experiencing purifying selection in *Sciaphila*, although this may have also been relaxed ($\omega_{\text{MHT}} = 0.288$ and $\omega_{\text{MHT}} = 0.240$, respectively for *clpP* and *rpl14* in *Sciaphila*; the corresponding values for green taxa are: $\omega_{\text{green}} = 0.154$ and $\omega_{\text{green}} = 0.096$).

In the branch-site tests, the null model of neutral evolution (H₀), which allows no sites to be under positive selection, appeared to fit the data better for 15 of the 18 genes tested (Table B.4). An alternative model of positive evolution (H₁), which allows some sites to be under positive selection, was a better fit for *accD*, *rpl20* and *rps18*, although the result was not significant for *rps18* after Bonferroni correction. The BEB test found two positively selected sites in each of the three genes. I located these sites in the alignments and speculate that this result is due to alignment difficulties for these parts of the genes, which are quite variable in *Sciaphila*. After staggering these hard-to-align sections in a revised alignment (effectively removing them from consideration) and re-running the PAML tests, I found no evidence of positive selection elsewhere in these genes (Table B.4). I therefore suspect that the positive selection results are artifacts. To ensure that these re-alignments did not affect the phylogenetic placement of *Sciaphila*, I substituted the re-aligned versions of these three genes in the original data matrix and re-ran two ML phylogenetic analyses for the nucleotide data, using nucleotide-substitution models, one with unpartitioned data and the second with the GxC partitioning scheme (see Table B.2d for partitioning scheme), repeating the phylogenetic procedures described above. These minor alterations in the alignment did not affect the placement or support for the placement of *Sciaphila* (<5% difference in support values; cf. Figs. B.1 vs. B.8; B.3 vs. B.9).

3.5 Discussion

3.5.1 Gene loss and retention in *Sciaphila* (Triuridaceae)

The retention of only 28 genes in total (18 protein-coding genes, four rDNA genes and six tRNA genes) makes the *Sciaphila densiflora* plastid genome one of the smallest ones known in land plants, at least in terms of the number of genes (tables 1, 2). *Sciaphila* therefore appears to be in the late stages of plastome reduction, and may be well on its way to full gene loss (Wicke et al. 2013). Photosynthetic land plants have a remarkable degree of conservation of gene content, and considering the non-duplicated genes in the IR, angiosperms typically have 79 protein-coding genes, four rDNA genes and 30 tRNA genes (e.g., Wicke et al. 2011). *Carludovica palmata* exemplifies a typical angiosperm plastome arrangement (Fig. 3.1, Table 3.1). In contrast, heterotrophic plants may show extensive gene loss, reflecting relaxed evolutionary constraints

following the loss of photosynthesis (Krause 2008; Barrett and Davis 2012; Wicke et al. 2011). For example, the mycoheterotrophic liverwort *Aneura mirabilis* has retained 125 genes, including duplicates in its IR region, and the holoparasitic species *Epifagus virginiana* and the mycoheterotrophic orchid *Rhizanthella gardneri* (Table 3.2) have retained only 55 genes and 37 genes, respectively (Wolfe et al. 1992; Delannoy et al. 2011). Although *Rafflesia lagascae* may have lost its plastome entirely (Molina et al. 2014), as in multiple lineages of secondarily heterotrophic unicellular eukaryotes (e.g., Abrahamsen et al. 2004; Smith and Lee 2014; Janouškovec et al. 2015), plastome loss remains to be definitively demonstrated in any land plant. Outside the land plants, the parasitic green alga *Helicosporidium* sp. has 54 genes (de Koning and Keeling 2006), and the malarial parasite *Plasmodium falciparum* has 68 genes (Wilson et al. 1996).

The common gene set retained across mycoheterotrophs includes ribosomal proteins (*rpl2*, 14, 16 and 36; *rps2*, 3, 4, 7, 8, 11, 14 and 19; the sequence for *rps18* in *Neottia* has a single in-frame internal stop codon which may be RNA edited, so this may also be consistent with a retention of *rps18*), other protein-coding genes (*accD* and *clpP*), all four ribosomal DNA loci (*rrn4.5*, 5, 16 and 23) and four transfer RNAs (*trnC*-GCA, E-UUC, I-CAU and *fM*-CAU). A slightly smaller set of genes is retained in heterotrophic plants in general, i.e., including holoparasitic plants (Li et al. 2013; Wicke et al. 2013; Barrett et al. 2014); *rps3*, *rps19* and *trnC*-GCA are not part of this broader list because they have been lost in some taxa. *Sciaphila* is evidently near the end of the degradation trajectory proposed by Barrett and Davis (2012) and Barrett et al. (2014), in which the only genes retained are those involved in housekeeping activities.

Commonly retained plastid genes in heterotrophs whose gene products are not involved in photosynthesis or translation include *accD*, *clpP* and *matK*. However, these loci have been lost individually from the plastid genome of at least one heterotrophic lineage (e.g., Delannoy et al. 2011; Logacheva et al. 2011; Wicke et al. 2013), and in some autotrophs; *accD* has been lost in several lineages of photosynthetic angiosperms (Jansen et al. 2007), and *clpP* in several lineages of photosynthetic eudicots (see Straub et al. 2011). Losses in photosynthetic lineages could be explained by functional transfer of the gene to the nuclear genome, which likely occurred for *accD* in Campanulaceae (Rousseau-Gueutin et al. 2013), for example, or by replacement of the plastid function by a distinct nuclear gene product with similar function, as with replacement of

plastid-encoded RNA polymerases, PEPs, with nuclear-encoded RNA polymerases, NEPs (e.g., Zhelyazkova et al. 2012).

The loss of the majority of the plastid tRNA genes in *Sciaphila* may indicate extensive modification in the functioning of its plastome translation apparatus. Plastid tRNAs may be replaced over evolutionary time by tRNAs imported from the cytosol (e.g., Alkatib et al. 2012), or may be functionally replaced by other tRNAs via the “superwobbling” effect (see Rogalski et al. 2008a). One tRNA gene, *trnE-UUC*, has been found to be retained in the plastid genomes of all heterotrophic plants to date (see Table 3.2 for mycoheterotrophs). Barbrook et al. (2006) hypothesized that this would be the last gene to be retained in the plastid genome of any heterotrophic plant, because of its essential additional role in heme biosynthesis. The precursor of heme, aminolevulinic acid (ALA), is synthesized in land-plant plastids via the C5-pathway, which begins with the ligation of plastid tRNA^{Glu} to glutamate. Secondarily heterotrophic eukaryotes that lack plastid genomes (unicellular eukaryotes: Abrahamsen et al. 2004; Smith and Lee 2014; Janouškovec et al. 2015; possibly the holoparasite *Rafflesia*: Molina et al. 2014), may either import a viable nuclear or mitochondrial tRNA^{Glu} into the plastid, or instead synthesize ALA via the Shemin pathway in mitochondria (Oborník and Green 2005; Barbrook et al. 2006; Smith and Lee 2014).

3.5.2 General retention of colinearity despite genome reduction in *Sciaphila*

Sciaphila exhibits relatively few changes in gene order, despite extensive gene loss (Table 3.2; Figs. 3.2, 3.3). My minimum size estimate of the *Sciaphila* plastome (21,485 bp) is smaller than most previously published heterotrophic plant genomes, with the exception of the orchid *Epipogium roseum*, which has a genome size of 19,047 bp (Schelkunov et al. 2015), although undocumented repeats may add to its size, as noted above. The non-repeated content of the *Sciaphila* plastid genome is smaller than that of the parasitic green alga *Helicosporidium*, with a genome size of 37,454 bp (de Koning and Keeling 2006) and the malarial parasite *Plasmodium falciparum*, with a genome size of 34,682 bp (Wilson et al. 1996), although the latter genome includes an inverted repeat.

Heterotrophic plant lineages often exhibit extensive changes in their plastomes in terms of gene order, compared to the extensively conserved genomes of photosynthetic land plants (Palmer and Stein 1986). Relaxed selective constraints (e.g., relaxation of selection against

repetitive elements that can trigger rearrangements) may contribute to plastid genome rearrangements in heterotrophic lineages (e.g., Wicke et al. 2013). Rearrangements may also be exacerbated by extensive modification or loss of the inverted repeat (IR), which may have occurred in *Sciaphila* (Figs. 3.2, 3.3). The IR is thought to act as a stabilizing factor during recombination-dependent replication of the plastome (e.g., Magee et al. 2010; Maréchal and Brisson 2010; Wicke et al. 2011; Sabir et al. 2014). A range of structural alterations have been observed in mycoheterotrophic monocots, including those that are apparently in the early stages of genome reduction, such as *Neottia* and *Corallorhiza* (Logacheva et al. 2011; Barrett and Davis 2012), which mostly show only gene loss, to those with more extensive and large-scale rearrangements, such as *Epipogium aphyllum*, which has lost its small single copy region (Schelkunov et al. 2015), and *Petrosavia*, which has multiple major rearrangements (Logacheva et al. 2014). In contrast, *Sciaphila* is largely colinear with green angiosperms, such as its close relative *Carludovica* in Cyclanthaceae (Fig. 3.3). Almost all of the differences can be explained by gene loss events; retained genes are shown in the figure (as numbered labels). The substantial colinearity observed here between *Sciaphila* and photosynthetic relatives (Fig. 3.3), ignoring gene losses, might point to retention of a cryptic IR (see above) as a stabilizer of genome structure.

3.5.3 Model-based tests of selective regime in plastid genes

Generally relaxed functional constraints resulting from the loss of photosynthesis may also affect plastid-encoded housekeeping genes (e.g., McNeal et al. 2009; Young and dePamphilis 2005). I detected little evidence of this effect here (Table B.3), as most retained genes in the *Sciaphila* plastome are inferred to be under strong purifying selection. Barrett et al. (2014) also found that housekeeping genes retained in the plastome of the fully mycoheterotrophic orchid *Corallorhiza* were under purifying selection, and the ω -ratios they observed were not significantly different from those of homologous genes in green relatives. Plastid ribosomal protein and tRNA genes are likely retained in the long term because of the general retention of *accD* and *clpP* (two non-photosynthesis related genes) in land-plant plastid genomes (e.g., Delannoy et al. 2011), which occurs regardless of autotrophy or heterotrophy status. As persistence of any essential plastid encoded-genes requires a functional apparatus for translation, translation apparatus genes would in turn be under strong purifying selection to be retained. Delannoy et al. (2011) hypothesized

that plastid-encoded *accD* and *clpP* are required for essential plastid-mediated regulation of the production of their respective multisubunit complexes. *ClpP* is part of the multi-subunit Clp protease, and *accD* codes for the β -carboxyltransferase subunit of ACCase; this subunit regulates fatty-acid biosynthesis in the plastid (see also Bungard 2004, who hypothesized a similar regulatory role for *accD*). *AccD* and *clpP* have both been lost from the plastid genomes of several plant lineages (see above), but these losses appear to be unrelated to the loss of photosynthesis, as they all occurred in green lineages.

Although the branch models test indicates elevated rates of nucleotide substitution in *Sciaphila* compared to homologous genes in green relatives (data not shown, but also evident in our ML phylograms, e.g., Fig. B.3), there appears to have been a proportional increase in both nonsynonymous and synonymous substitution rates, given that the ω -values of most retained genes are consistent with their photosynthetic relatives. Delannoy et al. (2011) also observed this pattern in the mycoheterotrophic orchid *Rhizanthella*. While my findings support the continued functionality of all or most retained genes in *Sciaphila*, recent losses of function cannot be completely ruled out, as there may be a lag between the loss of function/loss of purifying selection and our ability to detect it through pseudogenization, etc. (e.g., Leebens-Mack and dePamphilis 2002). A possible example of this phenomenon concerns the ribosomal protein *rps7*, which is retained here (Table 3.2) and in all heterotrophic plant plastomes sequenced to date (Li et al. 2013; Wicke et al. 2013; Barrett et al. 2014). This locus may be in the early stages of degradation in *Sciaphila*, as it has an ω -rate three times that of green taxa (*rps7*, $\omega = 0.611$; Table B.3), approaching $\omega \approx 1$, the rate expected under neutral evolution. The final expected fate of genes no longer under selective retention is the accumulation of stop codons and indels, leading eventually to complete deletion from the plastome (e.g., Barrett and Davis 2012).

As the genes retained in *Sciaphila* are mostly housekeeping genes involved in basic plastid processes, I did not expect to find substantial evidence of positive selection. My initial findings for evidence of positive selection using the branch-site model for three genes (*accD*, *rpl20* and *rps18*) are probably artifacts of alignment difficulties in highly variable regions, highlighting the sensitivity of this test to slight misalignment. Previous studies have identified sites under positive selection in plastid genes of heterotrophic plants using branch-site models. For example, Barrett et al. (2014) found evidence of positive selection in *atp* genes retained in fully mycoheterotrophic *Corallorhiza*. The *atp* gene complex plays a critical role in

photosynthesis, and changes in selective regime may be due to genes having additional or modified plastid functions (Wicke et al. 2013; Barrett et al. 2014). McNeal et al. (2009) found that positive selection may be acting on a codon of gene *matK* retained in the plastome of the parasitic plant *Cuscuta nitida*. The gene product of this locus is likely involved in splicing seven plastid group IIA introns (Zoschke et al. 2010); *Cuscuta nitida* has lost six of the seven, and may be undergoing positive selection in the *matK* X-domain (a putative RNA binding domain) in response to this (McNeal et al. 2009). Of the three group IIA introns retained in *Sciaphila* (one each in *clpP*, *rpl2*, 3'-*rps12*; Fig. 3.2), only *rpl2* and 3'-*rps12* are targets of *matK* (Zoschke et al. 2010). However, I found no signs of positive selection in *matK* in *Sciaphila* (Table B.4).

3.5.4 Resolution of the phylogenetic position of Triuridaceae in Pandanales

Until recently, most phylogenetic studies of mycoheterotrophs focused on mitochondrial and nuclear genes for phylogenetic inference (e.g., Neyland and Hennigan 2003; Davis et al. 2004; Merckx et al. 2006, 2009; Mennes et al. 2013, 2015), as rate elevation and the loss of mycoheterotroph plastid genes were thought to make them problematic for phylogenetic inference (e.g., Cronquist 1988, p. 467; Merckx et al. 2013a). Molecular data have been scarce for Triuridaceae (Mennes et al. 2013), and previous attempts to amplify plastid genes from the family were unsuccessful (Chase et al. 2000; Caddick et al. 2002). The only purported plastid marker available for Triuridaceae on GenBank is an unpublished *rbcL* sequence of *Sciaphila* sp. (FN870930.1) which is a probable contaminant (it has 97% BLAST match to Commelinaceae, and the gene is not retained in the species of *Sciaphila* we sequenced).

The phylogenetic affinities of members of Triuridaceae have proved to be elusive since the first species was described by Miers (1842). Previous studies suggested relationships with other mycoheterotrophic taxa, such as Petrosaviaceae (Cronquist 1988; Takhtajan 1997). In an early phylogenetic study based on morphology, Dahlgren and Rasmussen (1983) placed the family within Alismatales (as the sister group of the core alismatid families). Dahlgren et al. (1985) later considered its phylogenetic relationship to other families and even its placement within the monocots to be unclear, see also the overview of Triuridaceae systematics in Mennes et al. (2013). Chase et al. (2000) generated the first molecular data for this family (a nuclear 18S rDNA sequence of *Sciaphila*), placing it in Pandanales, a small order of monocots that includes the four photosynthetic families Cycolanthaceae, Pandanaceae, Stemonaceae and Velloziaceae

(APG 2009). Additional studies using one or a few mitochondrial and nuclear sequences (Caddick et al. 2002; Davis et al. 2004; Mennes et al. 2013) and morphological characters (Caddick et al. 2002; Rudall and Bateman 2006) added support for the inclusion of the family in Pandanales. Triuridaceae were therefore assigned to Pandanales in the most recent version of the Angiosperm Phylogeny Group classification system (APG 2009). However, the family's precise position within Pandanales has remained uncertain or poorly supported. Different studies have placed it with weak support as the sister group of Cyclanthaceae and Pandanaceae based on 18S rDNA (Chase et al. 2000), as the sister group of Velloziaceae or of a clade comprising Cyclanthaceae, Pandanaceae and Stemonaceae based on mitochondrial *atpA* (Davis et al. 2004), or even embedded within Stemonaceae based on morphological data (Rudall and Bateman 2006). More recently, Mennes et al. (2013) found it to be the sister group of Cyclanthaceae, Pandanaceae and Stemonaceae using nuclear and mitochondrial data (18S rDNA and three mitochondrial genes). The support for the latter relationship was weak, however, although their analyses strongly rejected a close relationship between Triuridaceae and Velloziaceae, or an origin of Triuridaceae within Stemonaceae. Mennes et al. (2013) also found strong support for the monophyly of Triuridaceae, considering five sampled genera.

Broad inferences of angiosperm phylogeny have relied extensively on a few plastid genes until relatively recently (e.g., APG 2009). It would therefore be valuable to integrate mycoheterotrophs into this large body of evidence. Here I demonstrated the utility of whole plastid genome data for the higher-order phylogenetic placement of Triuridaceae. I confirmed its inclusion in Pandanales (Fig. 3.4), as proposed in previous studies (Chase et al. 2000; Caddick et al. 2002; Davis et al. 2004; Rudall and Bateman 2006; Mennes et al. 2013). I also inferred a more confident placement of Triuridaceae among the four green families in Pandanales, as I found strong evidence that it is the sister group of a clade comprising Cyclanthaceae and Pandanaceae (Fig. 3.4). Stemonaceae are the sister group of this clade, and Velloziaceae are the sister group of all remaining families in the order. Whole plastome data may also be useful for inferring relationships among the nine genera in Triuridaceae, several of which have not been sampled to date (e.g., Mennes et al. 2013). However, I have not been able to retrieve plastid genes to date from two of the genera, *Kupea* and *Seychellaria*.

Elevated rates of substitution are generally observed across all three genomes of mycoheterotrophs compared to their green relatives (e.g., Merckx et al. 2006, 2009), but may be

particularly acute in the plastid genome. The resulting long branches may cause mis-inference in phylogenetic analyses, especially when taxon sampling is sparse (Felsenstein 1978). This may result in incorrect placements of heterotrophic lineages, and de-stabilizing effects on phylogenetic inference for neighboring clades. I demonstrated the latter effect here for parsimony by comparing the results with *Sciaphila* included vs. excluded from consideration in the analysis (Fig. B.7). Fortunately, long-branch effects are expected to be less acute for model-based methods such as maximum likelihood (e.g., Felsenstein 1988; Yang 1996; Huelsenbeck 1997, 1998; Swofford et al. 2001; Yang and Rannala 2012), and may also be minimized by using a dense sampling of species and by using the most realistic model of sequence evolution (Philippe et al. 2011). Here I used model-testing, different data partitioning schemes and examined substitution models that operate at the level of nucleotides, amino-acid residues or codons to explore this issue. Nucleotide and amino-acid models are commonly used and well-characterized methods for phylogenetic inference (for comprehensive reviews, see Swofford et al. 1996; Lío and Goldman 1998). Codon-based substitution models are less frequently implemented, as they consider sequence change at the level of codons, rather than nucleotides or amino-acid residues, and are considerably more computationally intensive. However, they may be more reliable and biologically meaningful than other methods (Goldman and Yang 1994), and may accommodate both closely related and highly divergent sequences in phylogenetic inference (Miyazawa 2013). Here the codon-based method recovered parallel results to the nucleotide and amino-acid based substitution models. This supports the idea that my phylogenetic inferences and the identification of the closest living green relatives of Triuridaceae are robust to these diverse analytical assumptions.

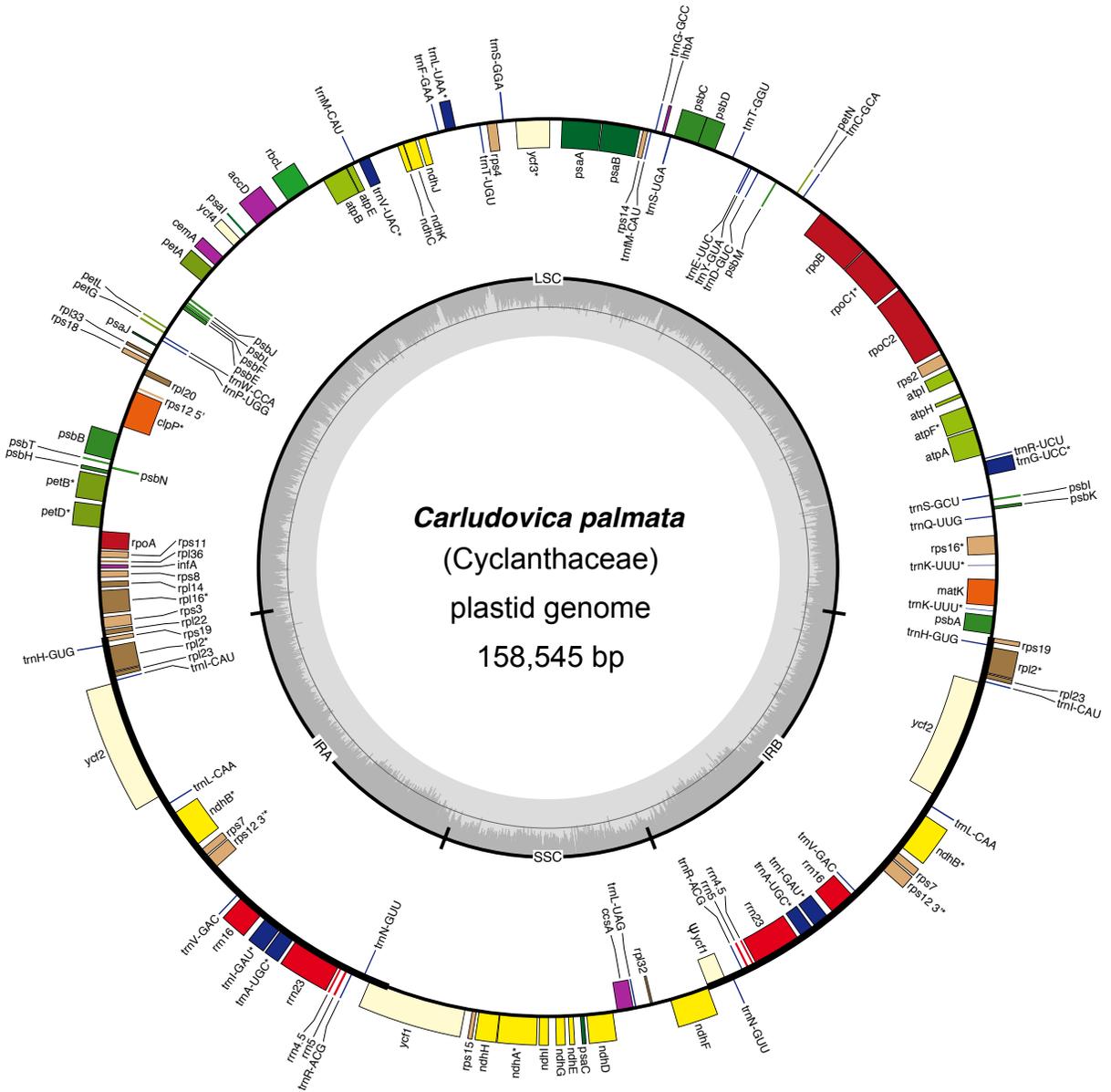
Table 3.1 Summary of genes retained in *Sciaphila* relative to *Carludovica*. Dash ('-') indicates the absence of all genes for that protein complex.

Function	<i>Carludovica palmata</i>	<i>Sciaphila densiflora</i>
Photosynthesis	<i>psaA, psaB, psaC, psaI, psaJ</i>	-
	<i>psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ</i>	-
	<i>atpA, atpB, atpE, atpF, atpH, atpI</i>	-
	<i>petA, petB, petD, petG, petL, petN</i>	-
	<i>rbcL</i>	-
	<i>ycf3, ycf4</i>	-
	<i>ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK</i>	-
Ribosomal proteins	<i>rpl2, rpl14, rpl16, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36</i>	<i>rpl2, rpl14, rpl16, rpl20, rpl36</i>
	<i>rps2, rps3, rps4, rps7, rps8, rps11, rps12, rps14, rps15, rps16, rps18, rps19</i>	<i>rps2, rps3, rps4, rps7, rps8, rps11, rps12, rps14, rps18, rps19</i>
RNA polymerase	<i>rpoA, rpoB, rpoC1, rpoC2</i>	-
Ribosomal DNAs	<i>rrn4.5, rrn5, rrn16, rrn23</i>	<i>rrn4.5, rrn5, rrn16, rrn23</i>
Transfer RNAs	<i>trnA-UGC, trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnJ⁺M-CAU, trnG-GCC, trnG-UCC, 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, trnV-UAC, trnW-CCA, trnY-GUA</i>	<i>trnC-GCA, trnE-UUC, trnJ⁺M-CAU, trnI-CAU, trnQ-UUG, trnW-CCA</i>
Other protein coding genes	<i>accD, ccsA, cemA, clpP, infA, matK, ycf1, ycf2</i>	<i>accD, clpP, matK</i>

Table 3.2 Summary of genes retained in published circular plastid genomes of mycoheterotrophic species. For protein-coding genes, only those found as open reading frames are included. Genes that are found in all species listed below are shown in **bold**.

	<i>Sciaphila densiflora</i> (Triuridaceae)	<i>Aneura mirabilis</i> (Aneuraceae)	<i>Corallorhiza striata</i> (Orchidaceae)	<i>Neottia nidus-avis</i> (Orchidaceae)	<i>Rhizanthella gardneri</i> (Orchidaceae)	<i>Epipogium aphyllum</i> (Orchidaceae)	<i>Epipogium roseum</i> (Orchidaceae)	<i>Petrosavia stellaris</i> (Petrosaviaceae)
Photosynthesis								
Photosystem I	-	<i>psaC</i> , I, J, M <i>ycf4</i>	<i>psaI</i> , J	-	-	-	-	-
Photosystem II	-	<i>psbA</i> , F, H, I J, L, M, N, T, Z	<i>psbH</i> , I, K, M, N, T, Z	-	-	-	-	<i>psbI</i> , Z
ATP synthase	-	<i>atpA</i> , B, E, F H, I	<i>atpA</i> , B, E, F H, I	-	-	-	-	<i>atpA</i> , B, E, F H, I
Cytochrome b ₆	-	<i>petD</i> , G, L, N	<i>petG</i> , L	-	-	-	-	<i>petG</i>
Rubisco	-	<i>rbcL</i>	-	-	-	-	-	<i>rbcL</i>
NAD(P)H dehydrogenase	-	-	<i>ndhJ</i>	-	-	-	-	-
Gene expression								
Ribosomal protein	<i>rpl2</i> , 14 , 16 20, 36	<i>rpl2</i> , 14 , 16 20, 21, 22, 23, 32, 33, 36	<i>rpl2</i> , 14 , 16 20, 22, 23 32, 33, 36	<i>rpl2</i> , 14 , 16 20, 22, 23 32, 33, 36	<i>rpl2</i> , 14 , 16 , 20 23, 36	<i>rpl2</i> , 14 , 16 36	<i>rpl2</i> , 14 , 16 20, 36	<i>rpl2</i> , 14 , 16 20, 22, 23 32, 33, 36
	<i>rps2</i> , 3 , 4 , 7 8 , 11 , 12 , 14 18 , 19	<i>rps2</i> , 3 , 4 , 7 8 , 11 , 12 , 14 15 , 18 , 19	<i>rps2</i> , 3 , 4 , 7 8 , 11 , 12 , 14 15 , 16 , 18 , 19	<i>rps2</i> , 3 , 4 , 7 8 , 11 , 12 , 14 15 , 19	<i>rps2</i> , 3 , 4 , 7 8 , 11 , 14 , 18 19	<i>rps2</i> , 3 , 4 , 7 8 , 11 , 12 , 14 18 , 19	<i>rps2</i> , 3 , 4 , 7 8 , 11 , 12 , 14 18 , 19	<i>rps2</i> , 3 , 4 , 7 8 , 11 , 12 , 14 15 , 16 , 18 , 19
RNA polymerase	-	<i>rpoA</i> , B, C1, C2	-	-	-	-	-	-
Ribosomal DNAs	<i>rrn4.5</i> , 5 , 16 , 23	<i>rrn4.5</i> , 5 , 16 , 23	<i>rrn4.5</i> , 5 , 16 , 23	<i>rrn4.5</i> , 5 , 16 , 23	<i>rrn4.5</i> , 5 , 16 , 23	<i>rrn4.5</i> , 5 , 16 , 23	<i>rrn4.5</i> , 5 , 16 , 23	<i>rrn4.5</i> , 5 , 16 , 23
Transfer RNAs	C GCA, E UUC, I CAU f M _{CAU} , QUUG WCCA	AUGC, C GCA, DGUC E UUC, FGAA, GGCC GUCC, HGUG, I CAU IGAU, KUUU, LCAA, L _{UAA} , LUAG, f M _{CAU} M _{CAU} , N _{GUU} , PUGG QUUG, RUCU, R _{ACG} , RCCG, S _{GCU} , S _{GGA} SUGA, TUGU, TGGU VGAC, VUAC, W _{CCA} Y _{GUA}	AUGC, C GCA, DGUC E UUC, FGAA, GGCC GUCC, HGUG, I CAU IGAU, KUUU, LCAA L _{UAA} , LUAG f M _{CAU} , M _{CAU}	AUGC, C GCA, DGUC E UUC, FGAA, GGCC GUCC, HGUG, I CAU KUUU, LCAA, L _{UAA} L _{UAA} , LUAG, f M _{CAU} M _{CAU} N _{GUU} , QUUG, R _{ACG} RUCU, S _{GCU} , SUGA TGGU, TUGU, W _{CCA} V _{GAC} , Y _{GUA}	C GCA, DGUC, E UUC FGAA, I CAU, f M _{CAU} QUUG, WCCA, Y _{GUA}	C GCA, E UUC, FGAA I CAU, f M _{CAU} , Y _{GUA}	C GCA, E UUC, FGAA I CAU, f M _{CAU} , QUUG Y _{GUA}	AUGC, C GCA, DGUC E UUC, FGAA, GGCC GUCC, HGUG, I CAU IGAU, KUUU, LCAA L _{UAA} , LUAG, f M _{CAU} M _{CAU} , N _{GUU} , PUGG QUUG, R _{ACG} , RUCU S _{GCU} , S _{GGA} , SUGA V _{GAC} , VUAC, TUGU WCCA, Y _{GUA}
Other protein coding genes	<i>accD</i> , <i>clpP</i> , <i>matK</i>	<i>accD</i> , <i>cemA</i> , <i>clpP</i> <i>infA</i> , <i>matK</i> , <i>chlB</i> <i>chlN</i> , <i>chlL</i> , <i>ycf1</i> &2	<i>accD</i> , <i>clpP</i> , <i>infA</i> <i>matK</i> , <i>ycf1</i> &2	<i>accD</i> , <i>clpP</i> , <i>infA</i> <i>ycf1</i> &2	<i>accD</i> , <i>clpP</i> , <i>infA</i> <i>ycf1</i> &2	<i>accD</i> , <i>clpP</i> , <i>infA</i>	<i>accD</i> , <i>clpP</i> , <i>infA</i>	<i>accD</i> , <i>clpP</i> , <i>infA</i> <i>matK</i> , <i>ycf1</i> &2

Figure 3.1 Circular plastome map of *Carludovica palmata* (Cyclanthaceae). 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 branches indicate inverted repeat (IR) copies. Genes with introns are indicated with asterisks (*). The short pseudogene copy of *ycf1* is marked as ‘ ψ ’.

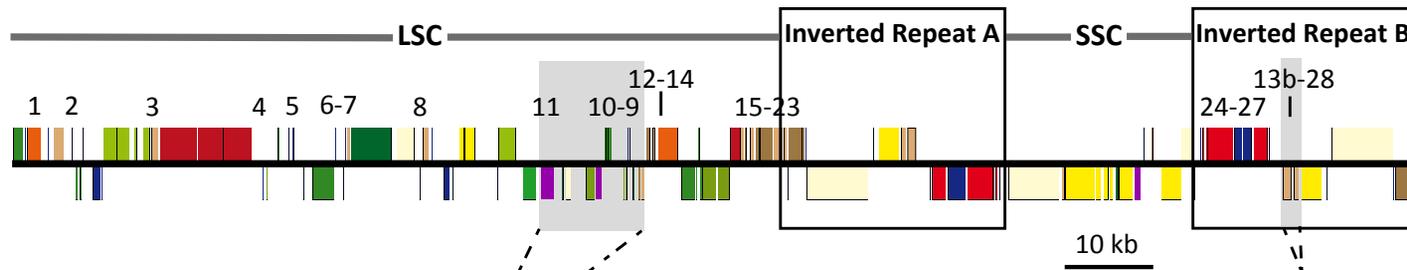


- | | |
|---|--|
| ■ photosystem I | ■ ribosomal proteins (SSU) |
| ■ photosystem II | ■ ribosomal proteins (LSU) |
| ■ cytochrome b/f complex | ■ clpP, matK |
| ■ ATP synthase | ■ other genes |
| ■ NADH dehydrogenase | ■ hypothetical chloroplast reading frames (ycf) |
| ■ RubisCO large subunit | ■ transfer RNAs |
| ■ RNA polymerase | ■ ribosomal DNAs |

Figure 3.2 Circular plastome map of *Sciaphila densiflora* (Triuridaceae). Genes located inside the circle are transcribed clockwise, those outside are transcribed counterclockwise. The exterior arc is a sector with possible repeats (thicker line indicates higher coverage, see main text and Fig. 3.3; the dotted line indicates Sanger sequence data). The grey circle marks the GC content: the inner circle marks a 50% threshold. Genes with introns are indicated with asterisks (*).

Figure 3.3 Comparison of linearized plastomes of *Carludovica palmata* (Cyclanthaceae) and *Sciaphila densiflora* (Triuridaceae). Boxes indicate inverted repeat regions (two copies, A and B) in *Carludovica*. Dashed arrows indicate predicted inversions of the small blocks highlighted in gray. Black lines below the *Sciaphila* plastome map indicate individual contigs (numbers below the lines indicate the estimated relative depth of coverage, see text). Gaps and contig overlap were respectively connected or confirmed using Sanger sequencing with primers at positions indicated with short arrows (primers not to scale; thin dashed lines are sequenced regions not represented in *de novo* contigs). A sector with higher read depth is indicated (the extent of higher-coverage is uncertain because this sector overlaps with a region produced using Sanger sequencing, indicated with a dotted line; 4X = four times coverage; 2X = two times coverage, see main text). LSC: large single copy region; SSC: small single copy region; numbers indicate the 28 genes retained in *Sciaphila*, 18 of which are protein-coding (note that *rps12* is a trans-spliced gene, noted here as 13a and 13b); asterisks (*) indicate genes with introns. The scale bars indicate relative plastome sizes of *Carludovica* and *Sciaphila* (kb = kilobase).

Carludovica palmata



Sciaphila densiflora

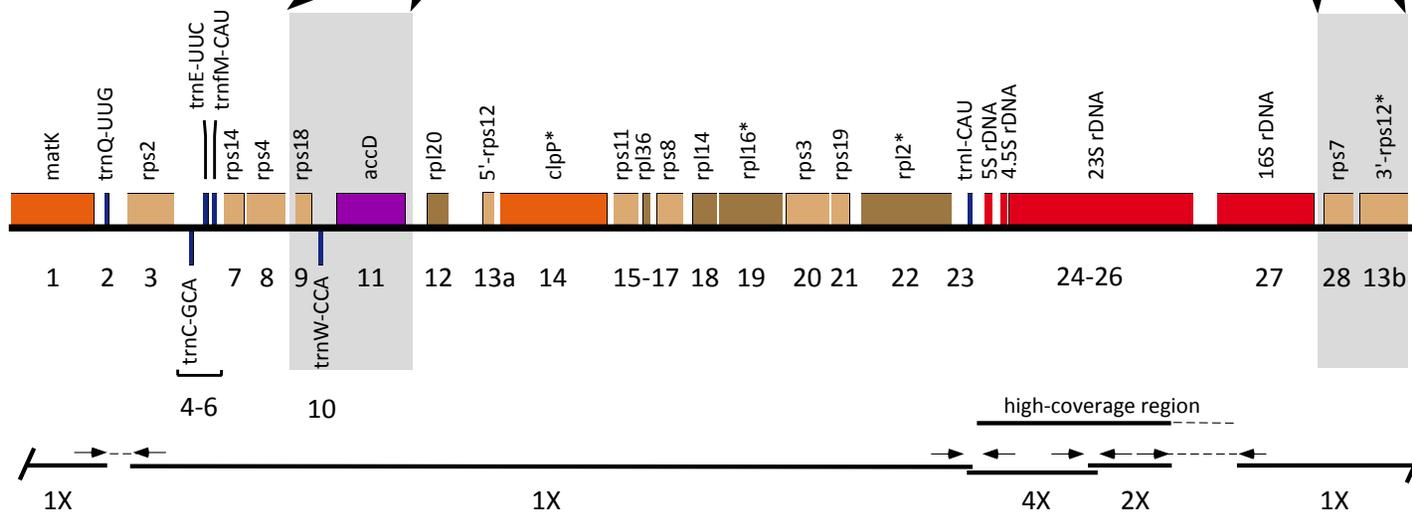
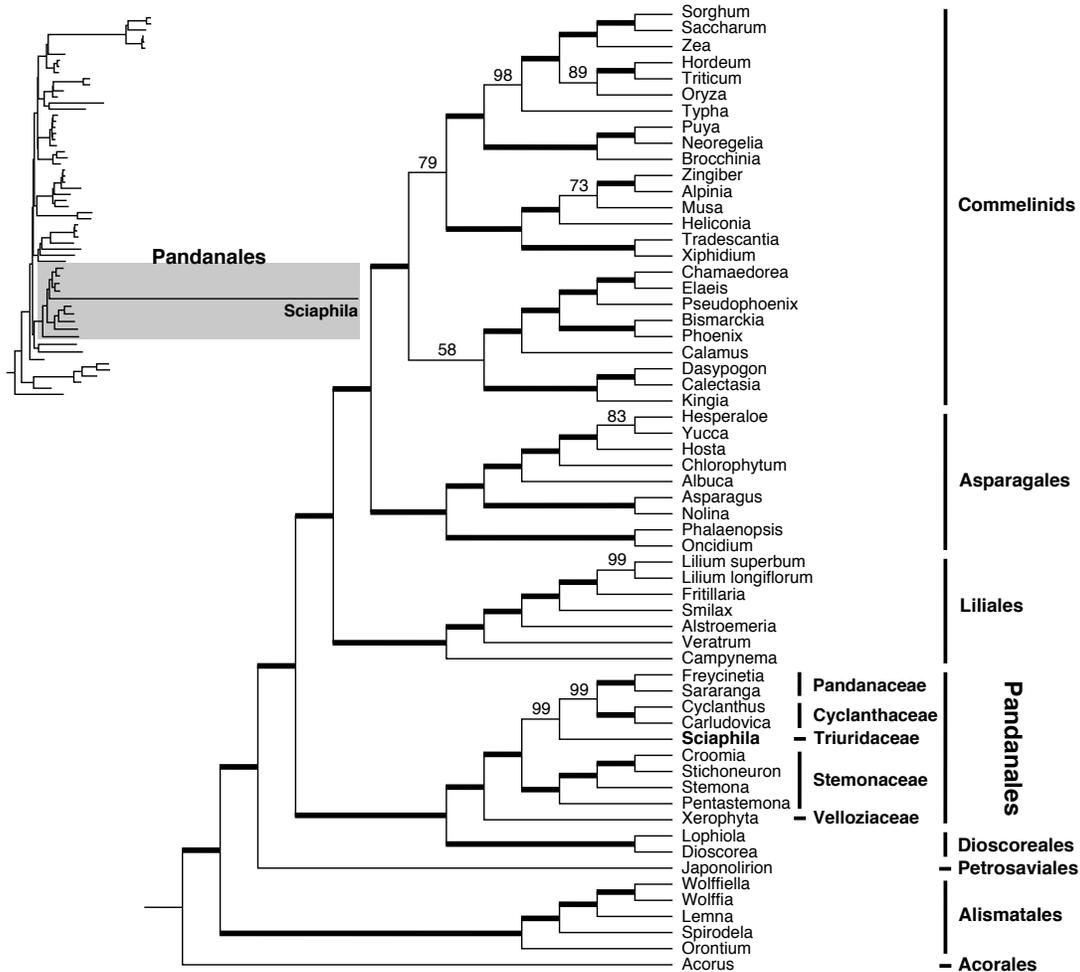


Figure 3.4 Phylogenetic relationships in Pandanales in the context of overall monocot phylogeny, based on plastid genome data (82 plastid genes in photosynthetic taxa; 22 in *Sciaphila*). The data matrix was partitioned using a G x C (gene by codon) partitioning scheme and analyzed using corresponding nucleotide substitution models (see text and table S2b for details). Thick lines indicate 100% bootstrap support; branches with lower support are indicated. The scale bar indicates estimated substitutions per site. This is a subset of a larger angiosperm-wide sampling (Fig. B.3, shown as an inset phylogram here; the shaded portion represents Pandanales).



Chapter 4: Comparative phylogenomics of mycoheterotrophic monocots¹

4.1 Summary

Most mycoheterotrophic plastid genomes (plastomes) sequenced to date belong to the monocot family Orchidaceae. Here I sequenced and characterized additional plastomes from ten additional fully mycoheterotrophic monocots, including multiple representatives of Burmanniaceae, Corsiaceae and Petrosaviaceae, a single representative each of Iridaceae and Triuridaceae (the latter discussed in more detail in Chapter 3), and six autotrophic relatives, including two in Burmanniaceae. I used these to infer the phylogenetic placements of these mycoheterotrophic lineages, and to characterize major plastome structural changes and patterns of gene loss and retention. Likelihood analysis strongly supports placement of these taxa in positions consistent with other molecular data sets (e.g., Chapter 2). Inferred relationships within Burmanniaceae are moderately well supported and consistent with independent losses of photosynthesis in the family. Plastid genes are generally maintained in colinear arrangements in the plastid genomes of mycoheterotroph and their autotroph relatives, pointing to contraction (associated with gene deletion) as the most significant cause of genome structural change in monocot mycoheterotrophs. This genome reduction is sometimes extreme: *Apteria aphylla* (Burmanniaceae) has one of the most reduced plastid genomes sequenced to date (15,715 bp). Additional changes include multiple large-scale inversions in *Arachnitis uniflora* (Corsiaceae) including inversion of its inverted repeat (IR) region, loss of the IR region (with at least four distinct losses, one between two closely related taxa in Petrosaviaceae), IR duplication (a direct repeat of one IR region in *Campylosiphon*, Burmanniaceae), and other generally smaller-scale shifts in the IR boundaries. Patterns of gene loss and retention are largely consistent with a recent hypothesis of ordered gene loss in heterotrophic taxa. They include occasional retention (representing delayed loss?) of some genes (the photosynthesis-related ATP synthase and Rubisco complexes, and plastid-encoded RNA polymerase), possibly supporting alternative functions of the corresponding plastid protein complexes. The common retention of additional

¹ A portion of the data obtained for this chapter (the plastid gene data set for Corsiaceae) was previously published in: Mennes, C.B., Lam, V.K.Y., Rudall, P.J., Lyon, S. P., Graham, S.W., Smets, E. F. and Merckx, V.S.F. 2015. Ancient Gondwana break-up explains the distribution of the mycoheterotrophic family Corsiaceae (Liliales). *Journal of Biogeography* 42:1123-1136.

genes with primary or secondary roles outside photosynthesis or protein translation (i.e., plastid-encoded *accD*, *trnE* and *trnfM*) indicates that these loci may be essential for plant survival regardless of trophic status, and explains retention of at least some plastid translation apparatus genes across all surveyed mycoheterotrophs here.

4.2 Introduction

Fully mycoheterotrophic plants no longer perform photosynthesis, and instead obtain their fixed carbon, water and other nutrients required for plant growth by tapping into mycorrhizal networks comprising soil fungi and their autotrophic (green-plant) partners, or occasionally from saprophytic fungi (Leake 1994, 2005). Mycoheterotrophy is distinct from parasitism, in which plants steal nutrients through direct physical connections (e.g., haustorial-like roots) to green plants (Heide-Jørgensen 2008); mycoheterotrophs instead attract, host and consume fungi in specialized underground organs (Imhof 2010). There are ~514 fully mycoheterotrophic species in the land plants (Merckx et al. 2013a). Most of these are monocots, which appear to be especially prone to this evolutionary transition (Imhof 2010): approximately 80% of all origins of full mycoheterotrophy (and corresponding losses of photosynthesis) belong to this major angiosperm clade (Merckx et al. 2013b). In total, ~91% of all fully mycoheterotrophic species are found in just seven monocot families (Burmanniaceae, Corsiaceae, Iridaceae, Orchidaceae, Petrosaviaceae, Triuridaceae and Thismiaceae), three of which are fully mycoheterotrophic (Corsiaceae, Triuridaceae and Thismiaceae; Leake 1994; Merckx et al. 2013b). Burmanniaceae and Orchidaceae have each experienced multiple independent origins of full mycoheterotrophy, associated with at least eight and 25 associated losses of photosynthesis, respectively (Merckx et al. 2013b; note that both families have autotrophic and mycoheterotrophic members). These separate origins provide evolutionarily independent reference points for studying the effects of mycoheterotrophy and photosynthesis loss on plant biology. The remaining full mycoheterotrophs belong to the conifers (*Parasitaxus usta* in Podocarpaceae, Chapter 5), eudicotyledons (multiple species in Ericaceae, Gentianaceae and Polygalaceae), liverworts (*Aneura mirabilis*), and possibly also the fern *Schizaea fluminensis* (Schizaeaceae) and moss *Buxbaumia aphylla* (Buxbaumiaceae) (see Leake 1994; Merckx et al. 2013b). Multiple lineages of ferns and lycophytes have fully mycoheterotrophic gametophytes, a form of initial

mycoheterotrophy analogous to that found in orchids, as they are mycoheterotrophic during establishment and most become photosynthetic at maturity (e.g., Gifford and Foster 1989; Wang and Qiu 2006; Winther and Friedman 2008). In addition, multiple lineages of Burmanniaceae, Ericaceae, Gentianaceae, Orchidaceae and possibly other lineages (e.g., sporophytes of *Botrychium*, Ophioglossaceae) are partially mycoheterotrophic, in that mature plants are thought to obtain only some of their fixed carbon using photosynthesis and the rest from fungal partners (e.g., Gifford and Foster 1989; Winther and Friedman 2007; Tedersoo et al. 2007; Zimmer et al. 2007, 2008; Cameron and Bolin 2010).

Convergent and divergent evolution resulting in vegetative reduction in mycoheterotrophs and sometimes spectacular floral modifications has made it difficult to place these lineages into higher-order classification based on morphological evidence alone (see the contrasting classification schemes of Cronquist, 1981, Dahlgren et al., 1985, and Takhtajan, 2009, for example). Full mycoheterotrophs are also not well integrated into recent angiosperm classification schemes, which are still mostly based on data from a few plastid and nuclear genes. For example, the Angiosperm Phylogeny Group (APG 2003, 2009) classification systems rely extensively on the photosynthetic plastid genes *rbcL* and *atpB*, in addition to the nuclear 18S rDNA locus (e.g., Chase et al. 1993, Soltis et al. 2000). Photosynthesis-related genes in heterotrophs are often pseudogenized or lost due to relaxed or released selection (Barrett et al. 2014), and the retained genes may exhibit highly elevated substitution rates (e.g., Chapter 2). Extensive rate elevation may interfere with phylogenetic placement due to long-branch attraction (e.g., Felsenstein 1987). However, I demonstrated in Chapter 2 that phylogenetic analyses based on only a few plastid genes (among those typically retained in heterotrophic plants) can permit integration of fully mycoheterotrophic lineages into monocot phylogeny, even when the genes are rapidly evolving or patchily retained. I also showed in Chapter 3 that phylogenetic inference based on gene sets retrieved from full plastome genomes using next-generation sequencing (NGS) can be used to place a highly modified mycoheterotroph (*Sciaphila*, Triuridaceae) with strong support in monocot phylogeny. However, only a limited subset of full mycoheterotrophs have had their plastid genomes included in phylogenomic analyses: these are *Petrosavia stellaris* (Logacheva et al. 2014), multiple lineages of orchids in the study of Schelkunov et al. (2015) that included data from Delannoy et al. (2011), Logacheva et al. (2011) and Barrett et al. (2014), and most recently the eudicot *Monotropa hypopitys*, Ericaceae (Gruzdev et al. 2016). Mennes et al.

(2015) also provided examples in Corsiaceae that used plastid gene sets extracted from full plastomes presented in more detail here. Thus, it would be useful to refine our understanding of the phylogenetic placement of these taxa by retrieving additional plastid genomes of mycoheterotrophs.

Comparative plastome analyses of mycoheterotrophic taxa and their autotrophic relatives should be useful for improving our understanding of plastid genome structure, evolution and function (e.g., Chapters 3, 5; Barrett and Davis 2012; Logacheva et al. 2014; Schelkunov et al. 2015). The typical or “canonical” land-plant plastome has a highly conserved gene content and order, and comprises two inverted repeat (IR) copies that separate large and small single copy regions (LSC and SSC, respectively; reviewed in Palmer et al. 1985). The inverted repeat may play a role in genome structural stability (e.g., Palmer and Thompson 1982; Márechal and Brisson 2010) and is subject to generally minor shifts in extent due to expansion and contraction, as single copy genes are incorporated into or expelled from it (e.g., Wang et al. 2008; Zhu et al. 2015). The quadripartite plastome structure (and underlying gene order) is conserved in multiple fully mycoheterotrophic lineages (e.g., Wickett et al. 2008; Logacheva et al. 2011; Barrett et al. 2012) but may be lost (or altered) in others (e.g., Logacheva et al. 2014; Triuridaceae; Chapter 3). However, the degree to which a functional plastome may deviate from the canonical angiosperm plastid structure is poorly known, and deserves additional attention.

In this chapter I address several major phylogenetic and evolutionary questions using newly retrieved full plastid genomes from multiple independent mycoheterotrophic monocot lineages. I use these genome-scale data sets to: (1) perform phylogenomic investigations of the placement of these taxa in the current framework of monocot phylogeny, refining pictures developed in Chapters 2 and 3; (2) test hypothesized trajectories of gene loss and retention, in particular the related models proposed by Barrett and Davis (2012) and Barrett et al. (2014), discussed in earlier chapters of my thesis (retentions of photosynthesis-related genes may be consistent with non-photosynthetic functions of these genes), and; (3) characterize the degree of plastid genome structural changes that occurred with each origin of full mycoheterotrophy including genome contraction, departures from genome colinearity, and changes in plastid inverted-repeat boundaries.

4.3 Materials and methods

4.3.1 Taxon sampling

I generated new plastid genome (plastome) sequences for 10 mycoheterotrophic monocots (this count includes the *Sciaphila* genome presented in more detail in Chapter 3) and six representative autotrophic relatives for this study (Table C.1). I added these to comparable sequences from taxa representing major monocot lineages, and included eudicots, magnoliids and the orders Amborellales, Nymphaeales and Austrobaileyales (i.e., ANA-grade taxa) as outgroups, with a total of ninety taxa considered for phylogenetic inference (Table C.1).

4.3.2 DNA extraction and sequencing

I extracted total genomic DNA using a modified CTAB protocol (Doyle and Doyle 1987; Rai et al. 2003) and prepared whole-genome shotgun libraries with several library sequencing kits, using NuGEN Ovation Ultralow Library System (NuGEN Technologies Inc., San Carlos, CA) for taxa with low amounts (<1 ng) of DNA (i.e., for *Campylosiphon*, *Campynema*, *Iris*, *Japonolirion* and *Lilium*), and Bioo Nextflex DNA sequencing kit (Bioo Scientific Corp., Austin, TX) and KAPA LTP Library Preparation kit (KAPA Biosystems, Boston, MA) for the remaining taxa. I sheared DNAs to ~400-bp fragments on a Covaris S220 sonicator (Covaris, Inc., Woburn, MA) for library preparation with all three kits, and prepared libraries following the methodology outlined in Chapter 3. Multiplexed libraries (Cronn et al. 2008) were sequenced on an Illumina HiSeq 2000 (Illumina, Inc., San Diego, CA) as 100-bp paired-end reads.

4.3.3 *De novo* assembly, gene annotation, and plastome reconstruction

Illumina reads were processed with CASAVA 1.8.2. (Illumina Inc.) to discard low-quality reads and sort multiplexed data by taxon. I assembled processed Illumina reads into *de novo* contigs with CLC Genomics Workbench v.6.5.1 (CLC Bio, Aarhus, Denmark) using default settings, and selected for contigs > 500 bp with at least 20X coverage. I used a custom Perl script (Daisie Huang, University of British Columbia; https://github.com/daisieh/phylogenomics/tree/master/filtering/filter_cp.pl) to BLAST contigs against a local database (Altschul et al. 1990) of plastid genes (*Dioscorea elephantipes*;

NC_009601.1; *Phoenix dactylifera* NC_013991) to select for plastid contigs. I used Primer3 (Koressaar and Remm 2007; Untergasser et al. 2007) to design taxon-specific primers to join contigs and confirm contig overlaps, amplifying DNA using Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific), and performing Sanger sequencing using BigDye Terminator v.3.1 sequencing chemistry (Applied Biosystems, Inc., Foster City, CA), with sequencing reactions run on an Applied Biosystems 3730S 48-capillary DNA analyzer (Applied Biosystems, Inc.). I assembled full circular plastome sequences from *de novo* contigs and Sanger-derived sequences for each taxon in Sequencher 4.2.2. (Gene Codes Corporation, Ann Arbor, US), and annotated plastid genes for the completed sequences in DOGMA (Wyman et al. 2004). I checked gene and exon boundaries for each protein-coding gene using the *D. elephantipes* and *P. dactylifera* plastomes as references, checking intron and intergenic spacer regions in mycoheterotrophic taxa in Sequencher to look for pseudogenes that may have been missed by BLAST. I also searched for potentially missing tRNAs using tRNAscan-SE search (Lowe and Eddy 1997). I illustrated plastomes using OGDRAW (Lohse et al. 2013) to generate plastome maps.

4.3.4 Matrix assembly and phylogenetic analyses

I extracted coding regions for up to 82 genes per taxon, representing 78 of 79 protein-coding genes in the plastid genome (I retrieved *ycf1*, but did not include it in downstream analyses due to difficulties in sequence alignment), and four plastid rDNA genes. I aligned each gene separately with Se-AL v.2.0a11 (Rambaut 2002) using alignment criteria laid out in Graham et al. (2000), and staggered regions that were difficult to align (Saarela and Graham 2010), checking that the protein-coding genes had open reading frames (pseudogenes were excluded from the alignment). I then concatenated individual gene matrices into a 113,652 bp matrix (based on 67,430 bp unaligned sequence, using photosynthetic *Burmannia bicolor* as a reference). I exported sequences from the completed matrix and crosschecked them in Sequencher with the original data for editing errors (none were found).

I performed heuristic tree searches using and maximum likelihood (ML) and parsimony methods. For parsimony, I used tree-bisection-reconnection branch swapping in PAUP* v4.0a134 (Swofford 2003) with 1,000 random stepwise addition replicates, holding 100 trees at each step (and otherwise using default settings) to search for the shortest tree(s). For likelihood, I

used RAxML v.7.4.2 (Stamatakis 2006), with a graphical user interface (Silvestro and Michalak 2012), conducting 20 independent searches for the best tree, and analyzing the data in an unpartitioned analysis (meaning that the same model and model parameters were used for all sites), or with the data partitioned by gene and codon position ('GxC' partitioning scheme). For the latter analysis I initially set up 241 data partitions (three codon-based partitions per gene, operationally counting trans-spliced *rps12* gene as two separate genes, 3'-*rps12* and 5'-*rps12*, and with each rDNA considered separately), and combined partitions with similar DNA substitution models and model parameters using PartitionFinder v.1.1.1 (Lanfear et al. 2012) and the Bayesian information criterion (BIC, Schwarz 1978). The PartitionFinder analysis yielded 18 distinct subpartitions with the GTR (general time reversible) + G (G = gamma) or GTR + G + I (I = proportion of invariant sites) models selected as optimal DNA substitution models (Table C.2). I used the GTR + G model for each individual subpartition in subsequent phylogenetic analysis, as the 'I' parameter (accounting for invariant sites) may be accommodated by gamma (Yang 2006). GTR + G was selected as the best model for the unpartitioned analysis.

I estimated branch support with bootstrap analysis (Felsenstein 1985), using 1,000 replicates with 100 random addition replicates per bootstrap replicate for the parsimony analysis. For the likelihood bootstrap analysis I used 500 rapid bootstrap and employed the models and schemes described above for the partitioned and unpartitioned analyses. For both parsimony and ML analyses, I consider strongly supported branches to have at least 95% bootstrap support, moderately supported branches to have 70-94% bootstrap support, and weakly supported branches less than 70% support (Zgurski et al. 2008).

4.3.5 Plastome comparison of mycoheterotrophic taxa and autotrophic relatives

I visualized plastome colinearity between pairs of mycoheterotrophic monocots and close autotrophic relatives (see phylogenetic analyses below) in two ways. Note that the SSC exists *in vivo* as two isomeric orientations (e.g., Walker et al. 2016), although I generally depicted the small single copy region in a consistent orientation (*Petrosavia sakuraii* and *Campylosiphon* are two exceptions). First, I used the webtool Kablammo (Wintersinger and Wasmuth 2015), to summarize sequence alignment output produced by query-subject pairwise BLAST (as xml outputs). For BLAST, I first adjusted the starting points of each plastome sequence in Sequencher to minimize the number of arrangements shown. I used the 'discontiguous

megablast' option in the pairwise BLAST with the default word size. I adjusted the 'show only alignments covering' display parameter from 0%-0.5% (depending on the species) to remove spurious results (poorly aligned regions shown as 'white spaces' were manually checked).

As the Kablammo output does not depict simple inversions (an individual block that is otherwise in the same sequential order) or very diverse rearrangements, I used Mauve to provide more detailed predictions of plastome rearrangement. I used Mauve 2.3.1 (Darling et al. 2004, 2010) to refine predictions of gene-order rearrangements. This version of Mauve (progressiveMauve) uses an algorithm based on a sum-of-pairs approach to align regions of shared homology between two or more sequences in an alignment (i.e., locally colinear blocks; LCBs). ProgressiveMauve is better optimized than previous versions for genomes that may have undergone gene loss and rearrangements (Darling et al. 2010). It positions the LCBs using progressive alignment based on the approach of CLUSTALW (Thompson et al. 1994; Darling et al. 2004). I used a seed weight of 19, and otherwise used default settings to examine several pairs of mycoheterotrophic taxa and their autotrophic relatives. I also conducted linear regression analysis in StatPlus:mac in Excel to examine a possible association between the number of unique genes in retained as open reading frames, and plastome size. For this analysis, I included a set of taxa for this analysis that represent five independent losses of photosynthesis, and one autotrophic member of Burmanniaceae, to derive a coefficient of determination (r^2) and p -value based on a corresponding F-test.

I also visually characterized the IR/SSC and IR/LSC boundaries of the inverted repeat region in mycoheterotrophic monocots and their close green relatives to those of the 'canonical' angiosperm IR as inferred by Zhu et al. (2015), noting additional genes that are included or excluded compared to the canonical IR typical of most taxa.

4.4 Results

4.4.1 Plastomes of autotrophic lineages

I assembled new full plastomes for six autotrophic taxa (two photosynthetic species of *Burmannia*, Burmanniaceae; *Campynema*, Campynemataceae, the sister group of Corsiaceae; *Iris*, Iridaceae; *Lilium*, Liliaceae and *Japonolirion*, Petrosaviaceae). Linearized maps of the photosynthetic *Burmannia* species are presented in Fig. 4.1; circular maps of the other four green

taxa are presented in Figs. C1-C4 (genome summary statistics noted in Table 4.1). The plastid genomes for green taxa are largely consistent in terms of size, gene order and content, and have the quadripartite structure (Table 4.1) typical of other autotrophic lineages (e.g., *Dioscorea elephantipes*; NC_009601.1; *Phoenix dactylifera* NC_013991). Minor differences in Liliales include the presence of pseudogenized versions of *infA* in *Campynema* and *Lilium*, and of *rps16* in *Campynema*. The plastomes of *Burmannia bicolor* and *B. capitata*, two photosynthetic taxa in Burmanniaceae, generally resemble those of typical angiosperm plastomes, except that *rps16* is absent in both species. *Burmannia capitata* has also likely experienced loss of function of plastid NAD(P)H dehydrogenase, as six of nine plastid-encoded subunits for this complex have internal stop codons and are probable pseudogenes (Fig. 4.2). The proportion of the plastome that is coding (~55.4-60.5%) and the plastome GC content (32.2-37.2%) both have relatively narrow ranges across these autotrophic taxa (Table 4.1).

4.4.2 Overview of surveyed fully mycoheterotrophic monocot plastomes

Linearized maps of the ten fully mycoheterotrophic monocots included here are presented in Fig. 4.1 (see also summary statistics in Table 4.1). *Campylosiphon* (Burmanniaceae), *Geosiris* (Iridaceae) and the two newly sequenced *Petrosavia* (Petrosaviaceae) taxa show moderately severe degrees of gene loss and plastome size reduction. The remaining taxa are even more reduced in size, with fewer retained genes (Table 4.1, Fig. 4.2). The plastome size of the fully mycoheterotrophic species range from those typical of some autotrophs (e.g., *Geosiris* is 123,620 bp, compared to *Dioscorea*, 152,609 bp), to almost one tenth of this (for *Apteria aphylla*, Burmanniaceae; 15,715 bp; Table 4.1). There is a significant positive relationship between the number of unique genes retained as open reading frames and plastome size (Fig. 4.3; $r^2=0.981$, $p<0.001$). *Geosiris* (Iridaceae) and *Campylosiphon* (Burmanniaceae) have many pseudogenes that contribute to the lower percentages of their plastomes retrieved as open reading frames compared to other mycoheterotrophic taxa (31.8% and 34.3%, respectively, compared to a range of 42.5-78.3% in other full mycoheterotrophs, Table 4.1; note that these values consider only the LSC, SSC and one copy of an IR, if present). *Apteria* is the most compact genome, with the lowest percentage of noncoding sequence (only 21.7% of the genome is noncoding) among the fully mycoheterotrophic taxa, many of which have coding fractions comparable to or higher than the green taxa included here. The GC content of the fully mycoheterotrophic plastomes varies

from 24.9% (*Gymnosiphon*, Burmanniaceae) to 39.9% in *Sciaphila* (Triuridaceae), compared to a narrower range among the six green taxa examined here (i.e., 32.2%-37.2% GC content; Table 4.1).

4.4.3 Gene loss and retention in mycoheterotrophic monocots

There are 17 commonly retained plastid genes across the ten fully mycoheterotrophic taxa that I sequenced, mostly translational apparatus genes: the core set of retained genes comprises 10 ribosomal proteins (*rpl2*, *rpl14*, *rpl16*, *rps3*, *rps4*, *rps7*, *rps8*, *rps12*, *rps14* and *rps19*), acetyl-coA carboxylase subunit (*accD*), four rDNAs (*rrn4.5*, *rrn5*, *rrn16*, *rrn23*) and two tRNAs (*trnE-UUC*, *trnFM-CAU*). Gene losses are somewhat idiosyncratic, and are documented below for each family and summarized in Fig. 4.2 (this figure includes comparisons to other published heterotrophic and autotrophic genomes, all arranged by genome size); *Sciaphila* (Triuridaceae) is discussed in more detail in Chapter 3, but is also included here for comparison.

In fully mycoheterotrophic Burmanniaceae, photosynthetic (*atp*, *ccsA*, *cemA*, *ndh*, *pet*, *psa*, *psb*, *rbcL*, *lhbA*, *ycf3*, *ycf4*) and plastid-encoded RNA polymerase genes (*rpo*) loci are either lost or have interruptions consistent with them being pseudogenes, see Fig. 4.2 (for the remainder of this section, interrupted genes are referred to as pseudogenes, though in practice this should be demonstrated with additional evidence). *Campylosiphon* retains the most housekeeping genes among full mycoheterotrophs in this family (only *rpl23*, *rps16*, *ycf2* and nine tRNAs are lost or pseudogenized) while *Apteria aphylla* has the fewest genes retained among all mycoheterotrophs surveyed here (21 genes, comprising 14 protein coding, four rDNA and three tRNA genes). *Apteria* also lacks detectable pseudogenes (Fig. 4.2).

In the two fully mycoheterotrophic members of Corsiaceae, *Arachnitis* and *Corsia*, all photosynthesis-related genes (*atp*, *ccsA*, *cemA*, *ndh*, *pet*, *psa*, *psb*, *rbcL*, *lhbA*, *ycf3*, *ycf4*) plastid-encoded RNA polymerase (*rpo*), and translation initiator (*infA*) genes are either lost or pseudogenized (Fig. 4.2). Of the housekeeping genes, *rpl23*, *rpl36*, *rps16*, 12 tRNA genes and *matK* are either lost or pseudogenized. The genes *rpl22*, *rpl32*, *rpl33*, *rps15*, *ycf1*, *ycf2* and 13 tRNA genes are also lost or pseudogenized in *Arachnitis*, which retains only 25 genes in open reading frame (16 protein coding genes, four rDNA, and five tRNA genes) and a single pseudogene (*rpl16*).

In the two species of fully mycoheterotrophic taxa in Petrosaviaceae surveyed here, *P. sakurarii* (the species-level identification of one taxon is unclear; the two specimens may be conspecific, and one taxon is referred to as ‘aff. *sakurarii*’) most photosynthesis-related genes (*ccsA*, *cemA*, *ndh*, *pet*, *psa*, *psb*, *rbcL*, *lhbA*, *ycf3*, *ycf4*) are lost or pseudogenized (Fig. 4.2). The *trnT*-GGU locus is lost in both taxa, but *trnL*-CAA is lost only in *Petrosavia* aff. *sakurarii*. Both taxa retain all of the plastid-encoded genes for plastid-encoded subunits of the ATP synthase (*atp*) and plastid-encoded RNA polymerase (*rpo*) as open reading frames.

In *Geosiris aphylla*, the sole fully mycoheterotrophic member of Iridaceae included here, all photosynthesis genes (*atp*, *ccsA*, *cemA*, *ndh*, *pet*, *psa*, *psb*, *rbcL*, *lhbA*, *ycf3*, *ycf4*) and plastid-encoded RNA polymerase (*rpo*) subunit loci are either pseudogenized or are lost (the only exception is *psbM*, which is retained in open reading frame) (Fig. 4.2). All of the remaining genes are retained as open reading frames.

In the single mycoheterotrophic member of Triuridaceae examined here, *Sciaphila densiflora*, all photosynthesis-related genes (*atp*, *ccsA*, *cemA*, *ndh*, *pet*, *psa*, *psb*, *rbcL*, *lhbA*, *ycf3*, *ycf4*), plastid-encoded RNA polymerase loci (*rpo*) and translation initiator (*infA*) genes are lost (Fig. 4.2; Chapter 3). *Sciaphila* retains most plastid-encoded ribosomal proteins but has lost several other housekeeping genes (*matK*, *ycf1*, *ycf2*), and 24 of 30 plastid tRNA genes.

4.4.4 Assessments of colinearity in mycoheterotrophic plastomes

The Kablammo and Mauve comparisons demonstrate that despite massive gene loss, most mycoheterotrophic plastomes exhibit near or absolute colinearity with their autotrophic relatives. This is the case in *Apteria*, *Campylosiphon* and *Gymnosiphon* in Burmanniaceae, *Geosiris* (Iridaceae), *Sciaphila* (Triuridaceae) and the two newly sequenced *Petrosavia* taxa (Figs. 4.4b, d, e, h-j, l; C.5, C.7, C.8, C.11, C.12, C.14); blue bands in the Kablammo figures indicate colinear blocks of genes shared by mycoheterotrophic taxa and autotrophic relatives. For Burmanniaceae, I used *B. bicolor* as the photosynthetic reference taxon; the two autotrophic members of *Burmannia* are also compared (Figs. 4.4a, C.13). *Burmannia itoana* (Burmanniaceae) and *Arachnitis* and *Corsia* (Corsiaceae) (Figs. 4.4c, f, g, C.6 C.9, C.10) have several major inversions relative to their autotrophic relatives. *Campylosiphon* has an additional directly repeated copy of one its inverted repeat regions relative to *Burmannia bicolor* (Fig. 4.4b; see also Figs. 4.1 and Fig. C.5), and *Burmannia itoana* has a novel or highly shifted IR (Table 4.2; Figs. 4.4c, C.6; see

below). *Arachnitis* has experienced an inversion of its inverted repeats (in terms of which ends face into the large and small single copy regions; Fig. 4.1). Areas between blue bands in Fig 4.4 ('white space') in pairwise Kablammo comparisons generally represent sequences with high sequence diversity (e.g., variable coding regions such as portions of *accD* and *ycf1*, high rates of pseudogenization, or intergenic spacers or intron regions); although these were not recognized by the BLAST cut-off I used, spot-checks indicated that they were colinear with the relevant green relative, considering features that could be checked manually.

4.4.5 Inverted repeat (IR) evolution in mycoheterotrophic plastomes

Several fully mycoheterotrophic taxa examined here lack a canonical quadripartite structure. Five species (*Apteria* and *Gymnosiphon*, Burmanniaceae; *Corsia*, Corsiaceae; *Petrosavia* aff. *sakurarii*, Petrosaviaceae, and *Sciaphila*, Triuridaceae) likely all lack an inverted repeat. As noted above, *Campylosiphon* (Burmanniaceae) has an additional direct copy of one of its two IR regions (Fig. 4.1), and in *Arachnitis* (Corsiaceae) the IR repeats are completely inverted relative to those of typical angiosperm plastomes (Fig. 4.1). Inverted repeat boundaries also vary among taxa compared to the canonical IR typical of most angiosperms, in which the IR/SSC boundary is demarcated by *trnN*-GUU in the IR, and IR/LSC boundary by *rpl2* in the IR, the arrangement in *Nicotiana* and most angiosperms (e.g., Goulding et al. 1996; Zhu et al. 2015). Note that only the last full gene included in the IR is noted for IR boundaries listed in Table 4.2 (see Fig. 4.1 for additional duplicated pseudogenes or partial genes in some cases).

All species of Burmanniaceae examined here have expanded or highly modified IR/LSC boundaries compared to autotrophic *Dioscorea elephantipes* (*trnH*-GUG in the IR; Table 4.2), an autotrophic representative of another family (Dioscoreaceae) in the order. The two surveyed photosynthetic members of Burmanniaceae, *Burmannia bicolor* and *B. capitata*, have no inferred changes in their IR/SSC boundary compared to the canonical IR arrangement in terms of gene content, but have additional genes in their IR at the IR/LSC boundaries (*rps19* and *rpl22*). *Campylosiphon* has retained the canonical *trnN*-GUU boundary at its IR/SSC end, and includes an additional gene (*rpl32*, likely after two inversions). There is an additional direct repeat of one of the IR regions that includes a pseudogenized *rpl32* gene in this species (Fig. 4.1). *Burmannia itoana* has a novel (or highly shifted) IR that incorporates genes from the typical IR (*rpl22*, *rps19*, *rpl2*, *trnI*-CAU), SSC (*ndhD*, *rpl32*, *rps15*, and *ycf1*), and parts of the adjacent IR (*ndhB*,

rrn5, 4.5, 23, 16). The redefined SSC region in this species includes pseudogenes of *rps7*, *rrn16* and intact copies of *rps7* and 3'-*rps12* (in that gene order).

In addition to having complete inversion of its IR copies, the *rps19*, *rpl2*, *rps7* and 3'*rps12* genes in *Arachnitis*, which are typically found in the IR, instead comprise a novel SSC in this species (which includes *rps3*, usually found in the LSC). The IR of *Arachnitis* is reduced to four rDNAs and parts of *rpl16*. The IR is lost completely in *Corsia boridiensis*, the other member of Corsiaceae (Fig. 4.1, Table 4.2). In Iridaceae, the IR of *Geosiris* (Iridaceae) has the same gene content and is colinear with its autotrophic relative, *Iris missouriensis* (although *ndhB* is pseudogenized in *Geosiris*; Figs. 4.1, 4.2). Both taxa have expanded IR/LSC boundaries (*rps19*) compared to the canonical angiosperm IR. In Petrosaviaceae, the IR/SSC boundary of *Petrosavia sakurarii* (which retains an IR, unlike its close relative *P. aff. sakurarii*) includes additional genes compared to the canonical angiosperm IR, specifically a pseudogenized *ccsA* gene and *trnL-UAG* (both genes are found in the SSC in the canonical IR). These two genes are part of a four-gene inversion (*ccsA-trnL-rpl32-ndhF*) that appears to have occurred in *P. sakurarii* compared to the closely related autotroph *Japonolirion*, but not in *P. stellaris* (Logacheva et al. 2014). The IR at the IR/LSC boundary in *P. sakurarii* has six fewer genes compared to the canonical IR, but the inclusion of a portion of *matK* in the IR at the IR/LSC boundary also points to an episode of expansion into the LSC, with subsequent gene losses.

4.4.6 Phylogenetic placement of mycoheterotrophic lineages

The likelihood analysis (see Fig. 4.5, which depicts the shortest tree from the partitioned ML analysis) strongly supports the monophyly of Corsiaceae (*Arachnitis* and *Corsia*) and their placement as the sister group of Campynemataceae in Liliales. *Sciaphila* (Triuridaceae) is strongly supported as the sister group of Cyclanthaceae-Pandanaceae in Pandanales (see also Chapter 3). The genus *Petrosavia* (considering two newly sequenced taxa and a previously published sequence) is strongly supported as the sister group of photosynthetic *Japonolirion*, and this clade (Petrosaviaceae, Petrosaviales) is strongly supported as the sister group of all monocots except Alismatales (Fig. 4.5). Fully mycoheterotrophic *Geosiris* is strongly supported as the sister group of the only other member of Iridaceae sampled here (i.e., *Iris*, Fig. 4.5), and the family has a well-supported placement in Asparagales. Burmanniaceae are strongly supported as the sister group of Dioscoreaceae (represented here by *Dioscorea*) in Dioscoreales.

Phylogenetic relationships within Burmanniaceae are moderately to strongly supported: *Apteria* and *Gymnosiphon* are strongly supported as sister taxa, and *Campylosiphon* is moderately supported as their sister group. *Burmannia itoana* and *B. bicolor* are inferred to be sister taxa among sampled taxa, with moderate support, and *B. capitata* is moderately well supported as their sister group. An unpartitioned likelihood analyses recovered an identical topology and similar support values (Fig. C.15).

The phylogeny inferred from the parsimony analysis (Fig. C.16) is generally similar to the one from likelihood, with a major exception. Several full mycoheterotrophs on long branches are apparently pulled into the most rapidly evolving clade in Burmanniaceae (*Apteria-Gymnosiphon*): *Arachnitis* (Corsiaceae) is depicted as the sister group of this group, and *Sciaphila* (Triuridaceae) as the sister group of these three. The remaining members of Burmanniaceae are then the sister group of this “fast” clade. The remaining fully mycoheterotrophic taxa have phylogenetic placements consistent with the likelihood analyses (i.e., *Corsia*, Corsiaceae; *Geosiris*, Iridaceae; *Petrosavia*, Petrosaviaceae; Figs. 4.5, C.15). There were otherwise no major differences in monocot relationships between the likelihood and parsimony analyses.

4.5 Discussion

4.5.1 Phylogenetic placement of mycoheterotrophic monocot lineages

Likelihood-based phylogenomic analysis provide well-supported placements of all included mycoheterotrophic lineages, despite substantial rate elevation in many of them (Fig. 4.5), patchy recovery of genes across the taxa (for example, *Apteria*, *Arachnitis* and *Sciaphila* have only 18, 20 and 19 genes protein-coding and rDNA genes, respectively; Fig. 4.2, Table 4.1), and regardless of whether complex or simple data partitioning schemes are employed (cf. Figs. 4.5, C.15). Inferred placements of *Geosiris* and *Petrosavia* spp. as members of Iridaceae (Asparagales) and Petrosaviales, respectively, (Fig. 4.5) are consistent with my findings in Chapter 2 and with previous studies (Reeves et al. 2001; Fay et al. 2000; Fuse and Tamura 2000; Davis et al. 2004; Chase et al. 2006; Goldblatt et al. 2008). My analysis resolves several problematic placements that were ambiguous or poorly supported (e.g., for Burmanniaceae in

Chapter 2 here, where I recovered only one to three plastid genes per taxon). These placements are also consistent with other studies (see below).

In contrast, parsimony analyses pool all members of Burmanniaceae (including photosynthetic taxa), *Arachnitis* of Corsiaceae and *Sciaphila* (Triuridaceae) into a single “fast” clade (Fig. C.16), similar to the corresponding three-gene analyses (Fig. A.4, Chapter 2), but here with strong (presumably misleading) bootstrap support. This result is likely to be a function of very strong long-branch attraction in the case of parsimony (see Felsenstein 1978; Hendy and Penny 1989), a consequence of rate elevation in retained plastid genes (see also Chapter 2).

Molecular phylogenetic analyses have yielded divergent placements for several monocot mycoheterotroph families. For example, a study based on the nuclear 26S rDNA locus recovered Corsiaceae as polyphyletic, with *Corsia* inferred to be the sister group of *Campynema* (Liliales), and *Arachnitis* recovered as the sister group of *Thismia* (Thismiaceae), embedded in a clade otherwise comprising members of Burmanniaceae and Thismiaceae (Neyland and Hennigan 2003). The latter result was likely a function of long-branch attraction resulting from limited taxon sampling, elevated substitution rates, and the use of parsimony (e.g., Chapter 2). Other studies have found *Arachnitis* to be weakly supported as sister to Liliales (e.g., Davis et al. 2004; Fay et al. 2006; Petersen et al. 2013), although interpretation of some of these studies is limited by poor outgroup sampling. A phylogenomic analysis that includes plastid gene sets extracted from the full plastome data presented here (Mennes et al. 2015) resolved Corsiaceae as the sister group of Campynemataceae (Liliales) with 100% likelihood-based bootstrap support. I recovered the same placement here using increased taxon sampling in Liliales (Fig. 4.5) (2015; see also Givnish et al. 2016).

Burmanniaceae *s.l.* have been treated as two distinct subtribes, Burmannieae and Thismieae (e.g., Jonker 1938), or as separate families, Burmanniaceae and Thismiaceae (e.g., Dahlgren et al. 1985; APG 1998). Modern treatments (APG 2003, 2009, 2016) combine Burmanniaceae and Thismiaceae as a single family. However, this appears to have been based on phylogenetic analyses (Caddick et al. 2002a, b) that included contaminant or completely missing plastid sequences, as noted in Chapter 2. Two studies that considered nuclear and mitochondrial data (Merckx et al. 2006, 2009) instead placed Thismiaceae as the sister group of *Tacca* (Taccaceae) and Burmanniaceae as the sister group of *Dioscorea* (Dioscoreaceae), supporting recognition of Burmanniaceae and Thismiaceae as separate families. I was unable to recover

plastid data from Thismiaceae (attempts from several taxa were unsuccessful), but a plastid genome has recently been reported for the family by Gwynne Lim (Cornell University, pers. comm.), and will be important to include in future analyses. Nonetheless, the few-gene analyses presented in Chapter 2 also suggest that Burmanniaceae and Thismiaceae are not each other's closest relatives. My study also provides an initial phylogenetic framework for Burmanniaceae based on whole plastome data, to which additional taxa could usefully be added (see Chapter 2 for a more broadly sampled plastid phylogeny based on several plastid loci). Merckx et al. (2006, 2008) inferred six to eight independent losses of photosynthesis in Burmanniaceae *sensu stricto*, and one-two losses in Thismiaceae. Multiple (at least two) independent losses of photosynthesis in Burmanniaceae are also supported here, based on my relatively limited taxon sampling of Burmanniaceae (Fig. 4.5).

Triuridaceae were sometimes placed in their own order, Triuridales, based on their morphological distinctiveness (e.g., Dahlgren and Clifford 1982; Maas-van de Kamer and Weustenfeld 1998). Their first clear placement came from the study of Chase et al. (2000), who recovered *Sciaphila* (Triuridaceae) as being closely related to Cyclanthaceae and Pandanaceae (Pandanales) in a combined analysis of several plastid genes and 18S rDNA (only the latter was recovered for *Sciaphila*). A phylogenetic analysis based on 39 morphological characters placed Triuridaceae within Stemonaceae, also in Pandanales (Rudall and Bateman 2006). More recently, Mennes et al. (2013) used mitochondrial and nuclear evidence to place Triuridaceae as the sister group of Pandanaceae-Cyclanthaceae-Stemonaceae with low support. Here I inferred a strongly supported placement of *Sciaphila densiflora* (Triuridaceae) as the sister group of Cyclanthaceae-Pandanales (Fig. 4.5), as in Chapter 3 (Fig. 3.4). This strongly supported placement was apparently not influenced by the inclusion of other rapidly evolving mycoheterotrophs here, in contrast to the analyses in Chapter 2 based on a few plastid genes, although a subset of analyses in that chapter recovered the same phylogenetic placement (e.g., Fig. 2.3e). Thus, including a sufficient number of characters appears to compensate for the effect of rapidly evolving characters here.

4.5.2 Structural diversity

4.5.2.1 Structural diversity: overall colinearity

Most published reports of plastome structure in heterotrophic plants depict relatively little change, apart from gene loss: the following taxa are essentially colinear with autotrophic relatives: *Cuscuta reflexa*, Convolvulaceae (Funk et al. 2007), *Epifagus virginiana*, Orobanchaceae (Wolfe et al. 1992), *Aneura mirabilis*, Aneuraceae (Wickett et al. 2008), *Corallorhiza* spp., Orchidaceae (Barrett and Davis 2014); *Cistanche deserticola*, Orobanchaceae (Li et al. 2013), *Epipogium aphyllum*, Orchidaceae, Schelkunov et al. 2015). A few plastomes are known that have intermediate levels of reorganization (e.g., *Cuscuta gronovii*, Convolvulaceae, Funk et al. 2007; *Rhizanthella gardneri*, Orchidaceae, Delannoy et al. 2011, some Orobanchaceae, Wicke et al. 2013). Only a few appear to have high levels of reorganization (e.g., *Petrosavia stellaris*, Petrosaviaceae, Logacheva et al. 2014; *Epipogium roseum*, Orchidaceae, Schelkunov et al. 2015). Consistent with this emerging picture, many of the mycoheterotrophic plastomes I sequenced are also nearly or completely colinear with close green relatives (Figs. 4.4, C.5, C.7, C.8, C.11, C.12, C.14), despite often massive deletion of photosynthesis genes (Fig. 4.4). *Sciaphila* (Triuridaceae) has two moderately sized inversions compared to a close autotrophic relative but is otherwise colinear (Fig. C.14, see also Fig. 3.3). None of the plastomes I sequenced were as severely rearranged as *Petrosavia stellaris* (Logacheva et al. 2014), summarized here in Fig. C.12 (note that the other *Petrosavia* plastid genomes are nearly colinear with *Japonolirion*, a green relative in the same family; Fig. C.12). It remains to be seen how unusual this pattern of structural evolution is in heterotrophs in general. It should also be noted that substantial plastome rearrangements have also been observed in a few autotrophic lineages, including Geraniaceae (Chumley et al., 2006; Guisinger et al. 2011, Weng et al. 2014), various Fabaceae (e.g., Cai et al. 2008; Martin et al. 2014) and Campanulaceae (e.g., Haberle et al. 2008; Knox 2014). In autotrophic Geraniaceae, it was suggested that mutations in the nuclear-encoded proteins involved with suppression of aberrant plastid recombination (e.g., Whirly proteins, Maréchal et al. 2009) may result in erroneous genome rearrangements between repetitive sequences, such as homopolymers (e.g., Guisinger et al. 2008, 2011; Maréchal et al. 2009). The mechanisms underpinning genome rearrangement in heterotrophic lineages remain unclear, although it has been hypothesized that they also likely

reflect relaxed selective pressure and the nonfunctionalization of genes in the plastome (Wicke et al. 2011), leading to increased generation of repeated sequences and anomalous plastid recombination between the repeats in heterotrophs compared to autotrophic lineages.

4.5.2.2 Structural diversity: IR loss

Several mycoheterotrophs examined here have lost a copy of their plastid inverted repeat (IR) region (*Corsia* in Corsiaceae, *Apteria* and *Gymnosiphon* in Burmanniaceae, *Petrosavia* aff. *sakurarii* in Petrosaviaceae, and *Sciaphila* in Triuridaceae; Fig. 4.1). These collectively represent at least four independent IR losses, as their closest relatives in each case have IRs (Fig. 4.5). Inverted repeat loss has been observed in a few other heterotrophic lineages (i.e., holoparasitic *Phelipanche ramosa* and *Conopholis americana* in Orobanchaceae, Wicke et al. 2013, mycoheterotrophic *Monotropa hypopitys*, Gruzdev et al. 2016). Thus, heterotrophs may be more prone to this major structural change than autotrophs, where it has been observed only sporadically (examples include cupressophyte conifers, Pinaceae and Fabaceae; Lavin et al. 1990; Wu et al. 2011). The presence of an IR has been hypothesized to stabilize genome structure during recombination-dependent plastome replication (e.g., Palmer and Thompson 1982; Maréchal and Brisson 2010), and so we would expect IR loss to be associated with major rearrangements in mycoheterotrophs. However, only one of the IR-lacking plastomes is highly rearranged (*Corsia* in Corsiaceae, Fig. C.9; vs. *Apteria* and *Gymnosiphon* in Burmanniaceae, *Sciaphila* in Triuridaceae, and *Petrosavia* aff. *sakurarii*, in Petrosaviaceae, see Figs 4.1, C.7, C.8, C.12, C.14) and several full mycoheterotrophs with inverted repeats are quite highly rearranged (e.g., *Arachnitis* in Corsiaceae, *Burmannia itoana* in Burmanniaceae, Figs 4.1, C.6, C.10; *Petrosavia stellaris*, Logacheva et al. 2014). (Of these, the loss in *Petrosavia* aff. *sakurarii* is presumably very recent indeed, Fig. 4.5, as the other representative of this species has retained an IR; Fig. 4.1). This suggests that the relationship between IR presence or absence and genome structural rearrangements, if it exists, is not a simple one.

4.5.2.3 Structural diversity: IR boundaries and composition

Plastid inverted repeat (IR) boundaries are generally stable in autotrophic monocots, although expansion of the IR to include the *trnH-rps19* gene cluster has occurred multiple times (Wang et al. 2008; Zhu et al. 2015). In contrast, IR boundary shifts occurred relatively frequently in

monocot mycoheterotrophs (Table 4.2), and the number of genes involved (Table 4.1) may also exceed those seen in autotrophic monocots that have some boundary shifts compared to canonical IRs (several of the latter are reported in Table 4.1). I documented four types of IR in the 10 mycoheterotrophic taxa here: (1) IR boundaries are identical to their closest autotrophic relatives, and the retained genes are colinear with them (*Geosiris*, Iridaceae); (2) expansion and/or contraction of IR boundaries has occurred, with colinearity maintained in the IR (e.g., *Arachnitis*, *Campylosiphon*, *Petrosavia sakuraii*); (3) IR gene content and arrangement varies greatly compared to autotrophs (*Burmannia itoana*), and an SSC may be formed that is novel in terms of its composition; (4) Gain of an additional direct copy of one of the IR copies, seen in *Campylosiphon* (Burmanniaceae). The latter is particularly surprising, given that large direct repeats are thought to be selected against, as they are thought to be destabilizing by leading to plastome recombination (Maréchal and Brisson 2010).

4.5.3 Models of gene loss and retention in heterotrophic plants

The plastomes of autotrophic land plants have a high degree of conservation in terms of gene content. They typically code for 79 protein-coding, four rDNA and 30 tRNA genes (e.g., Palmer et al. 1985; Wicke et al. 2011). Full mycoheterotrophs are expected to show extensive pseudogenization and loss of genes due to the release of selective pressure on the genes involved in photosynthesis (e.g., Krause 2008). Barrett and Davis (2012) proposed an ordered trajectory for gene loss based on sequenced plastomes of heterotrophic plants, as follows: (1) an initial loss of NAD(P)H dehydrogenase (*ndh*) genes, perhaps before the full loss of photosynthesis, (2) loss of photosynthesis genes (*cemA*, *ccsA*, *lhbA*, *psa*, *psb*, *pet*, *rbcL*, *ycf3*, *ycf4*), all associated with loss of this metabolic function; (3) loss of the plastid-encoded RNA polymerase (*rpo*) genes that contribute extensively to transcription of photosynthetic genes in autotrophs; (4) loss of plastid ATP synthase (*atp*) genes, and; (5) gradual loss of ribosomal proteins and other “housekeeping” genes (*rpl*, *rps*, *rrn*, *trn*, *accD*, *clpP*, *matK*, *ycf1*, *ycf2*). The genes for the plastid ATP synthase complex may be retained after loss of other photosynthetic genes because of secondary roles that the complex plays outside photosynthesis (Kohzuma et al. 2011; Kamikawa et al. 2015). A minor re-modeling of the Barrett and Davis (2012) hypothesis (Barrett et al. 2014) combined the loss of *rpo* genes with photosynthesis genes, and of *atp* genes with plastid housekeeping genes. Here I review gene absences in the newly and previously sequenced heterotroph plastid genomes

in the context of these hypotheses. I show that in general the patterns of gene absence are more consistent with the Barrett and Davis (2012) hypothesis than the modified one proposed by Barrett et al. (2014).

4.5.3.1 Loss of *ndh* genes

The plastid NAD(P)H dehydrogenase complex is composed of 15 protein subunits, 11 of which are plastid-encoded (i.e., *ndhA*, B, C, D, E, F, G, H, I, J and K; Martin and Sabater 2010). The complex is involved in cyclic electron transport, and appears to help functionally adapt the photosynthesis machinery to stressful conditions, including nutrient deprivation, high temperature, water shortage or intense light (Horváth et al. 2007; Rumeau et al. 2007; Suorsa et al. 2009; Martin and Sabater 2010; Wicke et al. 2011). It has been suggested that the plastid *ndh* genes can be functionally replaced by nuclear equivalents (e.g. Blazier et al. 2011). However, Ruhlman et al. (2015) found no evidence of this in reported plant lineages with *ndh* loss. The *ndh* complex appears to have been lost in several mycoheterotrophic, holo- and hemiparasitic lineages (Fig. 4.2). However, as the genes have also been lost in fully autotrophic lineages (e.g., Cactaceae, Gnetales, Geraniaceae, Lentibulariaceae, Pinaceae and multiple submerged aquatic lineages in Alismatales, see Wakasugi et al. 1994; Braukmann et al. 2009, Wu et al. 2010; Blazier et al. 2011; Wicke et al. 2013; Ruhlman et al. 2015; Sanderson et al. 2015; Ross et al. 2016), only a subset of taxa that have lost *ndh* genes are likely on a track to heterotrophy.

Barrett et al. (2014) proposed that *ndh* gene loss may occur before the loss of photosynthesis in low light conditions in the understory, possibly in plants that are already partially mycoheterotrophic (e.g., orchids). Loss of function of plastid NAD(P)H dehydrogenase may block these plants from reversion to high light conditions, which may inevitably ensure increasing reliance on fungal partners for plant nutrition and eventually the loss of photosynthesis. Examples of partial mycoheterotrophs that have lost *ndh* genes include multiple lineages of Orchidaceae (e.g., several lineages in Cypripedioideae, Epidendroideae, and Vanilloideae, Neyland and Urbatsch 1996; Chang et al. 2006; Wu et al. 2010; Barrett et al. 2014; Lin et al. 2015; Ruhlman et al. 2015; see Fig. 4.2). I infer that loss of function in NAD(P)H dehydrogenase has likely also occurred in photosynthetic *Burmannia capitata* (it has multiple plastid encoded *ndh* genes with internal stop codons Fig. 4.2), which may mark it as a partial mycoheterotroph, an hypothesis that is also supported by its morphological reduction (Merckx et

al. 2010); it thus appears to be a Burmanniaceae analog to *ndh* loss in photosynthetic Orchidaceae. The other autotrophic *Burmannia* species examined here, *B. bicolor*, has retained *ndh* open reading frames (Fig. 4.2), as have multiple photosynthetic orchid lineages (Neyland and Urbatsch 1996). All full mycoheterotrophs examined here have lost the complex.

4.5.3.2 Extensive loss of photosynthesis-related genes in heterotrophic lineages

Light-dependent and independent reactions in land plants require photosystems I and II (*psa*, *psb*), assembly factors (*ycf3*, *ycf4*), light harvesting complex proteins (*lhbA*), cytochrome *b₆f* (*pet*), Rubisco (*rbcL*) and other proteins involved in CO₂ uptake (*cemA*), and cytochrome *c* biogenesis protein (*ccsA*). Nearly all of these genes are lost or pseudogenized in all full mycoheterotrophs surveyed here, consistent with the Barrett and Davis (2012) or Barrett et al. (2014) hypotheses. However, *rbcL* (which codes for the large subunit of Rubisco) is retained as an open reading frame in all three *Petrosavia* taxa examined here (reported previously for *Petrosavia stellaris* by Chase et al. 2000; Logacheva et al. 2014). This may be consistent with this gene being functional in these taxa, despite loss of photosynthesis. The primary role of Rubisco (encoded by plastid *rbcL* and nuclear *rbcS*) in autotrophs is the carboxylation of ribulose-1,5-bisphosphate into phosphoglycerate in the Calvin cycle. Various heterotrophic taxa have retained *rbcL*, including *Cuscuta* spp. (e.g., Machado and Zetsche 1990; Funk et al. 2007), the liverwort *Aneura mirabilis* (Wickett et al. 2008), some Orobanchaceae (e.g., Randle and Wolfe 2005; Wickett et al. 2011; Wicke et al. 2013), and several heterotrophic protist lineages, including the euglenid *Astasia longa* (Gockel and Hachtel 2000; Seimeister and Hachtel 1990), *Cryptomonas paramecium* (Donaher et al. 2009) and *Prototheca wickerhamii* (Knauf and Hachtel 2002). Wickett et al. (2011) also noted the expression of plastid *rbcL* and nuclear *rbcS* in some holoparasitic Orobanchaceae.

A lag in the accumulation of stop codons and indels after initial loss of photosynthesis is expected (Leebens-Mack and dePamphilis 2002). However, this is an unlikely explanation for the persistence of putatively functional *rbcL* in multiple unrelated lineages. Instead, it may be retained in heterotrophic lineages because of secondary metabolic roles for Rubisco, including the synthesis of serine and glycine (Siemeister and Hachtel 1990), lipid biosynthesis (Schwender et al. 2004; McNeal et al. 2007) and possibly other cryptic roles (e.g., Osmond et al. 1975; Bungard 2004, Schwender et al. 2004). These functions are presumably eventually eliminated or

replaced, accounting for the absence of *rbcL* in most full mycoheterotrophs. If so, *rbcL* would be expected to be initially retained in lineages that have lost photosynthesis, but eventually to be lost. The possibility of a delayed loss is not directly accounted for in the Barrett and Davis (2012) or Barret et al. (2014) hypotheses, which could be modified accordingly.

4.5.3.3 Retention and loss of plastid-encoded RNA polymerase genes

Most plastid genes in autotrophic land plants are transcribed by a plastid-encoded RNA polymerase (PEP), including all of the photosynthetic genes (e.g., Hajdukiewicz et al. 1997). This polymerase is coded for by plastid *rpoA*, *rpoB*, *rpoC* and *rpoC1* genes. Many plastid genes (photosynthetic and non-photosynthetic) can also be transcribed by a nuclear-encoded RNA polymerase (NEP), although PEP is the predominant RNA polymerases in mature, active chloroplasts (Zhelyazkova et al. 2012). PEP genes have been lost or pseudogenized in most non-photosynthetic plants (the NEP likely serves as an alternative method for expression of expressed genes in the absence of PEP; Liere et al. 2011; Zhelyzakova et al. 2012). However, all four PEP subunit genes have been retained as open reading frames in *Petrosavia sakurarii* and *Petrosavia* aff. *sakurarii* (Petrosaviaceae), a contrast with their close relative *Petrosavia stellaris* (Fig. 4.2; Logacheva et al. 2014). This pattern supports the hypothesized gene-loss trajectory in Barrett and Davis (2012), but not Barrett et al. (2014), and also supports the idea that these genes experience delayed loss after the loss of most photosynthesis genes. Their retention may be linked to the retention of ATP synthase genes in these members of *Petrosavia*; however, this seems unlikely as *P. stellaris* has retained ATP synthase but not PEP (Fig. 4.2; Logacheva et al. 2014), see the next section.

4.5.3.4 Retention and loss of ATP synthase genes

Plastid ATP synthase in land plants is encoded by six plastid-encoded subunits (*atpA*, B, E, F, H and I), and three nuclear-encoded subunits (*atpC*, D and G). In photosynthetic plants, the main function of this enzyme is to catalyze the formation of ATP using the proton gradient, which is driven by the electron transport chain during photosynthesis, leading to generation of energy-rich compounds. ATP synthase genes are generally lost in heterotrophs, as is the case in most of the full mycoheterotrophs surveyed here (Fig. 4.2). However, retention of *atp* genes in open reading frame in full mycoheterotrophs here (*Petrosavia* spp., Fig. 4.2) and in several other full

mycoheterotrophs and holoparasites (*Aneura mirabilis*, Wickett et al. 2008; *Orobancha crenata*, *O. gracilis* and *Phelipanche ramosa*, Wicke et al. 2013) supports at least the initial retention of an active ATP synthase for processes that are not linked to photosynthesis. Additional plastid ATP synthase retentions are also known in heterotrophic protists: *Cryptomonas paramecium*, a non-photosynthetic cryptomonad alga, has also retained all of the required *atp* genes (including *atpC*, E and G found as nuclear genes in land plants), except *atpF* (Donaher et al. 2009), and the enzyme has been retained in the non-photosynthetic unicellular alga *Prototheca wickerhamii* (Knauf and Hachtel 2002), in which detectable levels of plastid-encoded *atp* transcripts have also been found.

Kohzuma et al. (2011) and Kamikawa et al. (2015) proposed that ATP synthase may be retained for ATP hydrolysis to maintain a proton gradient to power the twin arginine translocator (Tat) system for translocation of proteins or ion transport into the thylakoid lumen. Kamikawa et al. (2015) noted that the plastome of *C. paramecium* also codes for TatC, a core protein in the Tat system. They also documented the presence of homologs of plastid-encoded *tatC* and *atpC* in the transcriptome of *Orobancha aegyptiaca* and hypothesized that plastid-encoded *atp* genes are likely to be retained in this species. The retention of ATP synthase to generate a proton gradient may be common in the initial stages of heterotrophy (Wicke et al. 2013; Kamikawa et al. 2015), and is captured in the hypothesized trajectory of gene proposed by Barrett and Davis (2012). The evidence from *Petrosavia* suggests, however, that ATP synthase loss may precede loss of housekeeping genes, which conflicts with the modified hypothesis laid out in Barrett et al. (2014), who conjectured a longer retention of these genes. As most full mycoheterotrophs surveyed here lack ATP synthase (Fig. 4.2), the secondary (non-photosynthetic) functions of this enzyme may tend to be eliminated or replaced eventually, as with Rubisco.

4.5.3.5 Housekeeping genes involved in plastid translation

Translation in the plastid organelle occurs independently from that in the nucleus and other organelles. The plastid ribosomal apparatus include ribosomes constructed from multiple small ribosomal subunit proteins. The usual plastid encoded subunits in autotrophs include ribosomal small subunit genes (*rps2*, 3, 4, 7, 8, 11, 12, 14, 15, 16, 18, 19), large subunit genes (*rpl2*, 14, 16, 20, 22, 23, 32, 33, and 36), genes for the plastid ribosomal RNAs (*rrn4.5*, 5, 16, and 23), plastid translation initiation factor 1 (*infA*) and 30 transfer RNAs. I did not recover *infA* in several of the

mycoheterotrophic lineages here (Fig. 4.2). This gene is one of the most mobile genes in the plastome and has been functionally transferred to the nucleus in multiple autotrophic lineages (Millen et al. 2001), which may also be the case here (this would require additional follow-up work to confirm). Most mycoheterotrophs that I sequenced have otherwise retained multiple genes in each category (Fig. 4.2). The common retained subset comprises 10 ribosomal proteins (*rpl2*, *rpl14*, *rpl16*, *rps3*, *rps4*, *rps7*, *rps8*, *rps12*, *rps14* and *rps19*) all four rDNAs (*rrn4.5*, *rrn5*, *rrn16*, *rrn23*), and two tRNA genes (*trnE-UUC* and *trnFM-CAU*).

Knockout experiments in *Nicotiana tabacum* have demonstrated that *rpl33*, *rpl36*, and *rps15* may not be completely essential in experimental plants (Rogalski et al. 2008b; Fleischmann et al. 2011). The ribosomal proteins considered “essential” for cell viability in *N. tabacum* are *rpl20*, *rpl22*, *rpl23*, *rpl32*, *rps2*, *rps3*, *rps4*, *rps14*, *rps16* and *rps18* (the effect of losing *rpl2*, *rpl14*, *rpl16*, *rps7*, *rps8*, *rps11*, and *rps19* has not been determined; Rogalski et al. 2006, Fleischmann et al. 2011; Krause and Scharff 2014). However, several genes considered essential have been lost from the plastid genome in heterotrophs. For example, *rpl20* is lost in *Epipogium*, Orchidaceae (Schelkunov et al. 2015), and *Gymnosiphon* and *Apteria* in Burmanniaceae here (Fig. 4.1, 4.2). *Rpl22* has been lost in other heterotrophic taxa with reduced plastomes (Fig. 4.2). Plastid-encoded ribosomal proteins have been functionally transferred to the nucleus in some autotrophic rosids (i.e., *rpl22* in *Pisum sativum*, *Castanea mollissima* and *Quercus rubra*, Gantt et al. 1991; Jansen et al. 2011; *rps23* in *Populus alba*, Ueda et al. 2007). *Rps16* has been lost from *Populus* and *Medicago* and has been replaced by a nuclear-encoded copy (originally a mitochondrial copy) that is dual-targeted to the mitochondria and plastids (Ueda et al. 2008). Functional transfer of ribosomal proteins may also be the case here in heterotrophic lineages. However, attempts at locating nuclear copies of genes encoding plastid-targeted ribosomal proteins in some heterotrophs have been unsuccessful (Delannoy et al. 2011; Krause and Scharff 2014).

The four rDNA genes (*rrn4.5*, 5, 16, and 23) have been retained in nearly all heterotrophic plant plastome sequenced to date (Barrett et al. 2014; Li et al. 2013; Wicke et al. 2013; Schelkunov et al. 2015, and see Chapters 3 and 5), with the exception of endoparasitic *Pilostyles* (Bellot and Renner 2015), which has apparently lost *rrn4.5* and *rrn5*. It is unknown whether the rDNA loci in *Pilostyles* have been functionally transferred to other genomes, or functionally replaced by genes that originate in the other plant genomes. However, endoparasites

may evolve under substantially different evolutionary constraints than other heterotrophic plants, and may no longer have active plastid translation (Bellot and Renner 2016).

The loss of the majority of plastid tRNAs in many lineages sequenced here (Fig. 4.2) may indicate that they are replaced by imported tRNAs (e.g. Alkatib et al. 2012; Wolfe et al. 1992) or functionally substituted by others that undergo “superwobbling” (Rogalski et al. 2008a). The *trnE*-UCC and *trnfM*-CAU loci are retained across all mycoheterotrophic taxa I sequenced, and are retained in all heterotrophic plants sequenced to date (Fig. 4.2). These genes encode tRNA^{Glu} and tRNA^{fMet}, respectively, and have roles outside peptide chain elongation, likely explaining their persistence in heterotrophic plastid genomes (Howe and Smith 1991; Barbrook et al. 2006; Krause 2008). The tRNA^{Glu} product is ligated to glutamate to form aminolevulinic acid (ALA), the precursor of tetrapyrrole biosynthesis, which is involved in production of chlorophyll, heme, siroheme and phytychromobilin, all thought to be vital compounds for plant growth and function (Tanaka and Tanaka 2007). The replacement of the plastid tRNA^{Glu} with a cytosolic counterpart may not be possible, as glutamate is activated by ligation of this tRNA. The tRNA moiety interacts with a very specific binding site with glutamyl-tRNA reductase which catalyzes the formation of glutamate 1-semialdehyde, the next substrate required for ALA synthesis (Barbrook et al. 2004). In *Engelena*, a point mutation in the part of the *trnE* gene that codes for the T-loop of the tRNA renders the chloroplast incapable of heme biosynthesis, although protein synthesis remains unaffected (Stange-Thomann et al. 1994). tRNA^{Glu} is also thought to be involved in transcriptional regulation, by repressing NEP activity during chloroplast development (Hanaoka et al. 2005). Barbrook et al. (2006) hypothesized that plastomes should be retained as a replicating mini-circle encoding a single *trnE* in very late evolutionary stages of heterotrophy. The holoparasitic *Rafflesia lagascae* has been reported to have lost its plastome entirely (Molina et al. 2014). However, in *Rafflesia leonardi*, there is evidence of (plastid?) *trnE* in the nucleus or mitochondrion (unpublished data cited by Nickrent et al. in Molina et al. 2014), suggesting that this gene might be functionally transferred to one of the other plant genomes.

Initiation of translation in organelles requires tRNA^{fMet} as the starting methionine for protein synthesis (Barbrook et al. 2006), and translation initiation factor 1 (*infA*). The tRNA^{fMet} moiety is required for translation in eubacteria, in plastids and mitochondria, and differs from the plastid tRNA^{Met} used in peptide elongation (Barbrook et al. 2006). Knockout experiments of these genes resulted in cell death in *Nicotiana tabacum*, and so the *trnfM*-CAU gene product was

thought to be essential for both autotrophic and heterotrophic growth; superwobbling is not considered to be sufficient to replace *trnM*-CAU (Alkatib et al. 2012). As with *trnE*, the *trnM* gene product appears to play an indispensable role in the plastome. The retention of these two tRNAs underpins Barbrook's "essential tRNAs" hypothesis to explain the long-term retention of plastomes in heterotrophs.

4.5.3.6 Retention and loss of other housekeeping genes

Several genes that are not involved in the plastid translation apparatus are also frequently or occasionally retained in full mycoheterotrophs (Fig. 4.2). The *accD* locus (which encodes a plastidic subunit of ACCase involved in fatty acid production) has been retained in every heterotrophic land plant sequenced so far (Barrett et al. 2014; Li et al. 2013; Wicke et al. 2013; Schelkunov et al. 2015; Bellot and Renner 2016 – see also Chapters 2, 3 and 5). Physiological studies in *Nicotiana tabacum* point to the *accD* gene product as being essential for plastid development and cell viability (Kode et al. 2005). The *clpP* locus (which encodes a plastidic subunit of Clp protease) has been lost in *Apteris*, Burmanniaceae (Fig. 4.2) and pseudogenized in *Phelipanche* (Orobanchaceae) (Wicke et al. 2013). The *clpP* gene product is essential for shoot development (Kuroda and Maliga 2003). Both *accD* and *clpP* have also been lost from the plastid genome in several autotrophic lineages (Jansen et al. 2007; see Straub et al. 2011; functional transfer of the *accD* to the nucleus has been reported in at least two autotrophic lineages (Fabaceae, Magee et al. 2010; Campulanaceae, Rousseau-Gueutin et al. 2013). Their loss may therefore not be directly attributed to heterotrophy.

There are eight plastid genes with group IIA introns (*trnA*-UGC, *trnI*-GAU, *trnK*-UUU, *trnV*-UAC, *atpF*, *clpP*, *rpl2*, and 3'-*rps12*; Vogel et al. 1999; Zoschke et al. 2010). The *matK* locus codes for a group IIA intron maturase, and is thought to be required for correct splicing of seven of these introns (i.e., *trnA*-UGC, *trnI*-GAU, *trnK*-UUU, *trnV*-UAC, *atpF*, *rpl2*, and 3'-*rps12* transcripts; Zoschke et al. 2010) but apparently not the *clpP* intron, which has structural differences from the other introns and may not require it for splicing (Zoschke et al. 2010). *MatK* is typically retained in full heterotrophs that retain group IIA introns (e.g., *Epifagus*, Wolfe et al. 1992; *Phelipanche*, Wicke et al. 2013), but see below; loss of some or all group IIA introns was observed by McNeal et al. (2009) in several species of *Cuscuta*, some of which have lost *matK*, and by Wicke et al. (2013) in Orobanchaceae. Elimination of *matK* from the plastid genomes

was thought to be possible only when group IIA introns are all lost (e.g., McNeal et al. 2009). I observed loss of *matK* in several taxa surveyed here that are in advanced stages of gene loss: *Apteria*, *Burmannia itoana* and *Gymnosiphon* in Burmanniaceae, and *Arachnitis* in Corsiaceae; *matK* loss has also been noted in the plastomes of several published orchid full mycoheterotrophs (Table C.3). Of the eight genes that normally contain group IIA introns, only *rpl2* and 3'-*rps12* are retained in *Apteria*, and both genes have also lost their group IIA introns, consistent with the observed *matK* loss in this taxon (Table C.3). However, *B. itoana* has kept *rpl2* with an intact group IIA intron, and *Arachnitis* and *Gymnosiphon* have both retained group IIA introns in their *rpl2* and 3'-*rps12* genes (the only group IIA-containing genes remaining in their plastomes). Several fully mycoheterotrophic orchids that lack *matK* (*Epipogium* spp. and *Rhizanthella*; Table C.3) also retain intact group IIA intron in *rpl2*. These observations may imply that both *rpl2* and 3'-*rps12* do not require a maturase for splicing, or that they are spliced by a nuclear-imported maturase. *Neottia* (Orchidaceae) retains introns in two tRNA genes in addition to *rpl2* and 3'-*rps12*. Logacheva et al. (2011) speculated that *matK* in this species is pseudogenized (as it has a highly divergent 5'-end; Logacheva et al. 2011). However, it still retains a conserved region at its 3'-end (pers. obs.), that includes the 'domain X' of *matK*, which is thought to contain the active RNA binding site (Mohr et al. 1993; Hilu and Liang 1997). I therefore predict that *matK* is still functional in *Neottia*.

The two largest genes in autotrophic plastomes are *ycf1* and *ycf2*, which have been lost in most taxa surveyed here that have severe gene loss (i.e., *Apteria*, *Burmannia itoana*, *Gymnosiphon*, Burmanniaceae; *Arachnitis*, *Corsia*, Corsiaceae; *Sciaphila*, Triuridaceae; Fig. 4.2). Drescher et al. (2000) considered *ycf1* to be an essential gene that is not involved in photosynthesis, but was unable to determine its exact role. Kikuchi et al. (2013), and Nakai (2015) later associated its function with the translocation of nuclear-encoded proteins across the plastid membrane, specifically TIC of the TOC/TIC machinery. It was found to encode one of the proteins comprising the TIC complex, and *ycf1* was subsequently proposed to be renamed as *tic214* by Nakai (2015) (but see de Vries et al. 2015). The role of *ycf2* is still unclear, but it may be involved in drought response, based on one expression profile of water-use efficiencies in bean cultivars (Ruiz-Nieto et al. 2015). Several autotrophic lineages have also lost these genes (e.g., *Asclepias*, Straub et al. 2011; Geraniaceae; Guisinger et al. 2011; Poales, Jansen et al. 2007; Guisinger et al. 2010), and functionally transferred equivalents have yet to be found in the

other plant genomes (Downie et al. 1994). The retention of *ycf1* and *ycf2* in some full mycoheterotrophs, but not in those with extreme gene loss (Table 4.2) suggests that the eventual fate of these genes is also loss from heterotrophic plastomes.

Table 4.1 Overview of plastome size and properties of mycoheterotrophic monocots and photosynthetic relatives. Fully mycoheterotrophic lineages are indicated with an asterisk (*).

Order/species	Size (bp)	Inverted repeat present?	# of genes in IR (total) ^a	Number of unique genes (total) ^b	Coding percentage ^c	GC content (percentage) ^d
Asparagales						
<i>Iris missouriensis</i>	153,084	Yes	19	79 / 4 / 30 (113)	59.9	36.8
* <i>Geosiris aphylla</i>	123,620	Yes	18	28 / 4 / 30 (62)	31.8	36.6
Dioscoreales						
* <i>Apteris aphylla</i>	15,715	No	n/a	14 / 4 / 3 (21)	78.31	28.91
<i>Burmannia bicolor</i>	156,590	Yes	19	78 / 4 / 30 (112)	58.47	32.20
<i>Burmannia capitata</i>	149,769	Yes	19	72 / 4 / 30 (106)	55.41	33.69
* <i>Burmannia itoana</i>	44,790	Yes	11	21 / 4 / 13 (28)	42.53	29.93
* <i>Campylosiphon congestus</i>	121,733	Yes	18	24 / 4 / 21 (49)	34.32	33.04
* <i>Gymnosiphon longistylus</i>	21,289	Yes	n/a	19 / 4 / 5 (28)	66.46	24.86
Liliales						
* <i>Arachnitis uniflora</i>	24,678	Yes	n/a	16 / 4 / 5 (25)	66.12	34.27
* <i>Corsia</i> cf. <i>boridiensis</i>	50,347	No	n/a	22 / 4 / 18 (44)	57.59	32.58
<i>Campynema lineare</i>	156,261	Yes	19	77 / 4 / 30 (111)	58.27	35.78
<i>Lilium superbum</i>	152,070	Yes	18	78 / 4 / 30 (112)	60.50	35.90
Pandanales						
<i>Carludovica palmata</i>	158,545	Yes	19	79 / 4 / 30 (113)	58.00	36.70
* <i>Sciaphila densiflora</i>	21,485	No	n/a	15 / 4 / 6 (28)	68.70	39.90

Order/species	Size (bp)	Inverted repeat present?	# of genes in IR (total) ^a	Number of unique genes (total) ^b	Coding percentage ^c	GC content (percentage) ^d
Petrosaviales						
<i>Japonolirion osense</i>	157,897	Yes	18	79 / 4 / 30 (113)	58.67	37.18
* <i>Petrosavia sakurarii</i>	111,638	Yes	10	42 / 4 / 28 (74)	50.80	37.36
* <i>Petrosavia</i> aff. <i>sakurarii</i>	93,106	No	n/a	43 / 4 / 28 (75)	52.46	37.25

^a n/a: not applicable (i.e., lacking an inverted repeat). Count excludes partial genes or pseudogenes (*3'-rps12* counted as a single gene here)

^b Protein-coding, rDNA and tRNA genes, respectively; count excludes pseudogenes *5'-rps12* and *3'-rps12* counted collectively as a single gene.

^c Percentage includes protein coding, rDNA and tRNA loci, excludes pseudogenes and includes only one copy of the inverted repeat (IR) region.

Table 4.2 Inverted repeated (IR) boundaries in fully mycoheterotrophic monocots and close green relatives. Following Zhu et al. (2015), the last full gene included in the IR at the SSC and LSC boundaries is indicated (see Fig. 4.1 for any additional partially duplicated genes). Numbers in parentheses indicate the number of additional genes included into the IR (+) or absent from it (-) compared to canonical angiosperm IR boundaries. Fully mycoheterotrophic lineages are indicated with an asterisk (*).

Family	Species	IR/SSC boundary ^a	IR/LSC boundary	Notes:
Burmanniaceae	* <i>Apteria aphylla</i>	-	-	No IRs
	<i>Burmannia bicolor</i>	<i>trnN</i> -GUU	<i>rps19</i> (+2)	-
	<i>B. capitata</i>	<i>trnN</i> -GUU	<i>rpl22</i> (+3)	-
	* <i>B. itoana</i>	<i>rrn16</i> (+10)	<i>rpl22</i> (+3)	<i>rpl22</i> , <i>rps19</i> , <i>rpl2</i> , <i>trnI</i> -CAU (IR), <i>ndhD</i> , <i>rpl32</i> , <i>rps15</i> , <i>ycf1</i> (SSC), and <i>ndhB</i> , <i>rrns5</i> , 4.5, 23, 16 (from adjacent IR) comprise this novel or highly shifted IR
	* <i>Campylosiphon congestus</i>	<i>trnN</i> -GUU	<i>rpl3</i> (+4)	A small inversion may account for incorporation of <i>rpl32</i> into the IR; one IR copy has an additional duplication
	* <i>Gymnosiphon longistylus</i>	-	-	No IRs
Corsiaceae	* <i>Arachnitis uniflora</i>	<i>rrn16</i> (-8)	<i>rrn5</i> (-2)	The IR of <i>Arachnitis</i> is inverted relative to canonical angiosperm IRs
	<i>Campynema lineare</i>	<i>trnN</i> -GUU	<i>rps19</i> (+2)	-
	* <i>Corsia</i> cf. <i>boridiensis</i>	-	-	No IRs
	<i>Lilium superbum</i>	<i>trnN</i> -GUU	<i>rps19</i> (+2)	-

Family	Species	IR/SSC boundary ^a	IR/LSC boundary	Notes:
Iridaceae	<i>Iris missouriensis</i>	<i>trnN</i> -GUU	<i>rps19</i> (+2)	-
	* <i>Geosiris aphylla</i>	<i>trnN</i> -GUU	<i>rps19</i> (+2)	-
Pandanales	<i>Carludovica palmata</i>	<i>trnN</i> -GUU	<i>rps19</i> (+2)	-
	* <i>Sciaphila densiflora</i>	-	-	No IRs
Petrosaviaceae	<i>Japonolirion osense</i>	<i>trnN</i> -GUU	<i>trnH</i> -GUG (+2)	-
	* <i>Petrosavia sakurarii</i>	<i>trnL</i> -UAG (+1)	<i>rps7</i> (-6)	pseudogene of <i>matK</i> in IR may indicate an expansion into the LSC, rather than a contraction; An inversion at the SSC/IR boundary (<i>ccsA</i> , <i>trnL</i> -UAG, <i>rpl32</i> , <i>ndhF</i>) may account for <i>ccsA</i> pseudogene in the IR
	* <i>Petrosavia</i> aff. <i>sakurarii</i>	-	-	No IRs

^a No. of additional/fewer genes compared to canonical IRs

Figure 4.1 Linearized plastome maps of mycoheterotrophic monocots and two photosynthetic representatives highlighted in green. Sizes are indicated below each plastome in bp. Horizontal grey bars represent 10 kb increments. The inverted repeat (IR) region is shown in black, duplicated IR region is shown in dashed lines. Genes are colour-coded by function, see inset scheme. Inferred pseudogenes are indicated with red labeling, and genes with introns are indicated with an asterisk (*).

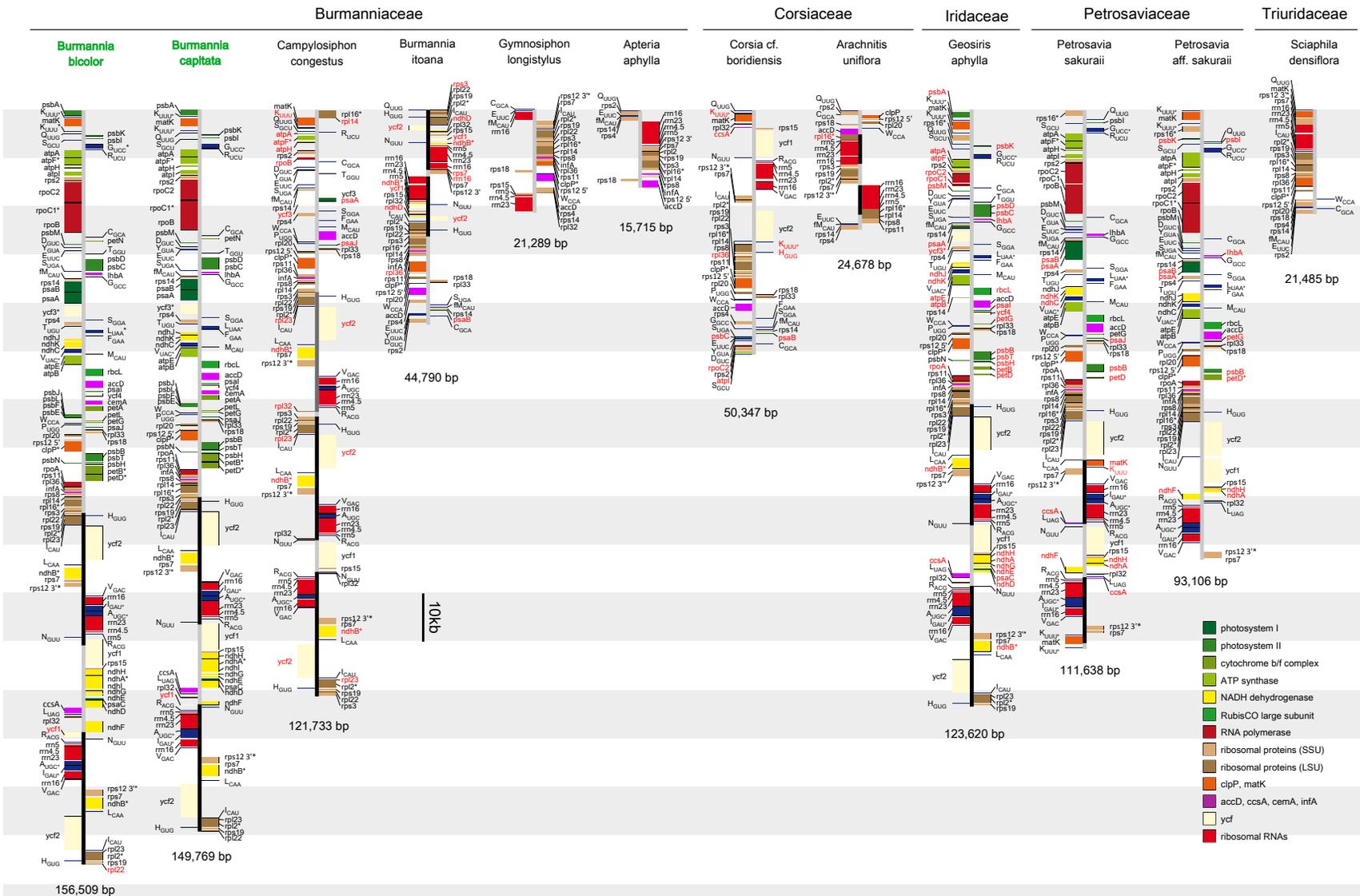


Figure 4.2 A “heat-map” showing gene loss and retention across full plastome gene sets from exemplar monocot mycoheterotrophs, eudicot hemi/holoparasites and autotrophic relatives (published plastomes: *Dioscorea elephantipes*, Hansen et al. 2007; *Cuscuta reflexa*, *C. gronovii*, Funk et al. 2007; *Phalaenopsis*, Chang et al. 2006; *Oncidium ‘Gower Ramsey’*, Wu et al. 2010, *Corallorhiza striata*, Barrett and Davis 2012; *Petrosavia stellaris*, Logacheva et al. 2014; *Cistanche deserticola*, Li et al. 2013; *C. phelypaea*, Wicke et al. 2013; *Neottia nidus-avis*, Logacheva et al. 2011; *Epifagus virginiana*, Wolfe et al. 1992; *Phelipanche ramosa*, Wicke et al. 2013; *Rhizanthella gardneri*, Delannoy et al. 2008; *Epipogium aphyllum*, *E. roseum*, Schelkunov et al. 2015; *Sciaphila densiflora*, Chapter 3). Plastomes are arranged by sizes from large to smallest. Dark blocks indicate missing genes, grey blocks indicate pseudogenes, and white blocks indicates retained genes. Abbreviations: AU=autotroph, Hemi=hemiparasite, Holo=holoparasite, MH=mycoheterotroph. **Phalaenopsis* and *Oncidium* defined here as non-heterotrophic taxa, and *Burmannia bicolor* and *B. capitata* as autotrophs, although all of these may be partial mycoheterotrophs at maturity.

Figure 4.3 Relationship between the number of retained unique genes and plastome size. The number of genes and plastome size includes only unique genes (excludes repeated genes found in the inverted repeat, IR, region). Plastome size is indicated in kb. A best-fit line is included that is based on the blue dots only.

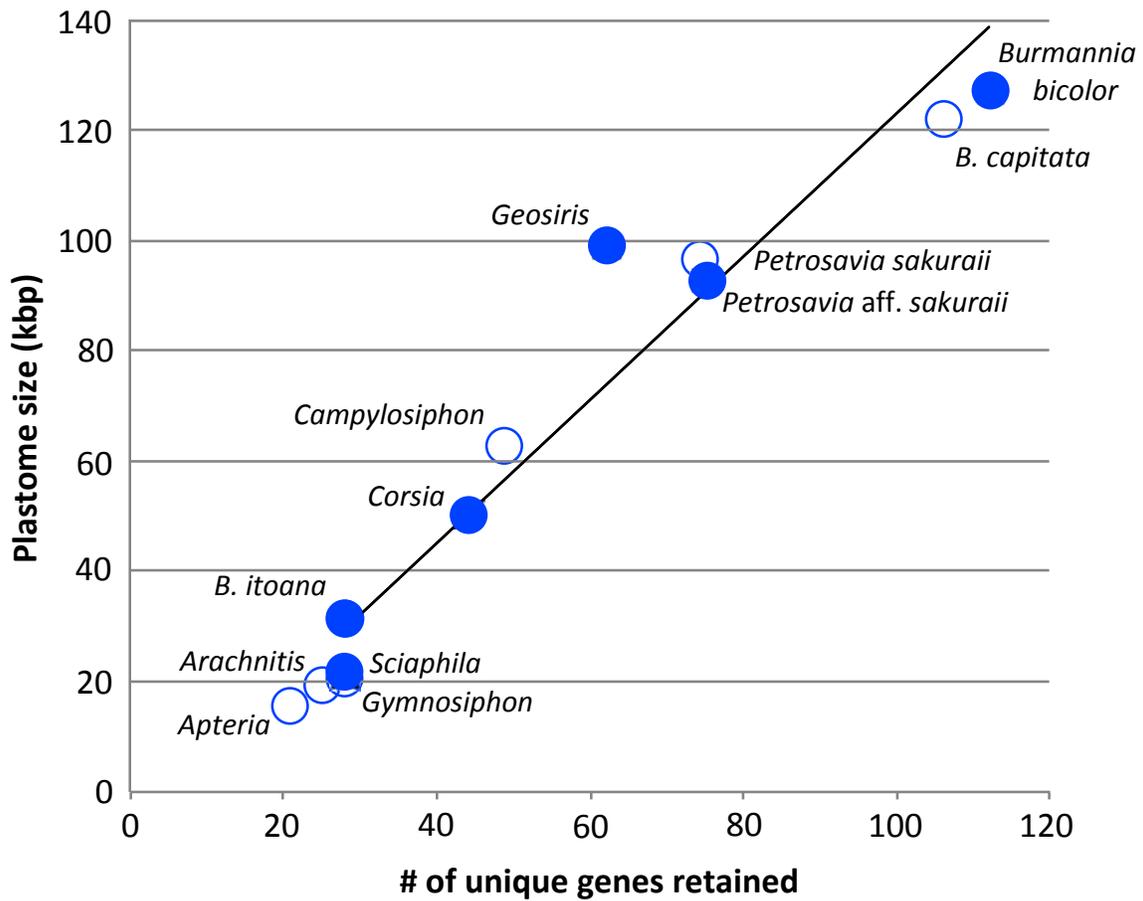


Figure 4.4 Diagrams showing approximate colinearity between plastomes of mycoheterotrophic monocots and autotrophic relatives (made using Kablammo); darker blue indicates blocks with higher sequence similarity; white areas indicate gene loss in the mycoheterotroph or areas of low similarity (e.g., intergenic spaces, introns, segments of highly variable genes such as *accD*, *ycf1*).

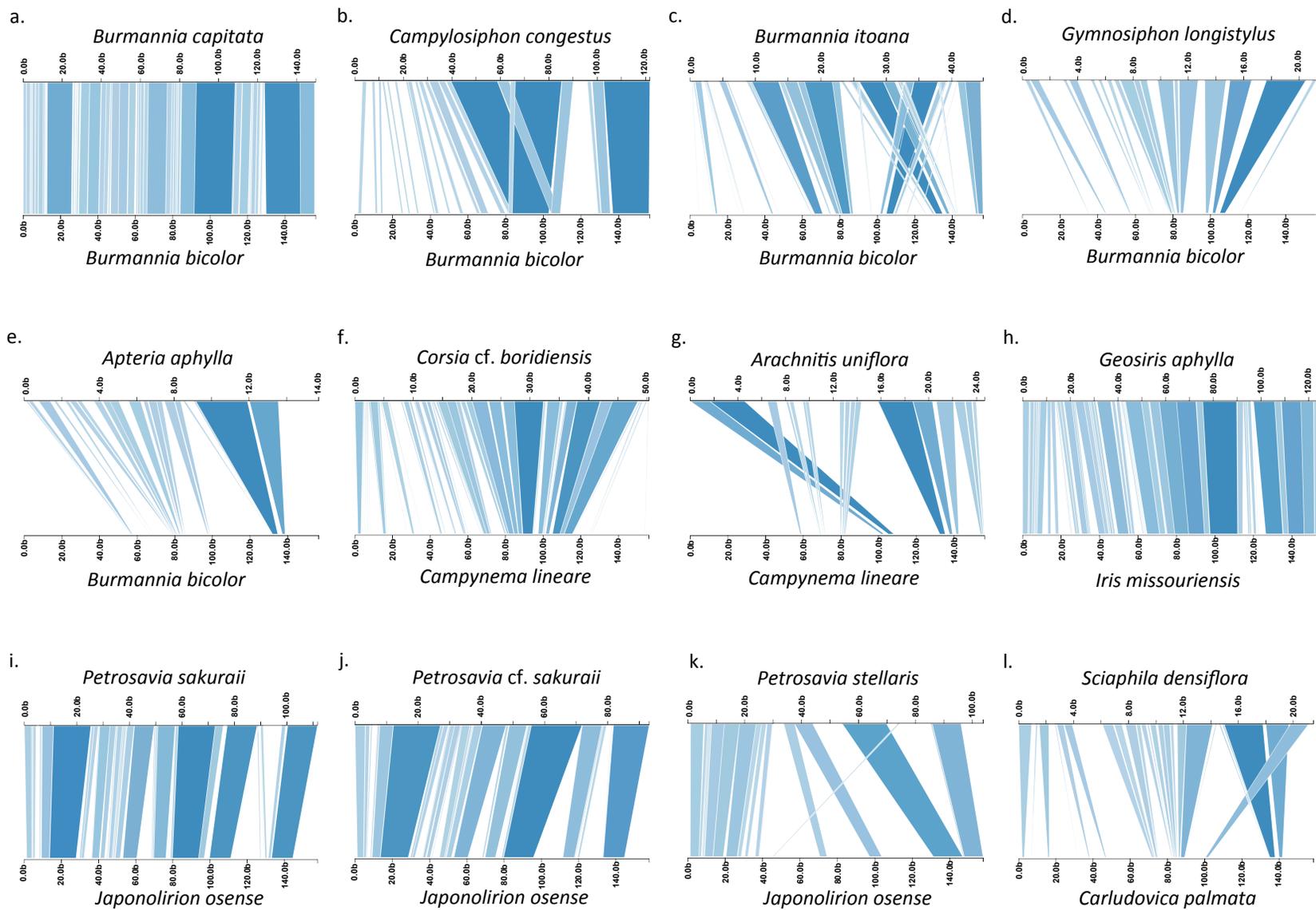
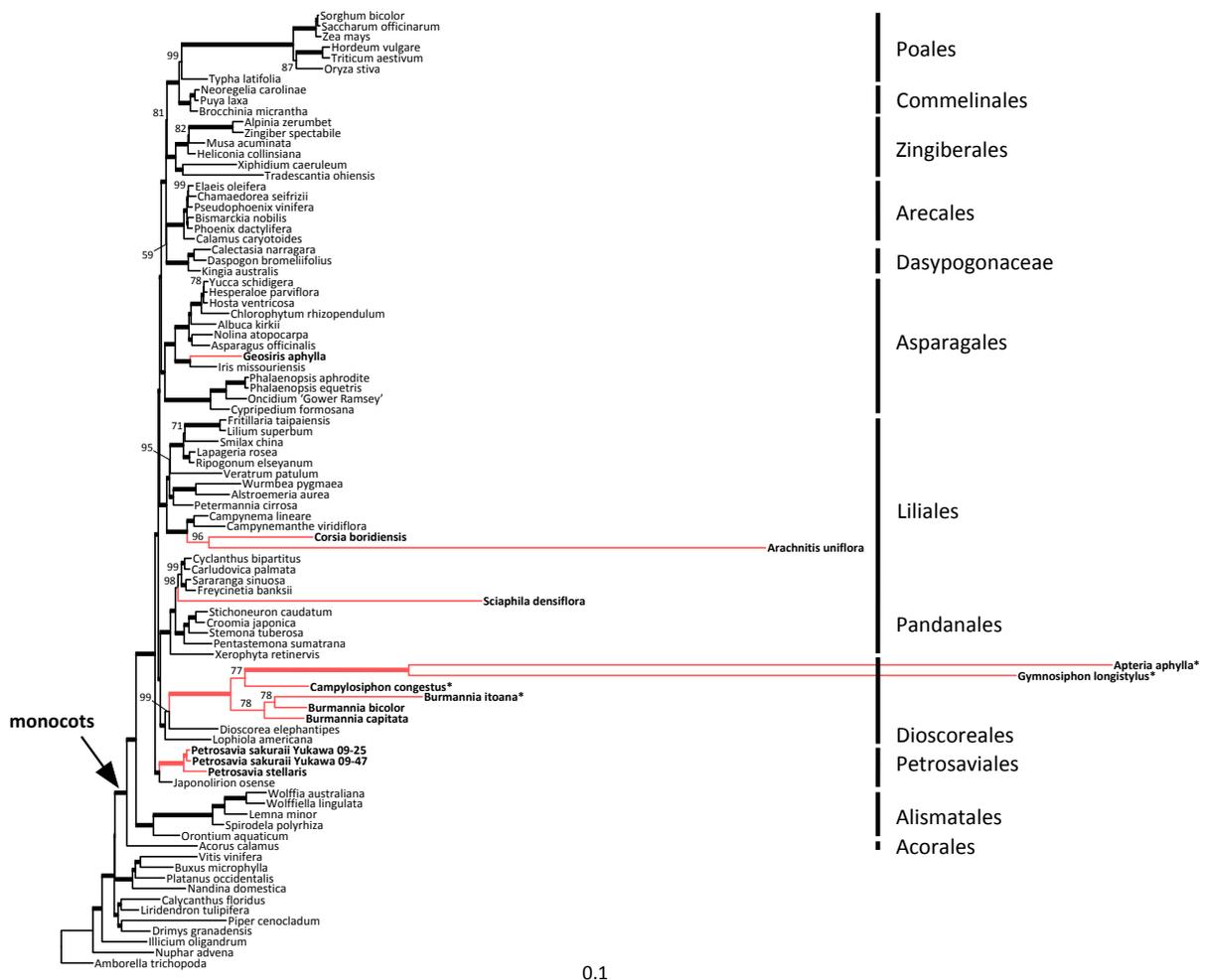


Figure 4.5 Phylogeny of photosynthetic and mycoheterotrophic monocots based on a partitioned gene-by-codon (G x C; see text for details) maximum-likelihood analysis of 83 plastid genes. Thick lines indicate branches with 100% bootstrap support, those with less than 100% bootstrap support are indicated numerically, with values <50% indicated with a dash ('-'). Lineages outside orchids with mycoheterotrophic taxa are indicated in red. The scale bar indicates inferred substitutions per site.



Chapter 5: Comparative plastome analysis of *Parasitaxus usta* (Podocarpaceae), a heterotrophic conifer

5.1 Summary

The heterotrophic conifer *Parasitaxus usta* (Podocarpaceae) has a unique nutritional mode that exhibits attributes of both mycoheterotrophy and plant parasitism. Here I sequenced and characterized the plastid genome of three accessions from this species and from two autotrophic members of Podocarpaceae (*Manoao* and *Lepidothamnus*). A plastid-based phylogenomic analysis supports *Parasitaxus* as the sister group of *Manoao* among included taxa, an arrangement consistent with other recent published studies. I also characterized major plastome structure changes and patterns of gene loss and retention in *Parasitaxus* compared to its closest photosynthetic relatives. Its plastid genome has contracted due to loss or pseudogenization of photosynthesis-related genes, and has several structural rearrangements compared to its closest green relatives (the latter also show multiple rearrangements compared to each other). A novel inverted repeat (IR) region has evolved in *Parasitaxus*, marking the reappearance of a structural feature apparently lost in the plastomes of all other conifers. Most photosynthesis-related genes have been lost or pseudogenized in *Parasitaxus*, but genes involved in the CF₁ portion of the plastid ATP synthase have been retained as open reading frames, as have genes encoding the plastid-encoded RNA polymerase (PEP). Comparative plastome analysis of three accessions of *Parasitaxus* reveals fairly substantial within-species variation that may be useful for future population-level studies of this species.

5.2 Introduction

Parasitaxus usta (Podocarpaceae) is an unusual conifer that parasitizes *Falcatifolium taxoides* (Podocarpaceae) via root-like sinkers that penetrate host roots or trunk to make direct xylem-to-xylem connections (Cherrier 1981; Feild and Brodribb 2005). The presences of high stomatal conductance and low water potential also support its status as a hemiparasite (Woltz et al. 1994; Feild and Brodribb 2005). Although *Parasitaxus* has plastids containing chlorophyll, it lacks the ability to photosynthesize, and measures of ¹³C enrichment suggest that it consumes fungal-metabolized carbon instead of obtaining fixed carbon directly from its conifer host (Feild and

Brodribb 2005). The presence of fungal hyphae at contact points in *Falcatifolium* tissue also suggests that *Parasitaxus* may be mycoheterotrophic on endophytic fungi rather than directly holoparasitic on the host conifer (Woltz et al. 1994; Feild and Brodribb 2005). *Parasitaxus* may therefore represent a unique hybrid mode of heterotrophy unknown in any other land plant lineage (Feild and Brodribb 2005). Successful recovery of plastid markers (e.g., the *trnL-F* intron and intergenic spacer region, Sinclair et al. 2002; and *matK*, Biffin et al. 2011) supports the retention of a plastid genome in *Parasitaxus*, and the accelerated evolutionary rate of these regions (Sinclair et al. 2002) may indicate generally accelerated evolution of its entire plastid genome (plastome), as observed in other heterotrophic lineages (e.g., *Corallorhiza*, Barrett et al. 2014; multiple lineages in Chapters 2-4).

The typical land-plant plastome has a quadripartite structure comprising a large single copy (LSC) and small single copy (SSC), separated by two inverted repeat (IR) copies (Palmer 1985). An IR is thought to be absent in all conifers, perhaps representing convergent loss in Pinaceae and cupressophytes (Wu et al. 2011). Relatively little is known about plastome evolution in Podocarpaceae, the second largest conifer family, as only four full plastome sequences have been obtained to date across its ~19 extant genera (*Nageia nagi*, *Retrophyllum piresii* and two species of *Podocarpus*, *P. lambertii* and *P. totara*). Plastome structure appears to be quite variable in the family based on these, even within a genus (*Podocarpus*, do Nascimento Vieira et al. 2014). Plastome genome reduction due to loss or pseudogenization of photosynthetic genes and other structural rearrangements is common in heterotrophic taxa (e.g., Wolfe et al. 1992; Wickett et al. 2008; Delannoy et al. 2011; Logacheva et al. 2011; Logacheva et al. 2014; Barrett and Davis, 2012; Schelkunov et al. 2015; Bellot and Renner 2016; Gruzdev et al. 2016; Naumann et al. 2016; Samigullin et al. 2016; Chapters 3, 4), and so the plastome of *Parasitaxus* may also be expected to exhibit substantial modification.

Parasitaxus thus provides a unique opportunity to study the effects of parasitism/mycoheterotrophy on a plastid genome of a conifer family that already appears to be prone to modification. Here I characterize newly sequenced plastid genomes of *Parasitaxus usta* from two populations and six autotrophic podocarps, two of the latter newly sequenced here, in order to characterize: (1) The local phylogenetic placement of *Parasitaxus* using full plastome data; (2) Patterns of gene loss and retention with the transition to heterotrophy in this conifer; (3) Major plastome structural changes in *Parasitaxus*, including genome contraction and major

structural rearrangements compared to its closest autotrophic relatives, and; (4) Within-species plastome variation in *Parasitaxus*.

5.3 Materials and methods

5.3.1 Plastome sequencing

I generated new plastome sequences for *Manoao colensoi*, *Lepidothamnus laxifolius* and three accessions of *Parasitaxus usta* (Table D.1). I produced sequencing libraries for *Manoao*, *Lepidothamnus* and *Parasitaxus* using the Bioo Nextflex DNA library prep kit (Bioo Scientific Corp., Austin, USA) following protocols outlined in Chapters 3 and 4. Multiplexed libraries (Cronn et al. 2008) were then sequenced on an Illumina HiSeq 2000 (Illumina, Inc., San Diego, USA) as 100-bp paired-end reads. I made *de novo* assemblies of the reads using CLC Genomics Workbench v. 6.5.1. (CLC bio, Aarhus, DK) with default settings, using *Podocarpus macrophyllus* (unpublished data, kindly provided by Linda Raubeson, Central Washington University) as a reference taxon to first select for plastid contigs using a custom Perl script (Daisie Huang, University of British Columbia; https://github.com/daisieh/phylogenomics/tree/master/filtering/filter_cp.pl). I produced full circular plastomes by connecting plastid contigs with Sanger sequencing products generated using taxon-specific primers (Primer3, Untergasser et al. 2007; Koressaar and Remm 2007) that were designed for amplification and sequencing, assembling the final contigs in Sequencher 4.2.2. (Gene Codes Corp., Ann Arbor, MI, USA). I annotated circular plastid genomes in DOGMA (Wyman et al. 2004), checked exon and intron boundaries in Sequencher, and illustrated them as plastome maps using OGDRAW (Lohse et al. 2013).

5.3.2 Matrix assembly and phylogenetic analysis

I retrieved up to 82 protein-coding genes, 4 rDNA, and 32 tRNA genes for *Manoao*, *Lepidothamnus* and *Parasitaxus* (representing up to 118 genes total; Table 5.1), and added these to corresponding genes from 32 previously published gymnosperm plastomes (Table D.1). I manually aligned protein-coding genes (excluding *ycf1* and *ycf2*, as these two genes were difficult to align) that were retained as open reading frames in each taxa using Se-AL v.2a11 (Rambaut 2002), following alignment criteria outlined in Graham and Olmstead (2000), and

staggered regions that were difficult to align (see Saarela and Graham 2010). The alignment includes 80 protein-coding genes in *Manoao* excluding *ycf1* and *ycf2* (and 31 of 33 protein-coding genes retained in open reading frame in *Parasitaxus*, also excluding its uninterrupted *ycf1* and *ycf2* genes, see Table 5.1). I also excluded the four rDNA genes in phylogenetic inference. I concatenated the aligned genes in a final matrix that is 87,912 bp in length. As a check for errors in matrix assembly, I exported individual genes for each taxon and checked them against their original sequence data in Sequencher (none were found). The *atpA* locus retrieved in *Parasitaxus* comprises an open reading frame that is substantially shorter than in *Manoao* at the 5'-end (see below). I compared this *atpA* sequence with the characterized binding domain sequence inferred for *Ipomoea batatas* (Shao et al. 2011) to confirm that it includes the active domain/binding site of this subunit.

I conducted phylogenetic inference using parsimony and maximum-likelihood (ML) methods. I ran a heuristic parsimony search for the shortest trees in PAUP* v4.0a134 (Swofford 2003) using tree-bisection-reconnection branch swapping (TBR) and 1000 random replicate stepwise addition replicates, holding one tree at each step, and otherwise used default settings. I conducted maximum-likelihood analyses using the graphical interface version of RaxML v.7.4.2 (Stamatakis 2006; Silvestro and Michalak 2012), employing 20 independent searches for the best tree. I analyzed the matrix as an unpartitioned version and as a partitioned version (separated into 240 initial partitions by gene and codon positions, 'G x C'). I used PartitionFinder v.1.1.1 (Lanfear et al. 2012) to find the best substitution models and partitioning schemes using the Bayesian Information Criterion (BIC; Schwarz 1978; Table D.2). For the unpartitioned matrix, PartitionFinder selected the GTR (general time reversible)+G+I model as the best substitution model. For the partitioned matrix, 16 final data partitions were inferred that did not have substantially different substitution models according to the hierarchical clustering method algorithm and the Bayesian information criterion (BIC), all with either the GTR+G or GTR+G+I as the best model. I used the GTR+G model for all partitions and analyses as the I parameter (invariant sites) may be adequately accounted for with the gamma parameter (Yang et al. 2006). I estimated branch support with bootstrap analyses (Felsenstein 1985) using 1000 replicates with one random addition replicates per bootstrap replicate for the parsimony search, and 500 "rapid" bootstrap replicates for the maximum-likelihood analyses. I consider well-supported branches to

have at least 95% bootstrap support, and poorly supported ones with less than 70% support for all bootstrap analyses (see Zgurski et al. 2008).

5.3.3 Identification of colinear regions in conifer plastomes

I used Mauve (Darling et al. 2010) to infer, compare and illustrate colinear segments between *Parasitaxus* and four other podocarps (*Lepidothamnus*, *Manoao* generated here, and published sequences of *Nageia* and *Podocarpus*) and *Ginkgo* (Ginkgoaceae), Table D.1. I also compared *Ginkgo* with *Manoao*, *Parasitaxus* and *Podocarpus* in individual pair-wise comparisons. I removed one copy of the inverted repeat (IR) region in *Ginkgo* and *Parasitaxus* for consistency with other sampled gymnosperms that lack an IR copy. I made these comparisons using the progressiveMauve alignment algorithm, with a seed weight of 19, and using seed families and otherwise used default settings for these analyses.

5.3.4 Identification of sequence variation between *Parasitaxus* populations

I used C-Sebelia (Minkin et al. 2013) to detect variation (i.e., indels and base substitutions) between the sequences of *Parasitaxus usta* 2170 and *Parasitaxus usta* 2107 (*Parasitaxus* 2170 and 2171 are from the same population and are essentially identical, Table D.4). The positions of indels and base substitution differences were loaded into the Blast Ring Image Generator (BRIG; Alikhan et al. 2011) to allow visualization of their relative positions in the plastid genome.

5.4 Results

5.4.1 Novel plastomes of autotrophic taxa

I assembled new plastomes for *Manoao colensoi* (135,808 bp, Fig 5.1), and *Lepidothamnus franklinii* (130,500 bp, Fig. 5.2). The size of these plastomes is typical of other podocarps, and the percentages of coding sequence (59.6%, 62.4%) and GC content (36.5%, 37.3%) in *Manoao* and *Lepidothamnus*, respectively, are also similar to those of other published gymnosperms (Table D.3). *Manoao* and *Lepidothamnus* are identical to each other in terms of their overall structure, with the exception of several tRNA duplications. *Manoao* has pseudogenized *trnD*-GUC and *trnY*-GUA copies adjacent to *ycf1*, a full copy of *trnD*-GUC adjacent to *rbcL*, and a

full copy of *rrn5* adjacent to *clpP* (Fig. 5.1). *Lepidothamnus* has an additional pseudogenized *trnD*-GUC copy adjacent to *ycf1* (Fig. 5.2).

5.4.2 Characterization of the *Parasitaxus* plastome

I assembled the plastomes of *Parasitaxus* from two different populations (provided by Michelle Hollingsworth, Royal Botanic Garden, Edinburgh; samples 2107, 2170/2171; the latter two are two samples from the same site), summarized in Fig. 5.3 (for the sample from population 2107), Fig. D.2 and Table D.4 (the latter two are summaries of variation between samples). Small in-frame indels are present in several genes (*accD*, *clpP* and *rps7*; data not shown) and most variation is in noncoding regions (Fig. D.2). Gene order and content are identical between populations. *Parasitaxus* has substantial genome contraction. At 83,156 bp, it is ~61% of the genome size of *Manoao*, presumably mostly reflecting loss of photosynthesis-related genes. The percentage of coding sequence (60.8%) and GC content (37.6%) of *Parasitaxus* is comparable to its autotrophic relatives (Table D.3).

The novel inverted repeat (IR) region in *Parasitaxus* comprises *matK*, *rpl2* and *rpl23* as duplicated genes maintained in full open reading frame, three tRNAs (*trnI*-CAU, *trnK*-UUU, and *trnQ*-UUG), a partial sequence of *rps19* (this forms an intact *rps19* in copy IRa) and three pseudogenized photosynthesis-related genes (*psbA*, *chlL* and *chlN*). Some of these genes are found in canonical land-plant inverted repeat regions (i.e., *rpl2*, *rpl23*, and *rps19*), and others in the LSC or SSC regions (i.e., *chlL*, *chlN*, *matK*, *psbA*, *trnK*, and *trnQ*) (Goulding et al. 1998; Zhu et al. 2015). Each repeat is 8,661 bp long, and separates a large single copy (LSC) region that is 38,836 bp long, and a small single copy (SSC) that is 28,978 bp in length. The repeats comprise genes typically found in the IR and SSC of land plant plastomes (Fig. 5.3). The gene content of the LSC is fairly typical of other land-plant plastomes.

Aside from *rps16*, instances of gene loss and pseudogenization in *Parasitaxus* are restricted to loci related to photosynthesis. There are eight pseudogenized genes (*psaB*, *psbC*, *psbA*, *psbB*, *petA*, *ndhA*, *ndhF*, and *cemA*) and 42 genes lost in total in *Parasitaxus* (Table 5.1). However, three of six photosynthesis-related genes (*atpA*, *atpB* and *atpE*) are retained as open reading frames. These encode protein subunits that comprise the CF₁ portion of the ATP synthase. The other lost plastid-encoded ATP synthase genes (i.e., *atpF*, *atpH* and *atpI*) comprise the CF₀ portion of the plastid ATP synthase. The *atpA* gene sequence retrieved from *Parasitaxus*

(1341 bp) is shorter by 228 bp at its 5'-end and 42 bases longer at the 3'-end compared to the *atpA* locus retrieved from *Manoao* (1527 bp). Despite the shorter length, the *Parasitaxus* sequence encompasses the entire ~550 bp binding domain (amino acid residues 149-365 in the α subunit of the CF₁ subunit of *Ipomoea batatas*) inferred by Shao et al. (2011), corresponding to nucleotides 307-1277 in *Parasitaxus atpA*. Additionally, all four plastid-encoded RNA polymerase (PEP) subunits, which are involved in transcription of photosynthetic and other plastid genes in autotrophic plants, are retained as open reading frames (Table 5.1; Fig. 5.3). Non-photosynthesis-related genes, including those that code for ribosomal proteins, rDNAs, *accD*, *clpP*, *infA*, *matK*, *ycf1* and *ycf2* are also retained; the protein-coding genes all have open reading frames. All tRNAs are present except for *trnR-CCG*.

5.4.3 Mauve-based assessments of colinearity

Mauve-based linear alignments of locally colinear blocks (LCBs) are shown in Fig. 5.4. Rearrangements can be identified as disruptions of colinear blocks; the fewer the colinear blocks, the fewer arrangements are present. In general, plastomes in podocarps appear to be slightly more compact (~130 kb) than *Ginkgo* (~142 kb), largely due to the absence of an inverted repeat in most of them (Table D.3). Two pairs of plastomes are completely colinear: *Nageia* and *Podocarpus*, and *Manoao* and *Lepidothamnus*. However, *Parasitaxus* appears to be only moderately more rearranged compared to other podocarps: pairwise analyses of *Ginkgo* with *Manoao*, *Parasitaxus* and *Podocarpus*, respectively (Fig. D.1), suggest that *Podocarpus* has the fewest rearrangements relative to *Ginkgo*, as they share seven colinear blocks (Fig. D.1a). *Manoao* shares eight colinear blocks with *Ginkgo* (Fig. D.1b), while *Parasitaxus* may be the most highly rearranged plastome relative to *Ginkgo* with 10 shared colinear blocks (Fig. D.1c).

5.4.4 Population-level plastome variation in *Parasitaxus*

I used C-Sibelia to identify 41 different indels and 33 nucleotide substitutions between *Parasitaxus* populations 2107 and 2170 (Table D.4). Indels varied from a single base (e.g., position 7247) pair to an insertion 60 bp in length (position 74172, located in the *rps7* gene), and were located in the large single copy (LSC) and small single copy (SSC) regions. No sequence variation was found between the two populations across the entire ~8.7-kb-long IR region (Fig. D.2).

5.4.5 Phylogenetic placement of *Parasitaxus* within Podocarpaceae

I inferred *Parasitaxus usta* (i.e., all three accessions) to be the sister group of *Manoao* in the parsimony and maximum-likelihood analyses, with strong bootstrap support for this arrangement (93-95%; Figs 5.5, D.3, D.4). The shortest trees inferred by these analyses are identical, and most branches in all three analyses are well-supported (92-100% bootstrap support).

5.5 Discussion

5.5.1 Phylogenetic placement of *Parasitaxus*

Parasitaxus is one of 19 extant genera in Podocarpaceae, a predominantly southern hemisphere family consisting mostly of understorey and canopy-occupying trees and some sprawling shrubs. Several genera (*Podocarpus*, *Dacrycarpus*, and *Dacrydium*) form major components of Australasian and Malesian forests (Farjon 2001; Salmon 1980). *Parasitaxus usta* is endemic to New Caledonia and is restricted to isolated populations (e.g., Jaffré 1995; Cherrier et al. 1992); its host podocarp *Falcatifolium taxoides* is also endemic to New Caledonia but is more widespread (Sinclair et al. 2002; Merckx et al. 2013). Although podocarps are ecologically and economically important conifers, they have been understudied historically (Farjon 2001; Hill and Brodribb 1999), and we have only recently gained a greater understanding of podocarp diversification using molecular dating analyses (e.g., Biffin et al. 2011; Leslie et al. 2012; Lu et al. 2014).

Early studies on the local placement of *Parasitaxus* in Podocarpaceae phylogeny were somewhat inconclusive: *Parasitaxus* was inferred to be the sister group of other scale-leaved taxa such as *Microstrobos* (= *Pherosphaera*), *Microcachrys* and *Halocarpus* based on morphological data (Kelch 1997). A later study using nuclear 18S rDNA and morphological data placed it as a close relative of *Lagarostrobos* and *Prumnopitys* (Kelch 1998). More recent few-loci datasets provided more consistent placements of *Parasitaxus* as the sister group of *Manoao* and/or *Lagarostrobos*, although with variable branch support (e.g., Sinclair et al. 2002; Biffin et al. 2011; Leslie et al. 2012). The study by Biffin et al. (2011) is the only one to recover this

relationship (i.e., *Parasitaxus* as the sister group of *Lagarostrobos-Manoao*) with strong support; although they inferred this using Bayesian inference, which is known to be prone to reporting inflated support values (e.g., Simmons et al. 2003). A study based on two nuclear loci (*LFY* and *NLY*) provided an alternative placement of *Parasitaxus*, as the sister group of *Phyllocladus*, but this was only weakly supported (Lu et al. 2014). Thus, my study provides the first robustly supported placement of *Parasitaxus* as the sister group of *Manoao* (*Lagarostrobos* was not sampled here) using whole plastome data or non-Bayesian methods. This placement corroborates one inferred by Sinclair et al. (2002), Biffin et al. (2011) and Leslie et al. (2012). The host species (*Falcatifolium taxoides*) is consistently inferred to belong to a distantly related clade of podocarps in all studies (e.g., Sinclair et al. 2002).

There are several possible scenarios to account for the current disjunct biogeographical distribution of *Manoao*, *Lagarostrobos* and *Parasitaxus* (these are endemic to New Zealand, Tasmania and New Caledonia, respectively; Molloy 1995; Sinclair et al. 2002). The break-up of Gondwana and climate change could have resulted in the current patchy distribution patterns observed for other Podocarpaceae, and here for *Parasitaxus* and its relatives (e.g., Hill and Brodribb 1999; Hill 2004). Modern distribution patterns and the fossil record of Podocarpaceae (Farjon 2001; McLoughin 2001) generally support this vicariance scenario for the family. If *Parasitaxus* and its host were formerly more widespread, they could then have become limited to their current distributions following extinction events outside New Caledonia (Sinclair et al. 2002). There is no direct macrofossil evidence for *Parasitaxus*, but its stem lineage likely arose during the Cretaceous, based on several molecular dating analyses (~100 Ma, Biffin et al. 2011; Leslie et al. 2012; Lu et al. 2014). Ancient long-distance dispersal (LDD) events provide an unlikely explanation of its distribution after the lineage switched to full heterotrophy, given its complete reliance on a single host species. However, LDD may have been possible if the host was formerly more widely distributed outside its current location (as may be the case in bird-mediated LDD of the mycoheterotrophic orchid *Cyrtosia septentrionalis*; Suetsugu et al. 2015), or if dispersal of an autotrophic ancestor of *Parasitaxus* resulted in its current distribution. Indeed, a postulated complete geological turnover of New Caledonia suggests that the island's biodiversity arose from multiple LDD events (e.g., Grandcolas et al. 2008; Espeland and Murienne 2011; Kranitz et al. 2014) after the purported re-emergence of New Caledonia around 37 Ma (e.g., Grandcolas et al. 2008). This may provide an upper limit for the timing of the loss

of photosynthesis in *Parasitaxus* if it arrived on New Caledonia as an autotrophic lineage along with its eventual host, and only later became parasitic.

Visual inspection of the phylograms inferred here (e.g., Fig. 5.5) suggest that *Parasitaxus* has a moderately elevated substitution rate that is not excessively fast compared to other heterotrophic plant lineages (e.g., Lemaire et al. 2011; Bronham et al. 2013; Mennes et al. 2013, 2015; Schelkunov et al. 2015; Naumann et al. 2016; Samigullin et al. 2016; Chapters 2, 3, and 4). There were no differences between parsimony and likelihood inferences here in terms of tree topology. As parsimony is known to be more prone to long-branch attraction (LBA; Felsenstein 1978; Hendy and Penny 1989) than model-based analyses (e.g., Huelsenbeck 1998; Swofford et al. 2001), this may suggest that LBA is not an important factor in phylogenomic/phylogenetic inference here, in contrast to other studies of heterotrophic taxa (e.g., Neyland and Hennigan 2003; Chapters 2-4). *Parasitaxus* is represented by only a fraction of the genes recovered in autotrophs here (Table 5.1), but this also did not appear to result in reduced branch support for its local placement. This finding is consistent with inferences for other mycoheterotrophic taxa presented in my thesis, also based on reduced plastid gene sets (Chapters 3, 4).

5.5.2 Evolution of an inverted repeat and plastome evolution in *Parasitaxus*

Inverted repeat (IR) copies in the plastome may serve as anchors during recombination-dependent replication events (Maréchal and Brisson 2010). Land-plant plastomes that have lost an inverted repeat (e.g., conifers, Wu et al. 2011; Wu and Chaw 2014; heterotrophic taxa, Chapters 3 and 4) are often reported to have rearrangements, assumed to be a consequence of destabilization in the plastome resulting from IR loss (e.g., Palmer and Thompson 1982; Strauss et al. 1988, but see Chapter 4). Heterotrophy may lead to structural rearrangements (discounting changes that reflect genome contraction due to gene loss) that are independent of the loss of the IR (e.g., Logacheva et al. 2014; Chapter 4), although many heterotrophic lineages have only limited or no major examples of these kinds of genome structural change (Chapter 4). As in other conifers, multiple plastome rearrangements have been recorded in autotrophic Podocarpaceae (e.g. Wu et al. 2011; Wu and Chaw 2014; do Nascimento Vieira et al. 2014, 2016; this chapter). It is hard to say which factors (including the relatively recent regain of an IR vs. the loss of photosynthesis), if any, are related to the rate of genome structural changes in *Parasitaxus*.

However, this species does not appear to have especially numerous major rearrangements compared to other members of Podocarpaceae (Figs. 5.4, D.1).

The novel IR in *Parasitaxus* likely arose on its stem branch, as all other conifers appear to lack one (the absence in all other conifers may reflect convergent losses in Pinaceae vs. the cupressophytes; Wu et al. 2011; if so, these losses likely reach back to crown clades that are at least 170 Ma and 260 Ma, respectively; Leslie et al. 2012). The lack of indels or point mutations in the inverted repeats in two different accessions of *Parasitaxus* (Fig. D.2) suggests that there are lower substitution rates in the IR compared to the LSC and SSC regions, which is consistent with what is observed in other land-plant IRs, and may be related to concerted evolution due to recombination between the repeats (e.g., Maréchal and Brisson 2010).

Duplication of tRNAs can generate repetitive sequences, that in turn may act as mutational hotspots for plastome rearrangements (e.g., Guisinger et al. 2011; Hipkins et al. 1995; Hirao et al. 2008). The duplication of *trnD*-GUC and *trnY*-GUA in *Manoao*, and *trnD*-GUC in *Lepidothamnus* resulted in new pseudogenized copies of these genes near *ycf1* in their respective plastomes in same orientation as the original genes. The common ancestor of *Manoao* and *Parasitaxus* may have possessed tandem copies of one or both genes. These or other duplicated genes may have become inverted and dispersed, acting as recombination sites in *Parasitaxus*, leading to the formation of the novel IR. The position of the pseudogenized *trnD*-GUC gene in *Manoao* and *Lepidothamnus* near the *ycf1* gene may support this hypothesis; the pseudogenized *trnD*-GUC gene (now lacking in *Parasitaxus*) could have acted as a recombination site resulting in the position of the *ycf1* gene adjacent to an IR/SSC boundary in *Parasitaxus* (Figs. 5.1, 5.3). *Juniperus* has two inverted copies of *trnQ*-UUG (as do other sequenced taxa from Cupressaceae, Cephalotaxaceae and Taxaceae) that may effectively act as small inverted repeats (Guo et al. 2014). Li et al. (2016) suggests that these may have arisen from tandem repeats formed from replication slippage; the repeats then likely became dispersed and experienced subsequent inversion. The novel inverted repeats in *Parasitaxus* are more substantial in length (8,661 bp) than these putative conifer “micro-IRs” (~250 bp). The inclusion of *trnQ*-UUG in the IRs of *Parasitaxus* suggests that these single-gene IRs could also have acted as nuclei for IR formation, which later grew and experienced additional shifts in IR boundaries (e.g., Chapter 4). The *rrn5* copies in *Manoao* are inverted copies of each other, and may also effectively act as tiny inverted repeats for the plastome of this conifer (see also Schelkunov et al. 2015, for another example of

tiny inverted repeats comprising a single tRNA gene in the orchid mycoheterotroph *Epipogium roseum*).

The identification of potential repetitive elements may help clarify the mechanisms behind the plastome rearrangements observed in *Parasitaxus*. An increased number of repetitive sequences (which I did not quantify here) or generally higher rates of substitution may explain the presence of extensive rearrangements in some autotrophic plastomes (e.g., Cai et al. 2008; Haberle et al. 2008; Blazier et al. 2011; Guisinger et al. 2011; Weng et al. 2014; Martin et al. 2013; Sloan et al. 2014; Zhu et al. 2015). However, these may be insufficient to account for plastome rearrangements in heterotrophic plant lineages: mycoheterotrophic *Petrosavia stellaris* possesses a highly rearranged plastome with novel inverted repeats, but has unremarkable substitution rates and a low proportion of repeated content (Logacheva et al. 2014).

5.5.3 Gene loss and retention in *Parasitaxus*

TrnR-CCG is the only non-photosynthetic gene lost in *Parasitaxus* (Table 5.1), and it may not be essential for plastid translation (Sugiura and Sugita 2004). This is the first recorded loss of this gene in Podocarpaceae, although it has been lost multiple times in some conifers (i.e., Cupressaceae, Cephalotaxaceae and Taxaceae), and retained in others (Wu et al. 2007; Hirao et al. 2008; Yi et al. 2013; do Nascimento Vieira et al. 2014; Yap et al. 2015; Li et al. 2016). Almost all of genes lost or pseudogenized in *Parasitaxus* are involved in photosynthesis (Table 5.1), which may indicate a somewhat recent transition to heterotrophy; the possible functional retention of several ATP synthase genes and all four PEP genes are also consistent with this hypothesis. The retention of *rpo* (plastid-encoded RNA polymerase, or PEP) genes is notable because this complex is typically active in photosynthetic plastids, and photosynthesis-related genes are the only plastid genes exclusively transcribed by PEP (e.g., Hajdukiewicz et al. 1997; Maliga 1998; Berg et al. 2004; Swiatecka-Hagenbruch et al. 2007). The remaining genes can be transcribed solely by NEP (nuclear-encoded RNA polymerase), although at lower efficiencies (Legen et al. 2002; Zhelyazkova et al. 2012). The retention of the PEP complex supports the hypothesis of Barrett and Davis (2012) in which *rpo* loss is separate from loss of photosynthesis genes, and contradicts that of Barrett et al. (2014), in which the two are presumed to be lost simultaneously with other sets of genes (see also Logacheva et al. 2014 and Chapter 4 or a

similar situation in the monocot *Petrosavia*). The retention of all four *rpo* genes in open reading frame in *Parasitaxus* is unlikely to be due to a lag in the accumulation of stop codons or indels (given the relatively advanced degradation of other photosynthesis genes), and is more likely to reflect it being a functional protein complex, an hypothesis that could be tested further with transcriptome data and functional assays.

5.5.4 Possible retention of ATP synthase CF₁ region suggests a novel function

ATP formation in the plastids of photosynthetic plants is catalyzed by ATP synthase, a membrane-bound protein complex that produces ATP using energy derived from a proton gradient generated from the photosynthetic electron transport chain. This complex is encoded by six plastid (*atpA*, B, E, F, H and I) and three nuclear (*atpC*, D, and G) genes, and comprises two portions: a multisubunit CF₀ transmembrane portion used to produce the proton gradient, and a multisubunit stromal CF₁ portion involved in ATP production (e.g., Strotmann et al. 1998; Weber 2006). This reflects the evolutionary history of ATP synthase, which was thought to have evolved from two separate enzymes in bacteria, an H⁺ motor and an ATPase, which later became CF₀ and CF₁ respectively (e.g., Walker and Cozens 1986).

ATP synthase has been retained in heterotrophic plant lineages (e.g., *Aneura mirabilis*, Wickett et al. 2008; *Orobancha crenata*, *O. gracilis* *Phelipanche ramosa*, Wicke et al. 2013; *Petrosavia sakurarii*, Chapter 4) and nonphotosynthetic alga (*Prototheca wickerhamii*, Knauf and Hachtel 2002; *Cryptomonas paramecium*, Donaher et al. 2009). Like these, *Parasitaxus* is non-photosynthetic (Feild and Brodribb 2005), but it has retained plastid-encoded *atpA*, B and E genes (as open reading frames) and nuclear *atpC* and *atpD* genes (based on unpublished transcriptome data, data not shown) which together code for subunits that comprise the entire CF₁ portion of ATP synthase. This may indicate that *Parasitaxus* has an active CF₁ protein, presumably with novel functionality given its lack of a CF₀ domain. The CF₁ complex bears similarity to hexameric DNA helicases with ATPase activity as observed in bacteria (e.g., Hingorani et al. 1997). If anchored to the plastid membrane via attachment to a membrane-bound protein complex, the CF₁ complex might still function to drive ATP hydrolysis to power translocation of proteins across membranes, similar to its hypothesized role in producing an H⁺ gradient to power an arginine translocator (Tat) system for protein translocation across the

thylakoid membrane (Kohzuma et al. 2012; Kamikawa et al. 2015). However, this speculative hypothesis requires examination with additional observation and experimentation.

5.5.5 Within-species plastome variation

The same set of genes is retained across the three accessions of *Parasitaxus* examined here (2170, 2171 and 2107; the first two are from the same population in the Mont Dzumac/Ouinne Valley, the latter from the vicinity of La Tranchée in New Caledonia). Only three genes have accumulated indels (*accD*, *clpP* and *rps7*). The variability in *rps7* corresponds to a series of tandem repeats located near the 5'-end of the gene identified as an “expansion hotspot” across the Podocarpaceae and Araucariaceae by Rai et al. (2008), resulting in gene lengths from 396 bp (i.e., *Podocarpus lambertii*, *P. totara*, *Retrophyllum*) to 840 bp (*Nageia*). There is variation between populations of *Parasitaxus* in this gene: accession 2107 is shorter (468 bp) than accessions 2170 and 2171 (527 bp), which had nearly identical *rps7* sequences. The latter two specimens were collected from the same population and may represent the same genet; in contrast, 2107 was collected approximately 17.5 km away (M. Hollingsworth, pers. comm.). The plastid genome may therefore provide useful population-level markers for demographic studies of *Parasitaxus*, for example to help resolve whether individual populations consist of one or multiple genetic individuals (underground connections between ramets within populations of *Parasitaxus* are difficult to infer visually; S.W. Graham, pers. comm.).

Table 5.1 Summary of genes retained in plastid genomes of *Parasitaxus* relative to *Manoao* and *Lepidothamnus*. *Manoao* and *Lepidothamnus* retained the same genes. All three populations of *Parasitaxus* retained the same genes – see text for details. Dash (‘-’) indicates the absence of all genes for any functional groups as categorized here.

Function	<i>Manoao colensoi</i> / <i>Lepidothamnus laxifolius</i>	<i>Parasitaxus usta</i>
Photosynthesis	<i>psaA, psaB, psaC, psaI, psaJ, psaM</i> <i>psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT</i> <i>atpA, atpB, atpE, atpF, atpH, atpI</i> <i>petA, petB, petD, petG, petL, petN</i> <i>ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK</i> <i>chlB, chlL, chlN, lhbA, rbcL, ycf3, ycf4</i>	<i>psaB(ψ), psaC(ψ)</i> <i>psbA(ψ), psbB(ψ)</i> <i>atpA, atpB, atpE</i> <i>petA(ψ)</i> <i>ndhA(ψ), ndhF(ψ)</i> -
Ribosomal proteins	<i>rpl2, rpl14, rpl16, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36</i> <i>rps2, rps3, rps4, rps7, rps8, rps11, rps12, rps14, rps15, rps18, rps19</i>	<i>rpl2, rpl14, rpl16, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36</i> <i>rps2, rps3, rps4, rps7, rps8, rps11, rps12, rps14, rps15, rps18, rps19</i>
RNA polymerase	<i>rpoA, rpoB, rpoC1, rpoC2</i>	<i>rpoA, rpoB, rpoC1, rpoC2</i>
Ribosomal DNAs	<i>rrn4.5, rrn5, rrn16, rrn23</i>	<i>rrn4.5, rrn5, rrn16, rrn23</i>
Transfer RNAs	<i>trnA-UGC, trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnG-GCC, trnG-UCC, trnH-GUG, trnI-CAU, trnI-GAU, trnK-UUU, trnL-CAA, trnL-UAA, trnL-UAG, trnM-CAU, trnM-CAU, trnN-GUU, trnP-UGG, trnP-GGG, trnQ-UUG, trnR-ACG, trnR-CCG, trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU, trnV-GAC, trnV-UAC, trnW-CCA, trnY-GUA</i>	<i>trnA-UGC, trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnG-GCC, trnG-UCC, trnH-GUG, trnI-CAU, trnI-GAU, trnK-UUU, trnL-CAA, trnL-UAA, trnL-UAG, trnM-CAU, trnM-CAU, trnN-GUU, trnP-UGG, trnP-GGG, trnQ-UUG, trnR-ACG, trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU, trnV-GAC, trnV-UAC, trnW-CCA, trnY-GUA</i>
Other protein coding genes	<i>accD, ccsA, cemA, clpP, infA, matK, ycf1, ycf2</i>	<i>accD, cemA(ψ), clpP, infA, matK, ycf1, ycf2</i>

Figure 5.1 Circular plastome map of *Manoao colensoi* (Podocarpaceae). Genes located inside the circle are transcribed clockwise, those outside counterclockwise. The gray circle marks the GC content: the inner circle marks a 50% threshold. Genes with introns are indicated with asterisks (*). Inferred pseudogenes of *trnD*-GUC and *trnY*-GUA are indicated with a ‘ ψ ’ symbol. Red boxes indicate duplicated genes.

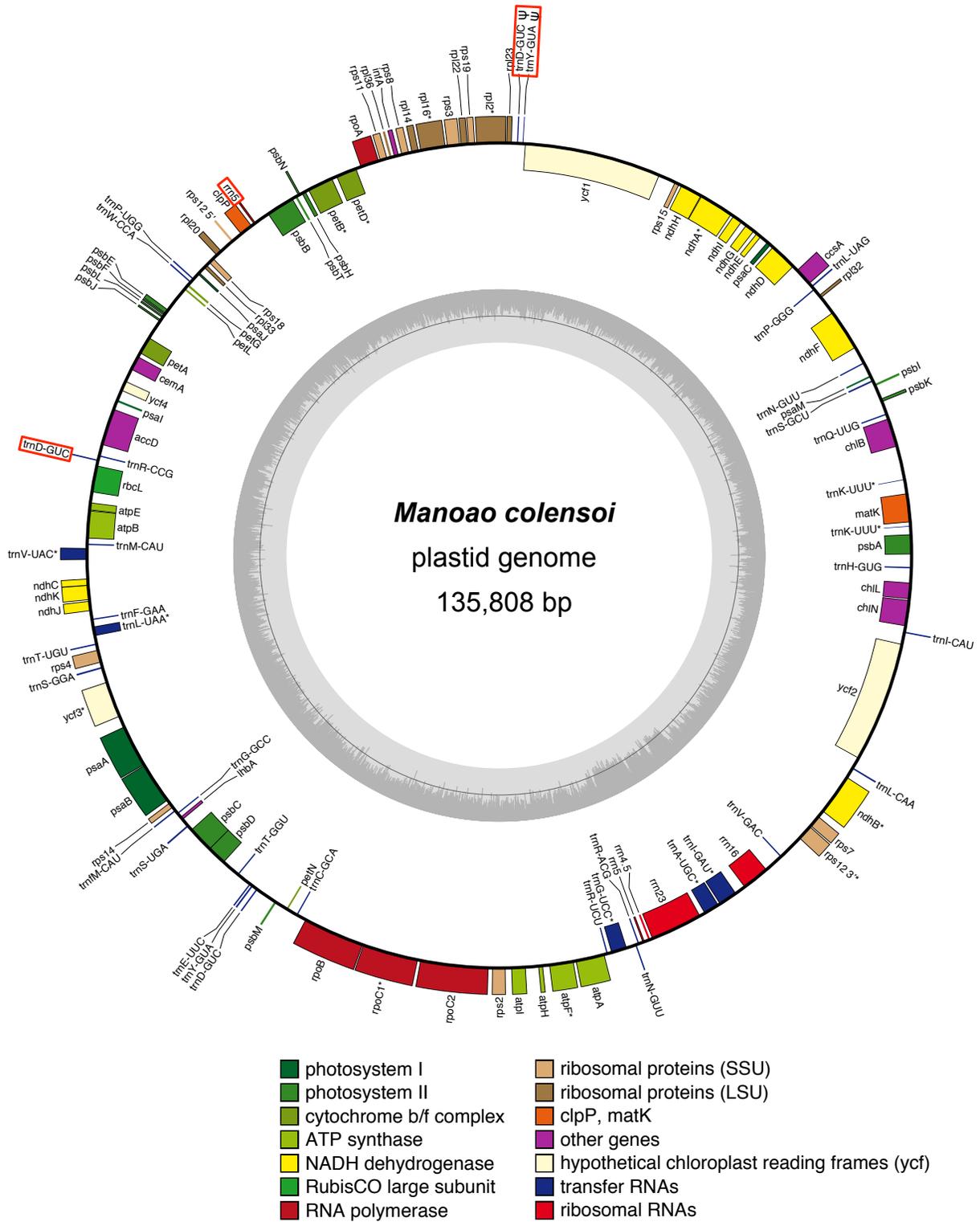
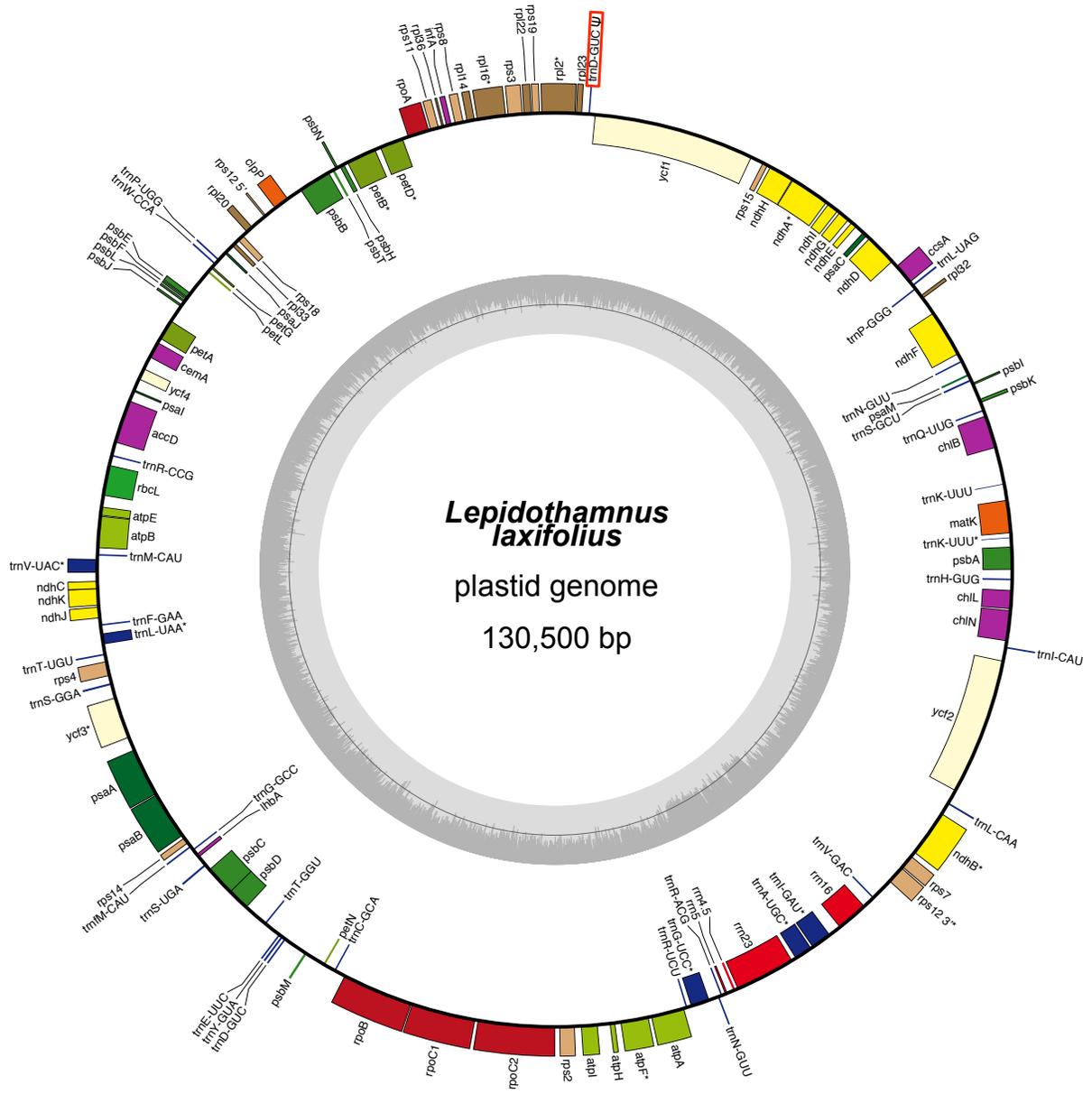


Figure 5.2 Circular plastome map of *Lepidothamnus laxifolius* (Podocarpaceae). Genes located inside the circle are transcribed clockwise, those outside counterclockwise. The gray circle marks the GC content: the inner circle marks a 50% threshold. Genes with introns are indicated with asterisks (*). An inferred *trnY*-GUA pseudogene is indicated with a ‘ ψ ’ symbol. Red boxes indicate duplicated genes.



- | | |
|---|--|
| ■ photosystem I | ■ ribosomal proteins (SSU) |
| ■ photosystem II | ■ ribosomal proteins (LSU) |
| ■ cytochrome b/f complex | ■ clpP, matK |
| ■ ATP synthase | ■ other genes |
| ■ NADH dehydrogenase | ■ hypothetical chloroplast reading frames (ycf) |
| ■ RubisCO large subunit | ■ transfer RNAs |
| ■ RNA polymerase | ■ ribosomal RNAs |

Figure 5.3 Circular plastome map of *Parasitaxus usta* (Podocarpaceae), accession 2170. Genes located inside the circle are transcribed clockwise, those outside counterclockwise. The gray circle marks the GC content: the inner circle marks a 50% threshold. Thick branches indicate IR copies. Genes with introns are indicated with asterisks (*). Inferred pseudogenes are indicated with an ‘ ψ ’.

Figure 5.4 Mauve-based alignments of gymnosperm autotrophs and *Parasitaxus* (a linear map of *Ginkgo biloba* appears first for reference). A single copy of the inverted repeat region was included for *Ginkgo* and *Parasitaxus* for comparison to the other taxa. Coloured blocks indicate shared gene order between two or more genomes, referred to as ‘locally colinear blocks’ (LCBs). LCBs appearing above the line are colinear and in the same orientation as the reference sequence; those below are reverse complements. Coloured lines link LCBs shared between taxa.

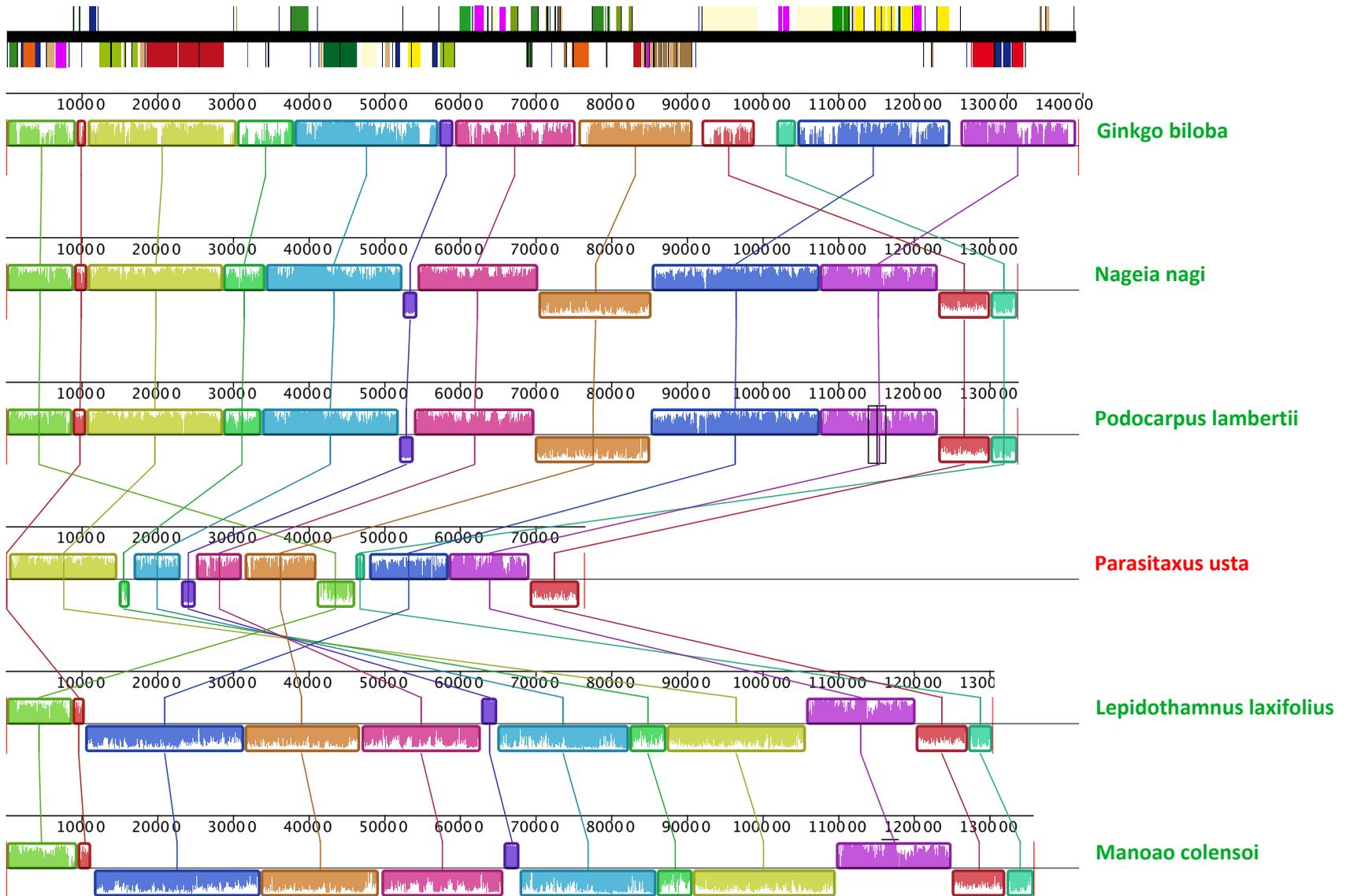
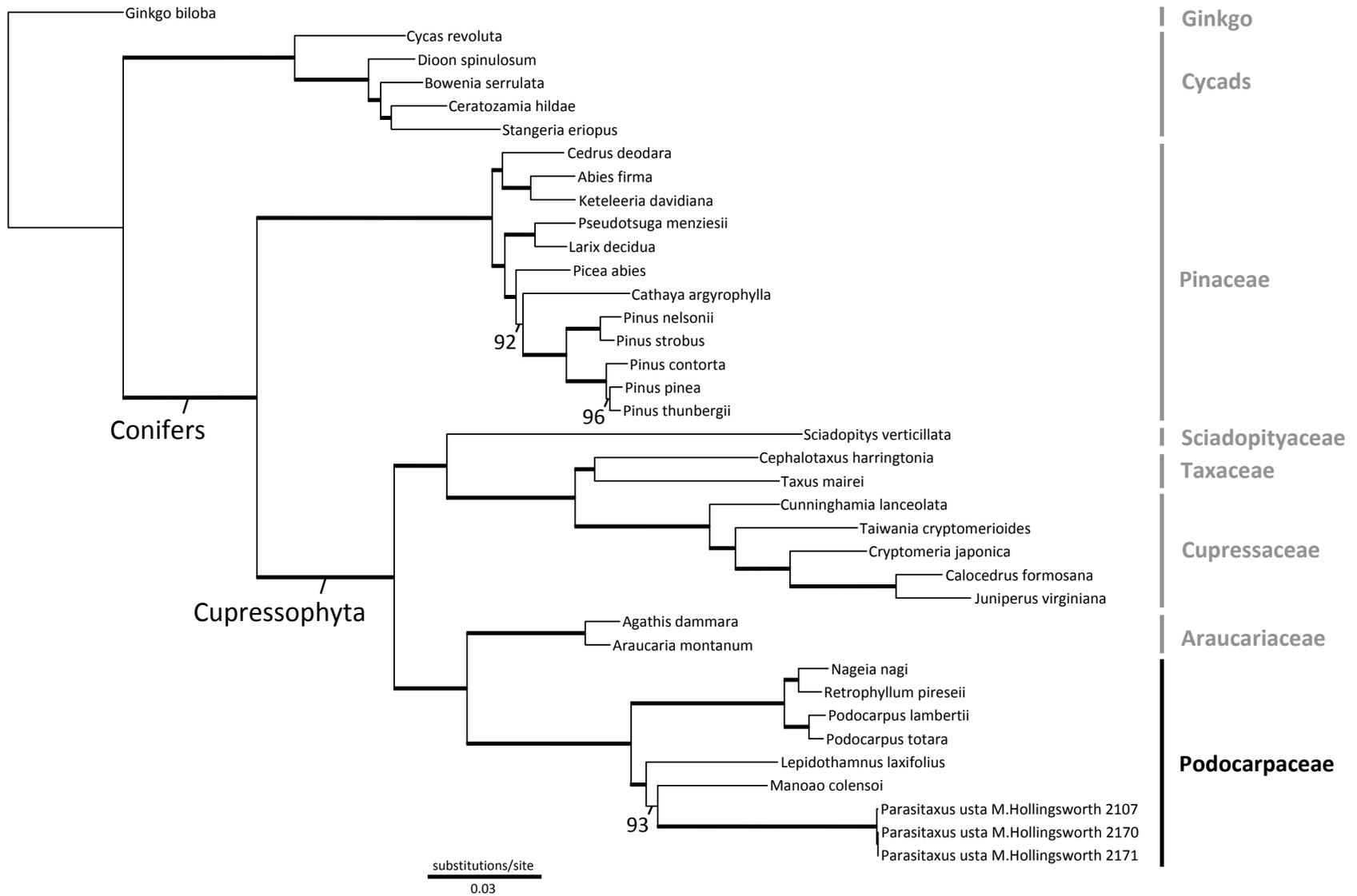


Figure 5.5 Gymnosperm phylogeny inferred in a maximum-likelihood analysis of 80 plastid genes using the 'G x C' partitioning scheme (best tree represented as a phylogram; see text and Table D.2 for details). Bootstrap support values are indicated beside branches: thick lines indicate branches with 100% bootstrap support. The scale bar indicates estimated substitutions per site.



Chapter 6: Conclusion

The mycoheterotrophic plastid genomes I recovered represent independent losses of photosynthesis in Burmanniaceae (with at least two losses among taxa considered here), Corsiaceae, Iridaceae, Petrosaviaceae, Podocarpaceae and Triuridaceae (Chapters 3-5). They therefore act as replicate data points for identifying and characterising general evolutionary processes acting on plastid genomes following loss of photosynthesis. For example, I used these evolutionarily independent data points to demonstrate a strong relationship between genome size and the degree of plastid gene loss in full mycoheterotrophs (Chapter 4). These genomes provide powerful new tools for phylogenetic inference (Chapters 3-5), building on what is possible to infer from few-gene studies (Chapter 2). They also allow us to probe molecular evolutionary processes in plastid genomes, including characterization of stasis and change in selection regimes acting on genes (Chapter 3), and common or lineage-specific patterns of gene loss/retention and genome structural change (Chapters 3-5). Consideration of multiple taxa following individual origins of mycoheterotrophy (in Burmanniaceae, Corsiaceae, Petrosaviaceae) also allows us to make comparisons of divergent evolution in fully mycoheterotrophic lineages after a given (homologous) loss of photosynthesis (Chapter 4).

6.1 Phylogenetics/phylogenomics of mycoheterotrophic taxa

My thesis includes plastid-based investigations into the placement of fully mycoheterotrophic taxa using a few genes that were recovered primarily using the polymerase chain reaction technique (Chapter 2), to those based on complete or nearly complete plastid gene sets recovered using next-generation sequencing methods for individual (Chapters 3, 5) or multiple mycoheterotrophic taxa (Chapter 4). These inferences provide strong plastid-based connections between mycoheterotrophs and the broadly sampled inferences of angiosperm/conifer phylogeny that underpin modern plant classification schemes, which otherwise rely heavily on a few plastid genes found in photosynthetic taxa (e.g., APG 1998-2016).

Perhaps surprisingly, phylogenetic inferences based on less than a handful of retained plastid genes from a broadly sampled taxon set (Chapter 2) provide substantial insights into the phylogenetic placement of these taxa, despite some contamination issues, the use of a relatively small number of genes that are sometimes patchily recovered, and the often highly elevated rates of evolution in individual lineages. These inferences are likely subject to strong long-branch effects, as demonstrated by highly implausible groupings of rapidly evolving heterotrophic lineages using parsimony, in strong contrast to likelihood results that are congruent with studies based on other genomes, and which have improved bootstrap support when additional long branches (other major mycoheterotrophic lineages) are removed by considering each lineage separately. Several of the placements I inferred were previously unclear, including the local position of Corsiaceae (in Liliales, sister to Campynemataceae), Triuridaceae (unstable, but possibly as the sister group of Cyclanthaceae-Pandanaceae in Pandanales) and Thismiaceae (unstable, but closely related to Taccaceae and Trichopodaceae, as in recent studies based on mitochondrial and nuclear data; e.g., Merckx et al. 2010). The recovery of a plastid gene (*accD*) from Thismiaceae points to possible retention of the plastid genome in this family. Other taxa in this study placed in positions that are widely recognized (i.e., *Geosiris*, *Petrosavia* and orchids; Burmanniaceae placed in Dioscoreales as expected, but with an unstable local position in the order; see Chapter 2).

The problem I pointed out with contamination may affect other PCR-based plastid studies of mycoheterotrophs; communications with Mark Chase (Royal Botanic Gardens, Kew) about my results likely contributed to the Angiosperm Phylogeny Group returning Burmanniaceae *sensu lato* (including Thismiaceae) to having a problematic taxonomic status (APG 2016). A possible close relationship between Burmanniaceae and Thismiaceae can not be ruled out by the limited gene sampling used in Chapter 2, but seems unlikely based on individual analyses here, and studies based on other genomes (Merckx et al. 2006, 2009, 2010). The inclusion of a recently produced plastid genome for Thismiaceae (pers. comm. Gwynne Lim, Cornell University) should settle the issue of whether the two families should be recognized separately. My few-gene study (Chapter 2) also provides the first family-wide analysis of Burmanniaceae using plastid data. Although inferred relationships in the family were largely poorly supported, I found

multiple points of phylogenetic congruence with other studies (Merckx et al. 2006, 2008), and was able to reject an hypothesis of a single loss of photosynthesis in the family, consistent with the same studies. In the future my sampling of full plastid genomes in Burmanniaceae (Chapter 4) could be readily augmented to allow inference of a well-sampled and strongly supported set relationships based on plastome evidence. Finally, the primers that I designed in Chapter 2 for genes commonly retained in mycoheterotrophic taxa (i.e., *accD*, *clpP* and *matK*) may be useful for other phylogenetic explorations in other heterotrophic taxa, without requiring full plastome sequencing.

Phylogenetic inferences using full plastid gene sets (focused on the ~80 protein coding genes retained in most seed plants; Chapter 4) are also subject to long-branch attraction for parsimony, as this study again grouped fast heterotrophic lineages in an implausible and (in contrast to Chapter 2) well-supported clade. This fast clade in parsimony trees is implausible because it conflicts strongly with likelihood-based inferences, and also groups together a highly taxonomically heterogeneous set of taxa that are distantly related in studies that consider other sources of evidence. In contrast, phylogenomic inferences made using likelihood analysis are consistent with other studies (e.g., Mennes et al. 2013; 2015) and with the few-gene study in Chapter 2. They are also well-supported here (Chapters 3-5), despite often extensive gene loss in individual taxa (Chapters 3-5). Advances on the phylogenetic inferences made in the few-gene study (Chapter 2) include much stronger support for Triuridaceae as the sister group of Cyclanthaceae-Pandanaceae in Pandanales, a result consistently found in a diverse range of DNA, amino-acid and codon-based likelihood analyses (Chapters 3, 4), and of Burmanniaceae as the sister group of Dioscoreaceae among sampled taxa in Dioscoreales, with moderate support for relationships within this family (Chapter 4). A phylogenomic analysis of Podocarpaceae places *Parasitaxus* with strong support as the sister group of *Manoao colensoi* among sampled taxa in Podocarpaceae (Chapter 5). This result is consistent with most other studies (e.g., Sinclair et al. 2002; Biffin et al. 2011; Leslie et al. 2012), but is the first time it has been recovered with strong support in likelihood analysis.

6.2 Gene loss (and retention) trajectory in mycoheterotrophs

The persistence of plastid genomes in fully heterotrophic land plants (e.g., Wolfe et al. 1992; Funk et al. 2007; Wickett et al. 2008; Delannoy et al. 2011; Barrett and Davis 2012; Logacheva et al. 2014; Schelkunov et al. 2015; Naumann et al. 2016; Samigullin et al. 2016; Gruzdev et al. 2016) supports additional essential roles of the plastid aside from photosynthesis. Core sets of plastid-encoded genes are commonly retained in the mycoheterotrophic lineages characterized here and in other heterotrophs (i.e., *accD*, *clpP*, *matK*, *trnE* and a minimal set of translational apparatus genes; the first three of these were employed in Chapter 2 as phylogenetic markers). This suggests that the continued maintenance of the plastome is (eventually) solely to allow for the expression of a few housekeeping genes outside those involved in the translation apparatus; successful functional replacement by nuclear or mitochondrial counterparts for at least some of these genes may be unlikely (although see Molina et al. 2014 and Janouškovec et al. 2015, respectively, concerning plastome loss in *Rafflesia* and heterotrophic protists; and Bellot and Renner 2016 for a contrasting case in endoparasitic holoparasites).

The lists of retained plastid genes inferred here (Chapters 3-5) across full mycoheterotrophs are generally consistent with a gene-loss trajectory proposed by Barrett and Davis (2012), in which genes are lost in successive stages after the initial transition to heterotrophy. Genes encoding plastid NAD(P)H dehydrogenase subunits may be lost before the full loss of photosynthesis (e.g., lost in autotrophic *Burmannia capitata*, Burmanniaceae; Chapter 4); subsequent full loss of photosynthesis then likely leads to a rapid loss of most photosynthesis genes, although the plastid-encoded RNA polymerase, PEP (which plays a significant role in transcription of photosynthetic genes) is retained at least initially when photosynthesis genes are lost (as in *Petrosavia sakurii* and *Parasitaxus*, Chapters 4, 5). Also consistent with their hypothesis is the functional retention of the plastid ATP synthase complex after the initial loss of photosynthesis (as found in the genus *Petrosavia*, and likely in *Parasitaxus*, Chapters 4, 5). The staggered loss of ATP synthase and PEP contradicts the more streamlined series of losses hypothesized in Barrett et al. (2014). The retention of *rbcL* (plastid-encoded large subunit of Rubisco) and *atp* (ATP synthase) genes in Petrosaviaceae, and *atp* genes in *Parasitaxus*, supports alternative or novel non-photosynthetic functions for the resulting

plastid protein complexes. In the final stages of plastid genome degradation, translational apparatus genes and other non-photosynthesis related genes (e.g., *accD*, *clpP*, *matK*) may be lost over time, consistent with many full mycoheterotrophs surveyed here (Chapters 3, 4). The plastomes retained in the most reduced mycoheterotrophic taxa studied here (e.g., *Apteria*, *Gymnosiphon*, and *Sciaphila*) may all be in this final stage of gene loss.

Surprisingly, *matK* (the group IIA intron maturase) has been lost in multiple sets of taxa that still retain one or more genes with group IIA introns (e.g., 3'-*rps12*, *rpl2*, the second intron of *clpP*), suggesting alternative splicing mechanisms for these genes (previously only suggested for *clpP*, see Zoschke et al. 2010). Only the highly reduced plastome of *Apteria* (Burmanniaceae) has lost all of its group IIA introns and *matK*.

6.3 Patterns of plastome rearrangement in mycoheterotrophs

The major cause of plastid genome structural arrangement here is simple gene loss, leading to genome compaction (Chapters 3-5). Despite often massive gene loss, most mycoheterotrophic plastomes exhibit absolute or near colinearity with those of autotrophic relatives considering the retained genes (i.e., *Apteria*, *Campylosiphon* and *Gymnosiphon* in Burmanniaceae; *Geosiris*, Iridaceae; *Petrosavia sakurarii* and *P. aff. sakurarii*, Petrosaviaceae; *Sciaphila*, Triuridaceae). A few display more substantial rearrangements (i.e., *Burmannia itoana* in Burmanniaceae; *Arachnitis* and *Corsia* in Corsiaceae; *Parasitaxus* in Podocarpaceae; *Petrosavia stellaris*). It is not clear whether the loss of an IR has a general effect on the stability of plastomes (e.g., Palmer and Thompson 1982; Maréchal and Brisson 2010), as several taxa with an IR have substantial structural changes (i.e., *Arachnitis* in Corsiaceae and *Burmannia itoana* in Burmanniaceae in Chapter 4, *Petrosavia stellaris* in Petrosaviaceae, Logacheva et al. 2014) and others lacking an IR here have little structural change (i.e., *Apteria* and *Gymnosiphon* in Burmanniaceae, *Sciaphila* in Triuridaceae and *Petrosavia aff. sakurarii* in Petrosaviaceae; Chapter 4). Different taxa in the same mycoheterotrophic lineages (Burmanniaceae, Corsiaceae, Petrosaviaceae) often show considerable differences in gene content and genome arrangement, supporting the idea that these processes are idiosyncratic both within and among mycoheterotrophic lineages. Aside from genome contraction, most changes in plastome architecture involve the inverted repeat (IR)

region: beyond expansions and contractions of IR boundaries (also observed in autotrophic lineages at a slower rate), several mycoheterotrophic taxa have independently lost an entire IR copy (i.e., *Apteria*, *Gymnosiphon*, *Corsia*, *Sciaphila* and *Petrosavia* aff. *sakurarii*). The gain of an additional tandem IR copy in *Campylosiphon*, and the formation of a novel inverted repeat in *Parasitaxus* is also noteworthy.

6.4 Future directions

Future studies could add substantial sets of new plastid genomes in Burmanniaceae and Orchidaceae, representing additional independent losses, and also examine plastomes of partial mycoheterotrophs (e.g., in Burmanniaceae), although the status of putative partial mycoheterotrophs in both of these families is often unclear, without physiological evidence this is inferred based on retention of chlorophyll and reductions in vegetative morphology or pigmentation (e.g., Merckx et al. 2006). Physiological testing to clarify the trophic states in these taxa (as in Bolin et al. 2015) would help us to better understand the functional basis of the development of mycoheterotrophy from autotrophic ancestors.

Although my current broad survey reveals several major commonalities (and much variation) in patterns of genome evolution related to gene loss and retention and genome structural change, additional data points would be useful, as the wide range of variation I observed here among ten fully mycoheterotrophic taxa makes it seem unlikely that I have captured all possible types of change after the loss of photosynthesis. Additional sampling could also capture additional lineages in the intermediate stages of genome degradation, to determine how general some patterns are (e.g., the putatively temporary retention of PEP, ATP synthase and Rubisco after the loss of other photosynthesis genes; whether these three protein complexes are always lost in a particular order), and at the other extreme the limits to plastome reduction. They would also help characterize how diverse genome degradation is after individual losses, as has been done recently in Orobanchaceae (Wicke et al. 2013), and as started here for Corsiaceae, Petrosaviaceae and Burmanniaceae, respectively (Chapter 4). These data would also be useful for further plastid phylogenomic inference, for example within Burmanniaceae, which was only lightly sampled here (Chapters 2, 4). I was not able to successfully retrieve plastomes from Thismiaceae and other taxa in Triuridaceae (aside

from *Sciaphila*), although there is evidence for plastome retention at least in Thismiaceae. If other plastomes from Thismiaceae *sensu lato* could be obtained, it would be useful to confirm the monophyly and number of losses of photosynthesis in this family (e.g., Merckx et al. 2009) using plastome data. Additional sampling might also reveal whether full mycoheterotrophs have any substantial differences from holoparasites in terms of how their genomes evolve after the loss of photosynthesis.

The retention of *atp* genes comprising the CF₁ (stromal) portion of ATP synthase (plastid-encoded *atpA*, B, and E; nuclear-encoded *atpD* and C) in *Parasitaxus* is particularly unexpected. This is unlikely to be a consequence of a temporal lag in accumulation of deleterious mutations, as other plastid-encoded *atp* genes in *Parasitaxus* are lost or pseudogenized. Transcriptome data (from the 1KP project) has shown that these genes are actively transcribed. Analysis of dN/dS ratios (as in Chapter 3 for *Sciaphila*) would help to understand the selective regime of the CF₁ portion of the complex, although it is unclear if it is acting with a new or previously unrecorded function, independently of the (now absent) CF₀ transmembrane domain. Ultimately, evidence from western blots and *in vitro* protein assays could be used to confirm the role of this CF₁ protein in *Parasitaxus*, and to help elucidate possible non-photosynthetic roles for it in land plants in general. This may be complicated by working with this plant in the field (it can not currently be cultivated). Further genomic surveys of other nuclear-encoded plastid genes in full mycoheterotrophs would also be extremely valuable; however, the plastid genome evidence presented here provides some of the first genome-level insights into the functioning and evolution of these cryptic and bizarre plants.

Bibliography

- Abrahamsen, M. S., T. J. Templeton, S. Enomoto, J. E. Abrahante, G. Zhu, C. A. Lancto, M. Deng, et al. 2004 . Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science* 304: 441–445.
- Alikhan, N. F., N. K. Petty, N. L. B. Zakour, and S. A. Beatson. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12: 402.
- 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, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 40–410 .
- Angiosperm Phylogeny Group. 1998. An ordinal classification for the families of flowering plants. *Annals of the Missouri Botanical Garden* 4: 531–553.
- Angiosperm Phylogeny Group. 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Botanical Journal of the Linnean Society* 141: 399–436.
- Angiosperm Phylogeny Group. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society* 161: 105–121.
- Angiosperm Phylogeny Group. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flower plants: APG IV. *Botanical Journal of the Linnean Society* 181: 1-20.
- Anisimova, A, and Z. Yang. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Molecular Biology and Evolution* 24:1219–1228.
- 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.

- Barkman, T. J., S.-H. Lim, K. M. Salleh, and J. Nais. 2004. Mitochondrial DNA sequences reveal the photosynthetic relatives of *Rafflesia*, the world's largest flower. *Proceedings of the National Academy of Sciences USA* 101: 787-792.
- 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. and J. V. Freudenstein. 2008. Molecular evolution of *rbcL* in the mycoheterotrophic coralroot orchids (*Corallorhiza* Gagnebin, Orchidaceae). *Molecular Phylogenetics and Evolution* 47: 665-679.
- Barrett C. F., J. I. Davis, J. Leebens-Mack, J. G. Conran, and D. W. Stevenson. 2013. Plastid genomes and deep relationships among the commelinid monocot angiosperms. *Cladistics* 29:65–87.
- 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 early-transitional heterotrophic orchids, and implications for heterotrophic angiosperms. *Molecular Biology and Evolution* 31: 3095–3112.
- Bellot, S., and S. S. Renner. 2016. The plastomes of two species in the endoparasite genus *Pilosyles* (Apodanthaceae) each retain just five or six possibly functional genes. *Genome Biology and Evolution* 8: 189–201.
- Bendich, A. J. 2004. Circular chloroplast chromosomes: the grand illusion. *The Plant Cell* 16: 1661-1666.
- Berg, S., Krause, K. and K. Krupinska. 2004. The *rbcL* genes of two *Cuscuta* species, *C. gronovii* and *C. subinclusa*, are transcribed by the nuclear-encoded plastid RNA polymerase (NEP). *Planta* 219: 541-546.
- Bernard, N. 1909. L'évolution dans la symbiose. Les orchidées et leurs champignons commensaux. *Annales des Sciences Naturelles. Botanique* 9: 1–196.
- Bidartondo, M. I., D. Redecker, I. Hijri, A. Wiemken, T. D. Bruns, L. Domínguez, A. Sérsic, J. R. Leake, and D. J. Read. 2002. Epiparasitic plants specialized on arbuscular mycorrhizal fungi. *Nature* 419: 389-392.

- 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. 2005. The evolutionary ecology of myco-heterotrophy. *New Phytologist* 167: 335-352.
- Biffin, E., J. Conran, J. and A. Lowe. 2011. Podocarp evolution: a molecular phylogenetic perspective. In: B. L. Turner, L. A. Cernusak [eds.] *Ecology of the Podocarpaceae in Tropical Forests*. Smithsonian Institution Scholarly Press 95 pp. 1-20.
- Blazier, J.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.
- Bock, R. 2007. Structure, function, and inheritance of plastid genomes. In R. Bock [ed.], *Cell and Molecular Biology of Plastids*. Springer Berlin, Heidelberg, pp 29-63.
- Bodin, S. S., J. S. Kim, and J.-H. Kim. 2016. Phylogenetic inferences and evolution of plastid DNA in Campynemataceae and the mycoheterotrophic *Corsia dispar* D.L. Jones & B. Gray (Corsiaceae). *Plant Molecular Biology Reporter* 34: 192–210
- Bolin, J. F., K. U. Tennakoon, M. B. A. Majid, and D. D. Cameron. In press. Isotopic evidence of partial mycoheterotrophy in *Burmannia coelestis* (Burmanniaceae). *Plant Species Biology* doi: 10.1111/1442-1984.12116
- Braukmann, T.W.A., M. Kuzmina, and S. Stefanovic. 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.
- Bromham, L., P. F. Cowman, and R. Lanfear. 2013. Parasitic plants have increased rates of molecular evolution across all three genomes. *BMC Evolutionary Biology* 13: 1.
- Bungard, R. A. 2004. Photosynthetic evolution in parasitic plants: Insight from the chloroplast genome. *BioEssays* 26: 235–247.
- Caddick, L. R., P. J. Rudall, P. Wilkin, and M. W. Chase. 2000. Yams and their allies: Systematics of Dioscoreales. In K. L. Wilson, D. A. Morrison [eds.], *Monocots: Systematics and evolution*. CSIRO Publishing, Victoria, Australia. pp. 475–487.

- Caddick, L. R., P. J. Rudall, P. Wilkin, T. A. J. Hedderson, and M. W. Chase. 2002 .
Phylogenetics of Dioscoreales based on analyses of morphological and molecular data.
Botanical Journal of the Linnean Society 138: 123–144.
- Cai, Z., M. Guisinger, H. G. Kim, E. Ruck, J. C. Blazier, V. McMurtry, J. V. Kuehl, J.
Boore, and R. K. Jansen. 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, K. M., M. W. Chase, and P. J. Rudall. 2003. Recircumscription of the
monocotyledonous family Petrosaviaceae to include *Japonolirion*. *Brittonia* 55: 214–
225.
- Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les, B. D. Mishler, M. R.
Duvall, et al. 1993. Phylogenetics of seed plants: An analysis of nucleotide sequences
from the plastid *rbcL*. *Annals of the Missouri Botanical Garden* 80: 528–580.
- Chase, M. W., D. E. Soltis, P. S. Soltis, P. J. Rudall, M. F. Fay, W. H. Hahn, S.
Sullivan, et al. 2000. Higher-level systematics of the monocotyledons: an assessment
of current knowledge and a new classification. In K. L. Wilson and D. A. Morrison
[eds.]. *Monocots: Systematics and Evolution*. CSIRO Publishing, Victoria, Australia.
pp. 3–16.
- Chase, M. W., M. F. Fay, D. S. Devey, O. Maurin, N. Rønsted, T. J. Davies, Y. Pillon, G.
Petersen, O. Seberg, M. N. Tamura, et al. 2006 Multigene analyses of monocot
relationships: a summary. *Aliso* 22: 63–75.
- Cherrier, J. F. 1981. Le Parasitaxus ustus (Vieillard) de Laubenfels. *Revue Forestière
Française* 33: 445–448
- Chumley, T.W., J. D. Palmer, J. P. Mower, H. M. Fourcade, P. J. Calie, J. L. Boore, and
R.K. Jansen. 2006. The complete chloroplast genome sequence of *Pelargonium x
hortorum*: organization and evolution of the largest and most highly rearranged
chloroplast genome of land plants. *Molecular Biology and Evolution* 23: 2175-2190.
- 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. 1981. *An Integrated System of Classification of Flowering Plants*. Columbia University Press, New York City, USA.
- Cronquist, A. 1988. *The evolution and classification of flowering plants*, 2nd Ed. New York Botanical Garden, Bronx, New York, USA.
- Cummings, M. P., and N. A. Welschmeyer. 1998. Pigment composition of putatively achlorophyllous angiosperms. *Plant Systematics and Evolution* 210: 105–111.
- Cusimano, N., and S. Wicke. 2015. Massive intracellular gene transfer during plastid genome reduction in nongreen Orobanchaceae. *New Phytologist* 2: 680-693.
- Dahlgren R. M. T., and F. N. Rasmussen. 1983. Monocotyledon evolution: characters and phylogenetic estimation. *Evolutionary Biology* 16:255–395.
- Dahlgren, R.M., H. T. Clifford, and P. F. Yeo. 1985. *The Families of the Monocotyledons: Evolution, and Taxonomy*. Springer Science & Business Media, Berlin. Germany.
- Darling, A. E., B. Mau, and T. N. Perna. 2010. progressiveMauve: multiple genome alignment with gene gain, loss, and rearrangement. *PLoS One* 5: e11147.
- Darriba D., G. L. Taboada R. Doallo, and D. Posada. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* 9: 772.
- Davis, J. I. , D. W. Stevenson, G. Petersen, O. Seberg, L. M. Campbell, J. V. Freudenstein, D. H. Goldman, et al. 2004 . A phylogeny of the monocots, as inferred from *rbcL* and *atpA* sequence variation, and a comparison of methods for calculating jackknife and bootstrap values. *Systematic Botany* 29: 467–510.
- Davis, C. C., and K. J. Wurdack. 2004. Host-to-parasite gene transfer in flowering plants: Phylogenetic evidence from Malpighiales. *Science* 305: 676–678.
- Delannoy, E., S. Fujii, C. C. Des Francs-Small, M. Brundrett, and I. Small. 2011. Rampant gene loss in the underground orchid *Rhizanthella gardneri* highlight 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, *rbcL* and *atpB*, in a non-photosynthetic parasitic plant. *Plant Molecular Biology* 29:1071-1079.

- Donaher, N., G. Tanifuji, N. T. Onodera, S. A. Malfatti, P. S. Chain, Y. Hara, and J. M. Archibald. 2009. The complete plastid genome sequence of the secondarily nonphotosynthetic alga *Cryptomonas paramecium*: reduction, compaction, and accelerated evolutionary rate. *Genome Biology and Evolution* 1: 439-448.
- Douglas, G. W., D. Meidinger, and J. Pojar. 2001. Illustrated flora of British Columbia. Volume 6: Monocotyledons (Acoraceae through Najadaceae). British Columbia Ministry of Environment, Lands and Parks.
- Downie, S.R., D. S. Katz-Downie, K. H. Wolfe, P. J. Calie, and J.D. Palmer. 1994. Structure and evolution of the largest chloroplast gene (ORF2280): internal plasticity and multiple gene loss during angiosperm evolution. *Current Genetics* 25: 367-378.
- Doyle, J. J., and J. L. Doyle. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19: 11–15.
- Drescher, A., S. Ruf, T. Calsa, H. Carrer, and R. Bock. 2000. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. *The Plant Journal* 22: 97-104.
- Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 29: 1969–1973.
- Ems, S. C., C. W. Morden, C. K. Dixon, K. H. Wolfe, C. W. dePamphilis, and J. D. Palmer. 1995. Transcription, splicing and editing of plastid RNAs in the nonphotosynthetic plant *Epifagus virginiana*. *Plant Molecular Biology* 29: 721–733.
- Engels, B. 1993. Amplify3x [online]. Website <http://engels.genetics.wisc.edu/amplify/> [accessed 15 May 2010].
- Espeland, M., and J. Murienne. 2011. Diversity dynamics in New Caledonia: towards the end of the museum model? *BMC Evolutionary Biology* 11: 254.
- Farjon, A. 2001. World Checklist and Bibliography of Conifers. Royal Botanic Gardens, Richmond, UK. p. 250.
- Fay, M. F., P. J. Rudall, S. Sullivan, K. L. Stobart, A. Y. De Bruijn, G. Reeves, F. Qamaruz-Zaman, et al. 2000. Phylogenetic studies of Asparagales based on four plastid DNA regions. In K. L. Wilson and D. A. Morrison [eds.], *Monocots: systematics and evolution*. CSIRO Publishing, Collingwood, Australia. pp 360–371.

- Fazekas, A. J., K. S. Burgess, P. R. Kesanakurti, S. W. Graham, S. G. Newmaster, B. C. Husband, D. M. Percy, et al. 2008. Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS One* 3: e2802.
- Feild, T. S. and T. J. Brodribb. 2005. A unique mode of parasitism in the conifer coral tree *Parasitaxus ustus* (Podocarpaceae). *Plant, Cell & Environment* 28: 1316-1325.
- 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.
- Felsenstein, J. 1988. Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Genetics* 22: 521–565
- Fleischmann, T.T., L. B. Scharff, S. Alkatib, S. Hasdorf, M. A. Schöttler, and R. Bock. 2011. Nonessential plastid-encoded ribosomal proteins in tobacco: a developmental role for plastid translation and implications for reductive genome evolution. *The Plant Cell* 23: 3137-3155
- Franke, T., L. Beenken, M. Döring, A. Kocyan, and R. Agerer. 2006. Arbuscular mycorrhizal fungi of the Glomus-group A lineage (Glomerales; Glomeromycota) detected in myco-heterotrophic plants from tropical Africa. *Mycological Progress* 5: 24-31.
- Freudenstein, J. V., and D. M. Senyo. 2008. Relationships and evolution of *matK* in a group of leafless orchids (*Corallorhiza* and Corallorhizinae; Orchidaceae: Epidendroideae). *American Journal of Botany* 95: 498–505.
- 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 species, *Cuscuta reflexa* and *Cuscuta gronovii*. *BMC Plant Biology*. 7:45
- Furness, C.A., P. J. Rudall, and A. Eastman. 2002. Contribution of pollen and tapetal characters to the systematics of Triuridaceae. *Plant Systematics and Evolution* 235: 209–218.
- Fuse, S., and M. N. Tamura. 2000. A phylogenetic analysis of the plastid *matK* gene with emphasis on Melanthiaceae sensu lato. *Plant Biology* 2: 415–427.

- Gandolfo, M. A., K. C. Nixon, W. L. Crepet, D. W. Stevenson, and E. M. Friis. 1998. Oldest known fossils of monocotyledons. *Nature* 394: 532–533.
- Gandolfo, M. A., K. C. Nixon, and W. L. Crepet. 2002. Triuridaceae fossil flowers from the Upper Cretaceous of New Jersey. *American Journal of Botany* 89: 1940–1957.
- Gantt, J. S., S. L. Baldauf, P. J. Calie, N. F. Weeden, and J. D. Palmer. 1991. Transfer of *rpl22* to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. *The EMBO Journal* 10: 3073
- Gebauer, G. and M. Meyer. 2003. ^{15}N and ^{13}C natural abundance of autotrophic and myco-heterotrophic orchids provides insight into nitrogen and carbon gain from fungal association. *New Phytologist* 160: 209-223.
- Givnish, T. J., J. C. Pires, S. W. Graham, M. A. McPherson, L. M. Prince, T. B. Patterson, H. S. Rai, et al. 2005. Repeated evolution of net venation and fleshy fruits among monocots in shaded habitats confirms *a priori* predictions: evidence from an *ndhF* phylogeny. *Proceedings of the Royal Society of London, B, Biological Sciences* 272: 1481–1490.
- Givnish, T. J., M. Ame, J. R. McNeal, M. R. McKain, P. R. Steele, C. W. dePamphilis, S. W. Graham, et al. 2010. Assembling the tree of the monocotyledons: plastome sequence phylogeny and evolution of Poales. *Annals of the Missouri Botanical Garden* 97: 584–616.
- Givnish, T. J., A. Zuluaga, I. Marques, V. K. Y. Lam, M. Soto Gomez, W. J. D. Iles, M. Ames, et al. 2016. Phylogenomics and historical biogeography of the monocot order Liliales: Out of Australia and through Antarctica. *Cladistics* doi: 10.1111/cla.12153
- Gockel, G. and W. Hachtel. 2000. Complete gene map of the plastid genome of the nonphotosynthetic euglenoid flagellate *Astasia longa*. *Protist* 151: 347-351.
- Goldblatt, P., A. Rodriguez, M. P. Powell, T. J. Davies, J. C. Manning, M. Van der Bank, and V. Savolainen. 2008. Iridaceae ‘Out of Australasia’? Phylogeny, biogeography, and divergence time based on plastid DNA sequences. *Systematic Botany* 3: 495–508.
- Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* 11: 725–736.
- Goulding, S. E., R. G. Olmstead, C. W. Morden and K. H. Wolfe. 1996. Ebb and flow of the chloroplast inverted repeat. *Molecular and General Genetics* 252: 195-206.

- 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. 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.
- Graham, S. W., J. M. Zgurski, M. A. McPherson, D. M. Cherniawsky, J. M. Saarela, E. S. C. Horne, et al. 2006. Robust inference of monocot deep phylogeny using an expanded multigene plastid data set. *Aliso* 22: 3–20.
- Grandcolas, P., J. Murienne, T. Robillard, L. Desutter-Grandcolas, H. Jourdan, E. Guilbert, and L. Deharveng. 2008. New Caledonia: a very old Darwinian island? *Philosophical Transactions of the Royal Society B: Biological Sciences* 363: 3309–3317.
- Gruzdev, E. V., A. V. Mardanov, A. V. Beletsky, E. Z. Kochieva, N. V. Ravin, and K. G. Skryabin. 2016. The complete chloroplast genome of parasitic flowering plant *Monotropa hypopitys*: extensive gene losses and size reduction. *Mitochondrial DNA Part B*: 1-2.
- Guindon, S., and O. Gascuel. 2003. A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. *Systematic Biology* 52: 696–704.
- Guisinger, M. M., J. V. Kuehl, J. L. Boore, and R. K. Jansen. 2008. Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. *Proceedings of the National Academy of Sciences* 105: 18424-18429.
- Guisinger, M. M., J. V. Kuehl, J. L. Boore, and R. K. Jansen. 2011. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. *Molecular Biology and Evolution* 28: 583-600.
- Guo, W., F. Grewe, A. Cobo-Clark, W. Fan, Z. Duan, R. P. Adams, A. E. Schwarzbach, A. E. and J. P. Mower. 2014. Predominant and substoichiometric isomers of the plastid genome coexist within Juniperus plants and have shifted multiple times during cupressophyte evolution. *Genome Biology and Evolution* 6: 580-590.

- 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: 4041-4048.
- Hanaoka, M., K. Kanamaru, M. Fujiwara, H. Takahashi, and K. Tanaka. Glutamyl-tRNA mediates a switch in RNA polymerase use during chloroplast biogenesis. *EMBO Reports* 6:545-550.
- Hansen, D. R., S. G. Dastidar, Z. Cai, C. Penaflor, J. V. Kuehl, J. L. Boore, and R. K. Jansen. 2007. Phylogenetic and evolutionary implications of complete chloroplast genome sequences of four early-diverging angiosperms: *Buxus* (Buxaceae), *Chloranthus* (Chloranthaceae), *Dioscorea* (Dioscoreaceae), and *Illicium* (Schisandraceae). *Molecular Phylogenetics and Evolution* 45:547–563.
- Hao, W., and J. D. Palmer. 2009. Fine-scale mergers of chloroplast and mitochondrial genes create functional, transcompartmentally chimeric mitochondrial genes. *Proceedings of the National Academy of Sciences USA* 106: 16728–16733.
- Heath, T. A., S. M. Hedtke, and D. M. Hillis. 2008. Taxon sampling and the accuracy of phylogenetic analyses. *Journal of Systematics and Evolution* 46: 239–257.
- Hedtke, S. M., T. M. Townsend, and D. M. Hillis. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Systematic Biology* 55: 522 – 529.
- Heide-Jørgensen, H. 2008. Parasitic Flowering Plants. Brill Academic Publishers, Leiden, Netherlands.
- Hendy, M. D., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. *Systematic Biology* 38: 297-309.
- Hertweck, K. L., M. S. Kinney, S. A. Stuart, O. Maurin, S. Mathews, M. W. Chase, M. A. Gandaldo, and J. C. Pires. 2015. Phylogenetics, divergence times and diversification from three genomic partitions in monocots. *Botanical Journal of the Linnean Society* 178: 375–393.

- Hill, R. S. and T. J. Brodribb. 1999. Southern conifers in time and space. *Australian Journal of Botany* 47: 639-696.
- Hillis, D. M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Systematic Biology* 47: 3–8.
- Hill, R. S. 2004. Origins of the southeastern Australian vegetation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 359: 1537-1549.
- Hilu, K. W. L. A., and H. Liang. 1997. The matK gene: sequence variation and application in plant systematics. *American Journal of Botany* 84: 830-830.
- Hilu, K. W., T. Borsch, K. Müller, D. E. Soltis, P. S. Soltis, V. Savolainen, M. W. Chase, et al. 2003. Angiosperm phylogeny based on *matK* sequence information. *American Journal of Botany* 90: 1758–1776.
- Hingorani, M. M., M. T. Washington, K. C. Moore, and S. S. Patel. 1997. The dTTPase mechanism of T7 DNA helicase resembles the binding change mechanism of the F1-ATPase. *Proceedings of the National Academy of Sciences* 94: 5012-5017
- Hipkins, V. D., K. A. Marshall, D. B. Neale, W. H. Rottmann, and S. H. Strauss. 1995. A mutation hotspot in the chloroplast genome of a conifer (Douglas-fir: *Pseudotsuga*) is caused by variability in the number of direct repeats derived from a partially duplicated tRNA gene. *Current Genetics* 27: 572-579.
- Hirao, T., A. Watanabe, M. Kurita, T. Kondo, and K. Takata. 2008. Complete nucleotide sequence of the *Cryptomeria japonica* D. Don. chloroplast genome and comparative chloroplast genomics: diversified genomic structure of coniferous species. *BMC Plant Biology* 8: 70
- Hitchcock, C. L. and A. Cronquist. 1973. Flora of the Pacific Northwest. University of Washington Press, Washington, USA.
- Hollingsworth, P. M., L. L. Forest, J. L. Spouge, M. Hajibabaei, S. Ratnasingham, M. van der Bank, M. W. Chase, et al. 2009. A DNA barcode for plants. *Proceedings of the National Academy of Sciences USA* 106: 12794–12797.
- Hollingsworth, P. M., S. W. Graham, and D. P. Little. 2011. Choosing and using a plant DNA barcode. *PLoS One* 6 : e19254.
- Howe, C. J. and A. G. Smith. 1991. Plants without chlorophyll. *Nature* 349:109.
- Huelsenbeck, J. P. 1997. Is the Felsenstein zone a fly trap? *Systematic Biology* 46: 69–74.

- Huelsenbeck, J. P. 1998 . Systematic bias in phylogenetic analysis: Is the Strepsiptera problem solved? *Systematic Biology* 47: 519–537.
- 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). *New Phytologist* 182: 719-726.
- Iles, W. J. D., S. Y. Smith, M. A. Gandolfo, and S. W. Graham. 2015. Monocot fossils suitable for molecular dating analyses. *Botanical Journal of the Linnean Society* 178: 346–374.
- Imhof, S. 2010. Are monocots particularly suited to develop mycoheterotrophy? In O. Seberg, G. Petersen, A. Barfod, and J. I. Davis [eds.], *Diversity, Phylogeny, and Evolution in the Monocotyledons*. Aarhus University Press, Copenhagen, Denmark. pp. 11–23.
- Jaffré, T. 1995. Distribution and ecology of the conifers of New Caledonia. In N. J. Enright, R. S. Hill [eds.] *Ecology of the Southern Conifers*. Melbourne University Press, Melbourne. pp.171-196
- Janouškovec, J., D. V. Tikhonenkov, F. Burki, A. T. Howe, M. Kolisko, 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 USA* 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 USA* 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* 28:835-847.
- Jansen, R. K. and T. A. Ruhlman. 2012. Plastid genomes of seed plants. In R. Bock, V. Knoop [eds.] *Genomics of Chloroplasts and Mitochondria*. Springer, Leiden, Netherlands. pp. 103-126.

- Jonker, F. P. 1938. A monograph of the Burmanniaceae. *Mededelingen van het Botanisch Museum en Herbarium van de Rijksuniversiteit te Utrecht* 51: 1-279.
- Kamikawa, R., G. Tanifuji, S. A. Ishikawa, K. I. Ishii, Y. Matsuno, N. T. Onodera, K. I. Ishida, T. Hashimoto, H. Miyashita, S. Mayama, and Y. Inagaki. 2015. Proposal of a twin-arginine translocator system-mediated constraint against loss of ATP synthase genes from nonphotosynthetic plastid genomes. *Molecular Biology and Evolution* 2: 2598-2604.
- Kelch, D. G. 1997. The phylogeny of the Podocarpaceae based on morphological evidence. *Systematic Botany* 22: 113-131
- Kelch, D. G. 1998. Phylogeny of Podocarpaceae: comparison of evidence from morphology and 18S rDNA. *American Journal of Botany* 85: 986-986.
- Kelchner, S. A. 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. *Annals of the Missouri Botanical Garden* 87: 482-498.
- Kikuchi, S., J. Bédard, M. Hirano, Y. Hirabayashi, M. Oishi, M. Imai, M. Takase, T. Ide, and M. Nakai. 2013. Uncovering the protein translocon at the chloroplast inner envelope membrane. *Science* 339: 571-574.
- Kim, J. 1996 . General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. *Systematic Biology* 45: 363-374.
- Kim, J. S., J.-K. Hong, M. W. Chase, M. F. Fay, and J.-H. Kim. 2013. Familial relationships of the monocot order Liliales based on a molecular phylogenetic analysis using four plastid loci: *matK*, *rbcL*, *atpB* and *atpF*. *Botanical Journal of the Linnean Society* 172: 5-21.
- 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* 267:492-497.
- Knox, E. B. 2014. The dynamic history of plastid genomes in the Campanulaceae sensu lato is unique among angiosperms. *Proceedings of the National Academy of Sciences USA* 111: 11097-11102.
- Kode, V., E. A. Mudd, S. Iamtham, and A. Day. 2005. The tobacco plastid *accD* gene is essential and is required for leaf development. *The Plant Journal* 44:237-244

- Kohzuma, K., C. Dal Bosco, A. Kanazawa, A. Dhingra, W. Nitschke, J. Meurer, and D.M. Kramer. 2012. Thioredoxin-insensitive plastid ATP synthase that performs moonlighting functions. *Proceedings of the National Academy of Sciences USA* 109: 3293-3298.
- Kolaczkowski, B., and J. W. Thornton. 2009. Long-branch attraction bias and inconsistency in Bayesian phylogenetics. *PLoS One* 4: e7891.
- de Koning A.P., and P. J. Keeling. 2006. The complete plastid genome of the parasitic green alga *Helicosporidium* sp. is highly reduced and structured. *BMC Biology* 4: 12.
- Koressaar, T., and M. Remm. 2007. Enhancements and modifications of primer design program Primer3. *Bioinformatics* 23: 1289–1291.
- Kranitz, M. L., E. Biffin, A. Clark, M. L. Hollingsworth, M. Ruhsam, M. F. Gardner, P. Thomas, R. R. Mill, R. A. Ennos, M. Gaudeul, A. J. Lowe, and P. M. Hollingsworth. 2014. Evolutionary diversification of New Caledonia *Araucaria*. *PLoS One* 9: e110308.
- Krause, K. 2008. From chloroplasts to “cryptic” plastids: evolution of plastid genomes in parasitic plants. *Current Genetics* 54: 111-121.
- Krause, K. 2011. Piecing together the puzzle of parasitic plant plastome evolution. *Planta* 234: 647-656.
- Krause, K. 2012. Plastid genomes of parasitic plants: a trail of reductions and losses. In C. E. Bullerwell [ed.]. *Organelle Genetics*. Springer Berlin Heidelberg. pp 79-103.
- Krause, K. and L. B. Scharff. 2014. Reduced genomes from parasitic plant plastids: templates for minimal plastomes? In: U. Lüttge, W. Beyschlag, J. Cushman [eds.] *Progress in Botany*. Springer Berlin Heidelberg. pp 97-115.
- Krause, K., Berg, S. and K. Krupinska. 2003. Plastid transcription in the holoparasitic plant genus *Cuscuta*: parallel loss of the *rrn16* PEP-promoter and of the *rpoA* and *rpoB* genes coding for the plastid-encoded RNA polymerase. *Planta* 216:815-823
- Kuijt, J. 1969. *The Biology of Parasitic Flowering Plants*. University of California Press, Berkeley.
- 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.

- Leake, J. R. 1994. The biology of myco-heterotrophic ('saprophytic') plants. *New Phytologist* 127: 171–216.
- Leake, J. R. 2004. Myco-heterotrophic/epiparasitic plant interactions with ectomycorrhizal and arbuscular mycorrhizal fungi. *Current Opinion in Plant Biology* 7: 422–428.
- Leake, J. R. 2005. Plants parasitic on fungi: Unearthing the fungi in mycoheterotrophs and debunking the 'saprophytic' plant myth. *Mycologist* 19: 113–122.
- Leebens-Mack, J. H., and C. W. 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.
- Legen, J., S. Kemp, K. Krause, B. Profanter, R. G. Herrmann, and R. M. Maier. 2002. Comparative analysis of plastid transcription profiles of entire plastid chromosomes from tobacco attributed to wild-type and PEP-deficient transcription machineries. *The Plant Journal* 31: 171-188
- Lemaire, B., S. Huysmans, E. Smets, and V. Merckx. 2011. Rate accelerations in nuclear 18S rDNA of mycoheterotrophic and parasitic angiosperms. *Journal of Plant Research*, 124: 561-576.
- Leslie, A. B., J. M. Beaulieu, H. S. Rai, P. R. Crane, M. J. Donoghue, and S. Mathews, 2012. Hemisphere-scale differences in conifer evolutionary dynamics. *Proceedings of the National Academy of Sciences USA* 109: 16217-16221.
- Li, D.-Z., L.-M. Gao, H.-T. Li, H. Wang, X.-J. Ge, J.-Q. Liu, Z.-D. Chen, et al. 2011. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences USA* 108: 19641–19646.
- Li, J., L. Gao, S. Chen, K. Tao, Y. Su, and T. Wang. 2016. Evolution of short inverted repeat in cupressophytes, transfer of *accD* to nucleus in *Sciadopitys verticillata* and phylogenetic position of *Sciadopityaceae*. *Scientific Reports* 6: 2093.
- Li, X., T.-C. Zhang, Q. Qiao, Z. Ren, J. Zhao, T. Yonezawa, M. Hasegawa, M. J. C. Crabbe, and J. Li. 2013. Complete chloroplast genome sequence of holoparasite *Cistanche deserticola* (Orobanchaceae) reveals gene loss and horizontal gene transfer from its host *Haloxylon ammodendron* (Chenopodiaceae). *PLoS One* 8:e58747.

- 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.
- Lin, C. S., J. J. Chen, Y. T. Huang, M. T. Chan, H. Daniell, W. J. Chang, C. T. Hsu, D. C. Liao, F. H. Wu, S. Y. Lin, and C. F. Liao. 2015. The location and translocation of *ndh* genes of chloroplast origin in the Orchidaceae family. *Scientific Reports* 5: 9040.
- Lio, P., and N. Goldman. 1998. Models of molecular evolution and phylogeny. *Genome Research* 8:1233–1244.
- 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.
- 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.
- 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:W575–W581.
- Lu, Y., J. H. Ran, D. M. Guo, Z. Y. Yang, and X. Q. Wang. 2014. Phylogeny and divergence times of gymnosperms inferred from single-copy nuclear genes. *PLoS One* 9: e107679.
- Maas-van de Kamer, H., and T. Weustenfeld. 1998. Triuridaceae. In: K. Kubitzki [ed.]The families and genera of vascular plants. III. Flowering plants: Monocotyledons. Berlin (Germany): Springer. pp. 452–458.
- 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.
- Magee, A. M., S. Aspinall, D. W. Rice, B. P. Cusack, M. Semon, A. S. Perry, S. Stefanovic, D. Milbourne, et al. 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Research* 20:1700–1710.
- Maliga, P. 1998. Two plastid RNA polymerases of higher plants: an evolving story. *Trends in Plant Science* 3: 4-6

- Mar, S. S. and R. M. K. Saunders. 2015. *Thismia hongkongensis* (Thismiaceae): a new myocheterotrophic species from Hong Kong, China, with observations on floral visitors and seed dispersal. *PhytoKeys* 46: 21-33.
- Margulis, L. 1981. Symbiosis in cell evolution: life and its environment on the early earth. Boston University, MA, USA.
- Maréchal, A., J. S. Parent, F. Véronneau-Lafortune, A. Joyeux, B. F. Lang, and N. Brisson, N. 2009. Whirly proteins maintain plastid genome stability in *Arabidopsis*. *Proceedings of the National Academy of Sciences USA* 106: 14693-14698.
- Maréchal, A., and N. Brisson. 2010. Recombination and the maintenance of plant organelle genome stability. *New Phytologist* 186: 299–317.
- Martin, W. and R. G. Herrmann. 1998. Gene transfer from organelles to the nucleus: how much, what happens, and why? *Plant Physiology* 118: 9-17
- Martin, W., B. Stoebe, V. Goremykin, S. Hansmann, M. Hasegawa, and K. V. Kowallik. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* 393: 162-165.
- Martin, M., and B. Sabater. 2010. Plastid *ndh* genes in plant evolution. *Plant Physiology and Biochemistry* 48:636–645.
- Martin, G. E., M. Rousseau-Gueutin, S. Cordonnier, O. Lima, S. Michon-Coudouel, D. Naquin, J. F. de Carvalho, M. Ainouche, A. Salmon, and A. Ainouche. 2014. The first complete chloroplast genome of the Genistoid legume *Lupinus luteus*: evidence for a novel major lineage-specific rearrangement and new insights regarding plastome evolution in the legume family. *Annals of Botany* 113: 1197-1210.
- Matsuo, M., Y. Ito, R. Yamauchi, and J. Obokata. 2005. The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flex. *Plant Cell* 17: 665–675.
- McLoughlin, S. 2001. The breakup history of Gondwana and its impact on pre-Cenozoic floristic provincialism. *Australian Journal of Botany* 49: 271-300.
- McNeal, J. R., J. V. Kuehl, J. L. Boore, and C. W. dePamphilis. 2007. Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus *Cuscuta*. *BMC Plant Biology* 7: 57.

- 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: e5982.
- Mennes, C. B., E. F. Smets, S. N. Moses, and V. S. F. T. Merckx. 2013. New insights in the long-debated evolutionary history of Triuridaceae. *Molecular Phylogenetics and Evolution* 69: 994–1004.
- 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.
- Merckx, V., P. Schols, H. Maas-Van De Kamer, P. Maas, S. Huysmans, and E. Smets. 2006. Phylogeny and evolution of Burmanniaceae (Dioscoreales) based on nuclear and mitochondrial data. *American Journal of Botany* 93: 1684–1698.
- Merckx, V., L. W. Chatrou, B. Lemaire, M. N. Sainge, S. Huysmans, and E. F. Smets. 2008 . Diversification of myco-heterotrophic angiosperms: Evidence from Burmanniaceae. *BMC Evolutionary Biology* 8: 178.
- Merckx, V. S., S. B. Janssens, N. A. Hynson, C. D. Specht, T. D. Bruns, and E. F. Smets. 2012. Mycoheterotrophic interactions are not limited to a narrow phylogenetic range of arbuscular mycorrhizal fungi. *Molecular Ecology* 21: 1524-1532.
- Merckx, V. S. F. T., C. B. Mennes, K. G. Peay, and J. Geml. 2013a. Evolution and diversification. In V. Merckx [ed.], *Mycoheterotrophy: The biology of plants living on fungi*. Springer-Verlag, New York, New York, USA. pp 222–226.
- Merckx, V. S. F. T., J. V. Freudenstein, J. Kissling, M. J. M. Christenhusz, R. E. Stotler, B. Crandall-Stotler, N. Wickett, P. J. Rudall, H. Maas-van de Kamer, and P. J. M. Maas. 2013b. Taxonomy and classification. In V. Merckx [ed.], *Mycoheterotrophy: The Biology of Plants Living on Fungi*. Springer-Verlag, New York, New York, USA. pp. 19-102.
- Merckx, V., M. I. Bidartondo, and N. A. Hynson. 2009a . Myco-heterotrophy: when fungi host plants. *Annals of Botany* 104: 1255–1261.

- Merckx, V. S. F., T. Bakker, S. Huysman, and E. Smets. 2009b. Bias and conflict in phylogenetic inference of myco-heterotrophic plants: A case study in Thismiaceae. *Cladistics* 25: 64–77.
- Merckx, V., S. Huysmans, and E. Smets. 2010. Cretaceous origins of mycoheterotrophic lineages in Dioscoreales. In: O. Seberg, G. Petersen, A. Barfod, and J. I. Davis [eds.], *Diversity, Phylogeny, and Evolution in the Monocotyledons*. Aarhus University Press, Copenhagen, Denmark. pp 39–53.
- 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. Merckx [ed.], *Mycoheterotrophy: The Biology of Plants Living on Fungi*. Springer-Verlag, New York, New York, USA. pp 1–18.
- Miers, J. 1842. Description of a new genus of plants from Brazil. *Transactions of the Linnean Society of London* 19:77–80.
- Millen, R. S., R. G. Olmstead, K. L. Adams, J. D. Palmer, N. T. Lao, L. Heggie, T. A. Kavanagh, J. M. Hibberd, J. C. Gray, C. W. Morden, and P. J. Calie. 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. In: *Proceedings of the Gateway Computing Environments Workshop (GCE)*. 2010 Nov 14. New Orleans, LA. pp. 1-8.
- Minkin, I., H. Pham, E. Starostina, N. Vyahhi, and S. Pham. 2013. C-Sibelia: an easy-to-use and highly accurate tool for bacterial genome comparison. *F1000Research* 2: 258
- Miyazawa, S. 2013. Superiority of a mechanistic codon substitution model even for protein sequences in phylogenetic analysis. *BMC Evolutionary Biology* 13: 257
- Mohr, G., P. S. Perlman, and A. M. Lambowitz. 1993. Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic Acids Research* 21: 4991-4997.
- 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.
- Molloy, B.P. 1995. *Manoao* (Podocarpaceae), a new monotypic conifer genus endemic to New Zealand. *New Zealand Journal of Botany* 33: 183-201.
- Motomura, H., M. A. Selosse, F. Martos, A. Kagawa, A. and T. Yukawa. 2010. Mycoheterotrophy evolved from mixotrophic ancestors: evidence in *Cymbidium* (Orchidaceae). *Annals of Botany* 106: 573-581.
- Mower, J. P., S. Stefanović, G. J. Young, and J. D. Palmer. 2004. Plant genetics: Gene transfer from parasitic to host plants. *Nature* 432: 165–166.
- Nakai, M. 2015. The TIC complex uncovered: The alternative view on the molecular mechanism of protein translocation across the inner envelope membrane of chloroplasts. *Biochimica et Biophysica Acta (BBA)-Bioenergetics* 1847: 957-967.
- do Nascimento Vieira, L., H. Faoro, M. Rogalski, H. P. de Freitas Fraga, R. L. A. Cardoso, E. M. de Souza, F. de Oliveira Pedrosa, R. O. Nodari, and M. P. Guerra. 2014. The complete chloroplast genome sequence of *Podocarpus lambertii*: genome structure, evolutionary aspects, gene content and SSR detection. *PLoS One* 9: e90618.
- do Nascimento Vieira, L., M. Rogalski, H., Faoro, H. P. de Freitas Fraga, K. G. dos Anjos, G. F. A. Picchi, R. O. Nodari, F. de Oliveira Pedrosa, E. M. de Souza, and M. P. Guerra. 2016. The plastome sequence of the endemic Amazonian conifer, *Retrophyllum piresii* (Silba) CN Page, reveals different recombination events and plastome isoforms. *Tree Genetics & Genomes* 12: 1-11
- Naumann, J., J. P. Der, E. K. Wafula, S. S. Jones, S. T. Wagner, L. A. Honaas, P. E. Ralph, J. F. Bolin, E. Maass, C. Neinhuis, S. Wanke, and C. W. dePamphilis. 2016. Detecting and characterizing the highly divergent plastid genome of the nonphotosynthetic parasitic plant *Hydnora visseri* (Hydnoraceae). *Genome Biology and Evolution*, p.evv256.
- Newmaster, S. G., A. J. Fazekas, R. A. D. Steeves, and J. Janovec. 2008. Testing candidate plant barcode regions in the Myristicaceae. *Molecular Ecology Resources* 8: 480–490.

- Neyland, R. and L. E. Urbatsch. 1996. Phylogeny of subfamily Epidendroideae (Orchidaceae) inferred from *ndhF* chloroplast gene sequences. *American Journal of Botany* 83: 1195-1206.
- Neyland, R., and M. Hennigan. 2003. A phylogenetic analysis of large-subunit (26S) ribosome DNA sequences suggests that the Corsiaceae are polyphyletic. *New Zealand Journal of Botany* 41: 1–11.
- Neyland, R. 2002. A phylogeny inferred from large-subunit (26S) ribosomal DNA sequences suggest that Burmanniales are polyphyletic. *Australian Systematic Biology* 15: 19–28.
- Obornik, M., and B. R. Green. 2005. Mosaic origin of the heme biosynthesis pathway in photosynthetic eukaryotes. *Molecular Biology and Evolution* 22:2343–2353.
- 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. 1983. Chloroplast DNA exists in two orientations. *Nature* 301: 92–93
- Palmer, J. D. and W. F. Thompson. 1982. Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell* 29: 537-550.
- Palmer, J. D. 1985. Comparative organization of chloroplast genomes. *Annual Review of Genetics* 19: 325-354.
- Palmer, J. D. and D. B. Stein. 1986. Conservation of chloroplast genome structure among vascular plants. *Current Genetics* 10: 823-833.
- Palmer, J. D., and L. A. Herbon. 1988. Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. *Journal of Molecular Evolution* 28: 87– 97.
- dePamphilis, C. W., N. D. Young, and A. D. Wolfe. 1997. Evolution of plastid gene *rps2* in a lineage of hemiparasitic and holoparasitic plants: many losses of photosynthesis and complex patterns of rate variation. *Proceedings of the National Academy of Sciences* 94: 7367-7372.
- Parks, S. L., and N. Goldman. 2014. Maximum likelihood inference of small trees in the presence of long branches. *Systematic Biology* 63: 798–811.

- Petersen, G., O. Seberg, and J. I. Davis. 2013. Phylogeny of the Liliales (Monocotyledons) with special emphasis on data partitioning congruence and RNA editing. *Cladistics* 29: 274-295
- Petersen, G., A. Cuenca, and O. Seberg. 2015. Plastome evolution in hemiparasitic mistletoes. *Genome Biology and Evolution*. 9: 2520-2532.
- Philippe, H., H. Brinkmann, D. V. Lavrov, D. T. J. Littlewood, M. Manuel, G. Wörheide, and D. Baurain. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biology* 9: e1000602.
- Pojar, J., A. MacKinnon, and P. B. Alaback. 1994. Plants of coastal British Columbia. Lone Pine Publishing, AB, Canada.
- Pollock, D. D., D. J. Zwickl, J. A. McGuire, and D. M. Hillis. 2002. Increased taxon sampling is advantageous for phylogenetic inference. *Systematic Biology* 51: 664.
- Rai, H. S., H. O'Brien, P. A. Reeves, R. G. Olmstead, and S. W. Graham. 2003. Inference of higher-order relationships in the cycads from a large chloroplast data set. *Molecular Phylogenetics and Evolution* 29: 350–359.
- Rambaut, A., M. A. Suchard, D. Xie, and A. J. Drummond. 2014. Tracer v1. 6. Available from <http://beast.bio.ed.ac.uk/Tracer> [accessed 15 January 2016].
- Rambaut, A. 2002. Se-AL v. 2.0a11: Sequence alignment program. Available from <http://tree.bio.ed.ac.uk/software/seal/> [accessed 15 January 2016].
- Rambaut, A. 2006. FigTree. Available from <http://tree.bio.ed.ac.uk/software/figtree/> [accessed 15 January 2016].
- Randle, C. P. and A. D. Wolfe. 2005. The evolution and expression of *rbcL* in holoparasitic sister-genera *Harveya* and *Hyobanche* (Orobanchaceae). *American Journal of Botany* 92: 1575-1585
- Rasmussen, H. N. 1995. Terrestrial orchids: From seed to mycotrophic plant. Cambridge University Press, Cambridge, UK.
- Reeves, G., M. W. Chase, P. Goldblatt, P. Rudall, M. F. Fay, A. V. Cox, B. Lejeune, and T. Souza-Chies. 2001. Molecular systematics of Iridaceae: Evidence from four plastid DNA regions. *American Journal of Botany* 88: 2074–2087.
- Rogalski, M., S. Ruf, and R. Bock, R. 2006. Tobacco plastid ribosomal protein S18 is essential for cell survival. *Nucleic Acids Research* 34: 4537-4545.

- Rogalski, M., D. Karcher, and R. Bock. 2008a. Superwobbling facilitates translation with reduced tRNA sets. *Nature Structural and Molecular Biology* 15:192–198.
- Rogalski, M., M. A. Schöttler, W. Thiele, W. X. Schulze, and R. Bock. 2008b. *Rpl33*, a nonessential plastid-encoded ribosomal protein in tobacco, is required under cold stress conditions. *The Plant Cell* 20: 2221-2237.
- Ross, T. G., C. F. Barrett, M. Soto Gomez, V. K. Y. Lam, C. L. Henriquez, D. H. Les, J. I. Davis, A. Cuenca, G. Petersen, O. Seberg, and M. Thadeo. 2016. Plastid phylogenomics and molecular evolution of Alismatales. *Cladistics* 32: 160–178.
- 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 (*accD*) gene by recent transfer to the nucleus in some angiosperm lineages. *Plant Physiology* 161: 1918–1929.
- Rudall, P. J., M. Alves, and M. das Graças Sajo. 2016. Inside-out flowers of *Lacandonia brasiliensis* (Triuridaceae) provide new insights into fundamental aspects of floral patterning. *PeerJ*. doi: 10.7717/peerj.1653
- Rudall, P. J., and R. M. Bateman. 2006. Morphological phylogenetic analysis of Pandanales: Testing contrasting hypotheses of floral evolution. *Systematic Botany* 31: 223–238.
- Ruhfel, B. R., M. A. Gitzendanner, P. S. Soltis, D. E. Soltis, and J. G. Burleigh. 2014. From algae to angiosperm – 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, J. C. Blazier, X. Jin, M. C. Shih, and R.K. Jansen. 2015. NDH expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. *BMC Plant Biology* 15: 100.
- 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.

- Rumeau, D., G. Peltier, and L. Cournac. 2007. Chlororespiration and cyclic electron flow around PSI during photosynthesis and plant stress response. *Plant, Cell & Environment* 30: 1041-1051.
- Rychlik, W. 2007. OLIGO 7 primer analysis software. In A. Yuryev [ed.], *Methods in molecular biology*, vol. 402. Humana Press, New York, New York, USA. pp. 35–59.
- 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: 743–754.
- Salmon, J. T. 1980. *The Native Trees of New Zealand*. Reed Books, New Zealand. pp. 50-92.
- Samigullin, T. H., M. D. Logacheva, A. A. Penin, and C. M. Vallejo-Roman. 2016. Complete plastid genome of the recent holoparasite *Lathraea squamaria* reveals earliest stages of plastome reduction in Orobanchaceae. *PloS One* 11: e0150718.
- Sanderson, M. J., D. Copetti, A. Búrquez, E. Bustamante, J. L. Charboneau, L. E. Eguiarte, S. Kumar, H. O. Lee, J. Lee, M. McMahon, and K. Steele. 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.
- de Santis-Maciossek, D., W. Kofer, A. Bock, S. Schoch, R. M. Maier, G. Wanner, W. Rüdiger, H. U. Koop, and R. G. Herrmann. 1999. Targeted disruption of the plastid RNA polymerase genes *rpoA*, B and C1: molecular biology, biochemistry and ultrastructure. *The Plant Journal* 18:477-489.
- Schelkunov, M. I., V. Y. Shtratnikova, M. S. Nuraliev, M.-A. Selosse, A. A. Penin, and M. D. 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* 7: 1179–1191.
- Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.
- 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
- Shanklin, J., N. D. DeWitt, and J. M. Flanagan. 1995. The stroma of higher plant plastids contain ClpP and ClpC, functional homologs of *Escherichia coli* ClpP and ClpA: An archetypal two-component ATP-dependent protease. *Plant Cell* 7: 1713.
- Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. *Systematic Biology* 51: 492–508.
- Shimodaira, H., and M. Hasegawa. 2001. CONSEL: For assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17: 1246–1247.
- Siemeister, G., and W. Hachtel. 1990. Structure and expression of a gene encoding the large subunit of ribulose-1, 5-bisphosphate carboxylase (*rbcL*) in the colourless euglenoid flagellate *Astasia longa*. *Plant Molecular Biology* 14: 825-833.
- Silvestro, D., and I. Michalak. 2012. raxmlGUI: A graphical front-end for RAxML. *Organisms, Diversity & Evolution* 12: 33 –337.
- Simmons, M. P., K. M. Pickett, and M. Miya. 2004. How meaningful are Bayesian support values? *Molecular Biology and Evolution* 21: 188–199.
- Simmons, M. P., and H. Ochoterena. 2000. Gaps as characters in sequence based phylogenetic analyses. *Systematic Biology* 49: 369–381.
- Sinclair, W. T., R. R. Mill, M. F. Gardner, P. Woltz, T. Jaffré, J. Preston, M. L. Hollingsworth, A. Ponge, and M. Möller. 2002. Evolutionary relationships of the New Caledonian heterotrophic conifer, *Parasitaxus usta* (Podocarpaceae), inferred from chloroplast *trnL-F* intron/spacer and nuclear rDNA ITS2 sequences. *Plant Systematics and Evolution* 233: 79-104.
- Sloan, D. B., D. A. Triant, N. J. Forrester, L. M. Bergner, M. Wu, and D. R. Taylor. 2014. A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae (Caryophyllaceae). *Molecular Phylogenetics and Evolution* 72: 82-89
- Smith, D. R., and R. W. Lee. 2014. A plastid without a genome: evidence from the nonphotosynthetic green algal genus *Polytomella*. *Plant Physiology* 164: 1812–1819.
- Soltis, D. E., P. S. Soltis, M. W. Chase, M. E. Mort, D. C. Albach, M. Zanis, V.

- Savolainen, et al. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. *Botanical Journal of the Linnean Society* 133 :381–461.
- Soltis, D. E., and P. S. Soltis. 2004. *Amborella* not a “basal angiosperm”? Not so fast. *American Journal of Botany* 91: 997–1001.
- Soltis, D. E., S. A. Smith, N. Cellinese, K. J. Wurdack, D. C. Tank, S. F. Brockington, N. F. Refulio-Rodriguez, et al. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany* 98: 704–730.
- Stamatakis, A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22: 2688 – 2690.
- Stange-Thomann, N., H. U. Thomann, A. J. Lloyd, H. Lyman, and D. Söll. 1994. A point mutation in *Euglena gracilis* chloroplast tRNA (Glu) uncouples protein and chlorophyll biosynthesis. *Proceedings of the National Academy of Sciences USA* 91: 7947-7951
- Straub, S. C. K., M. Fishbein, T. Livshultz, Z. Foster, M. Parks, K. Weitmier, R. Cronn, and A. Liston. 2011. Building a model: Developing genome resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics* 12: 211.
- Strauss, S. H., J. D. Palmer, G. T. Howe, and A. H. Doerksen. 1988. Chloroplast genomes of two conifers lack a large inverted repeat and are extensively rearranged. *Proceedings of the National Academy of Sciences USA* 85: 3898-3902.
- Strotmann, H., N. Shavit, and S. Leu. 1998. Assembly and function of the chloroplast ATP synthase. In J.-D. Rochaix, M. Goldschmidt-Clermont, S. Merchant [eds.] *The Molecular Biology of Chloroplasts and Mitochondria in Chlamydomonas*. Kluwer Academic Publishers, Netherlands. pp. 477-500.
- Suetsugu K., A. Kawakita, and M. Kato. 2015. Avian seed dispersal in a mycoheterotrophic orchid *Cyrtosia septentrionalis*. *Nature Plants* 1: 15052.
- Sugiura, C., and M. Sugita. 2004. Plastid transformation reveals that moss tRNA Arg-CCG is not essential for plastid function. *The Plant Journal* 40: 314-321.
- Sullivan, J., and P. Joyce. 2005. Model selection in phylogenetics. *Annual Review of Ecology, Evolution and Systematics* 36: 445- 466.
- Suorsa, M., S. Sirpiö, and E. M. Aro. 2009. Towards characterization of the chloroplast.

- NAD(P)H dehydrogenase complex. *Molecular Plant* 2: 1127-1140.
- Swiatecka-Hagenbruch, M., K. Liere, and T. Börner. 2007. High diversity of plastidial promoters in *Arabidopsis thaliana*. *Molecular Genetics and Genomics* 277: 725-734.
- Swofford, D. L., P. J. Waddell, J. P. Huelsenbeck, P. G. Foster, P. O. Lewis, and J. S. Rogers. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Systematic Biology* 50: 525– 539.
- Swofford, D. L. 2003. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sinauer, Sunderland, Massachusetts, USA.
- Takhtajan, A. 1997. Diversity and classification of flowering plants. New York: Columbia University Press.
- Takhtajan, A. 2009. Flowering Plants. Springer Science & Business Media, Berlin, Germany.
- Tamura, M. N., J. Yamashita, S. Fuse, and M. Haraguchi. 2004. Molecular phylogeny of monocotyledons inferred from combined analysis of plastid *matK* and *rbcL* gene sequences. *Journal of Plant Research* 117: 109-120.
- Tanaka, R., and A. Tanaka. 2007. Tetrapyrrole biosynthesis in higher plants. *Annual Review of Plant Biology* 58: 321-346
- Timmis, J. N., M. A. Aylliffe, C. Y. Huang and W. Martin. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nature Reviews Genetics* 5: 123-135.
- Trudell, S. A., P. T. Rygielwicz, and R. L. Edmonds 2003. Nitrogen and carbon stable isotope abundances support the myco-heterotrophic nature and host-specificity of certain achlorophyllous plants. *New Phytologist*, 160: 391-401.
- Ueda, M., M. Fujimoto, S. I. Arimura, J. Murata, N. Tsutsumi, and K.I. Kadowaki. 2007. Loss of the *rpl32* gene from the chloroplast genome and subsequent acquisition of a preexisting transit peptide within the nuclear gene in *Populus*. *Gene* 401:51-56.
- Ueda, M., T. Nishikawa, M. Fujimoto, H. Takanashi, S. I. Arimura, N. Tsutsumi, and K.I. Kadowaki. 2008. Substitution of the gene for chloroplast *rps16* was assisted by generation of a dual targeting signal. *Molecular Biology and Evolution* 25: 1566-1575.

- Untergasser, A., H. Nijveen, X. Rao, T. Bisseling, R. Geurts, and J. A. Leunissen. 2007. Primer3Plus, an enhanced web interface to Primer 3. *Nucleic Acid Research* 35:W71–W74.
- Vogel, J., T. Hübschmann, T. Börner, and W. R. Hess. 1997. Splicing and intron-internal RNA editing of *trnK-matK* transcripts in barley plastids: support for *matK* as an essential splice factor. *Journal of Molecular Biology* 270: 179–187.
- Vogel, J., T. Börner, and W. R. Hess. 1999. Comparative analysis of splicing of the complete chloroplast group II introns in three higher plants. *Nucleic Acids Research* 27: 3866-3874.
- de Vries, J., F. J. Sousa, B. Bölter, J. Soll, and S. B. Gould. 2015. YCF1: a green TIC?. *The Plant Cell* 27: 1827-1833.
- Wakasugi, T., J. Tsudzuki, S. Ito, K. Nakashima, T. Tsudzuki, and M. Sugiura. 1994. Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. *Proceedings of the National Academy of Sciences USA* 91: 9794-9798.
- Walker, J. E., and A. L. Cozens. 1986. Evolution of ATP synthase. *Chemica Scripta* 26: 263-272.
- Wang, R. J., C. L. Cheng, C. C. Chang, C. L. Wu, T. M. Su, and S.M. Chaw. 2008. Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. *BMC Evolutionary Biology* 8:36.
- Wang, B., and Y. L. Qiu. 2006. Phylogenetic distribution and evolution of mycorrhizas in land plants. *Mycorrhiza* 16: 299-363.
- Waterman, R. J., M. R. Klooster, H. Hentrich, and M. I. Bidartondo. 2013 . In V. Merckx [ed.], *Mycoheterotrophy: The Biology of Plants Living on Fungi*. Springer-Verlag, New York, New York, USA. pp 267–296.
- Weber, J. 2006. ATP synthase: subunit–subunit interactions in the stator stalk. *Biochimica et Biophysica Acta –Bioenergetics* 1757: 1162-1170
- Weng, M.L., J. C. Blazier, M. Govindu, and R.K. Jansen. 2014. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats and nucleotide substitution rates. *Molecular Biology and Evolution* 31: 645-659.

- 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.
- Wicke, S., K. F. Müller, C. W. dePamphilis, 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. *Plant Cell* 25: 3711–3725.
- 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* 31:529545.
- Wickett, N. J., Y. Zhang, S. K. Hansen, J. M. Roper, J. V. Kuehl, S. A. Plock, P. G. Wolfe, 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.
- Wiens, J. J., and J. Tiu. 2012. Highly incomplete taxa can rescue phylogenetic analyses from the negative impacts of limited taxon sampling. *PLoS One* 7: e42925.
- Wiens, J. J. 2005. Can incomplete taxa rescue phylogenetic analyses from long branch attraction? *Systematic Biology* 54: 731–742.
- Wiens, J. J. 2006. Missing data and the design of phylogenetic analyses. *Journal of Biomedical Informatics* 39: 34–42.
- Wilson, R. J., P. W. Denny, K. Rangachari, K. Roberts, A. Roy, A. Whyte, M. Strath, D. J. Moore, P. W. Moore and D. H. Williamson. 1996. Complete gene map of the plastid-like DNA of malaria parasite *Plasmodium falciparum*. *Journal of Molecular Biology* 261:155–172.
- Wintersinger, J. A., and J. D. Wasmuth. 2015. Kablammo: an interactive, web-based BLAST results visualizer. *Bioinformatics* 31:1305-1306.
- Wolfe, K. H., W.-H. Li, and P. M Sharp. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences USA* 84: 9054-9058.

- 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, USA* 89: 10648–10652.
- Woltz, P., R. A. Stockey, M. Gondran, J. F. Cherrier, and J. Bernard. 1994. Interspecific parasitism in the gymnosperms: unpublished data on two endemic New Caledonian Podocarpaceae using scanning electron microscopy. *Acta Botanica Gallica* 141: 731-746.
- Wu, F. H., M. T. Chan, D. 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.
- Wu, C. S., Y. N. Wang, C. Y. Hsu, C. P. Lin, and S. M. Chaw. 2011. Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. *Genome Biology and Evolution* 3: 1284-1295.
- Wu, C. S., and S. M. Chaw. 2014. Highly rearranged and size-variable chloroplast genomes in conifers II clade (cupressophytes): evolution towards shorter intergenic spacers. *Plant Biotechnology Journal* 12: 344-353.
- Wyman, S. K., R. K. Jansen, and J. L. Boore. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20:3252–3255.
- Yang, Z., and B. Rannala. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. *Systematic Biology* 54: 455–470.
- Yang, Z., and B. Rannala. 2012. Molecular phylogenetics: principles and practice. *Nature Reviews. Genetics* 13: 303–314.
- Yang, Z. 1996. Phylogenetic analysis using parsimony and likelihood methods. *Journal of Molecular Evolution* 42: 294–307.
- Yang, Z. 2006. Computational molecular evolution. Oxford University Press, Oxford, UK.
- Yang, Z. 2007. PAML4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24: 1586–1591.

- Yap, J. Y. S., T. Rohner, A. Greenfield, M. van der Merwe, H. McPherson, W. Glenn, G. Kornfeld, E. Marendy, A. Y. Pan, A. Wilton, and M. R. Wilkins. 2015. Complete chloroplast genome of the Wollemi pine (*Wollemia nobilis*): structure and evolution. *PLoS One* 10: e0128126.
- Yi, X., L. Gao, B. Wang, Y. Su, and T. Wang. 2013. The complete chloroplast genome sequence of *Cephalotaxus oliveri* (Cephalotaxaceae): evolutionary comparison of cephalotaxus chloroplast DNAs and insights into the loss of inverted repeat copies in gymnosperms. *Genome Biology and Evolution* 5: 688–698.
- Yoshida, S., S. Maruyama, H. Nozaki, and K. Shirasu. 2010. Horizontal gene transfer by the parasitic plant *Striga hermonthica*. *Science* 328: 1128.
- Young, N. D., and C. W. dePamphilis. 2005. Rate variation in parasitic plants: correlated and uncorrelated patterns among plastid genes of different function. *BMC Evolution and Biology* 5:16.
- 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.
- Zhang, J., R. Nielsen, and Z. Yang. 2005. Evolution of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution* 22:2472–2479.
- Zhang, D., and R. M. K. Saunders. 1999. *Corsiopsis chinensis* gen. et sp. nov. (Corsiaceae): first record of the family in Asia. *Systematic Botany* 24: 311-314.
- Zhelyazkova, P., C. M. Sharma, K. U. Förstner, K. Liere, J. Vogel, and T. Börner. 2012. The primary transcriptome of barley chloroplasts: noncoding RNAs and the dominating role of the plastid encoded RNA polymerase. *Plant Cell* 24:123–136.
- Zhu, A., W. Guo, S. Gupta, W. Fan, and J. P. Mower. 2016. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytologist* 4: 1747-1756.
- 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 Science USA* 107:3245–3250.

Zwickl, D. J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion [Ph.D. dissertation]. The University of Texas at Austin.

Appendices

Appendix A: Supplementary tables and figures for Chapter 2

Table A.1 Accessions and herbarium information for all newly sequenced species; sequences of several mycoheterotrophic taxa retrieved from GenBank are also noted and sources of other published sequences are footnoted. Fully mycoheterotrophic (achlorophyllous) lineages are indicated with a ‘‡.’ Genes retrieved by next-generation sequencing are indicated in bold; all other new sequences were retrieved by PCR and direct (Sanger) sequencing. Genes known to be missing are noted as ‘lost’ based on evidence from full plastid circles (noted as published or unpublished data). Instances where we recovered contaminant sequences are also noted (see text). A dash (‘-’) indicates that a region was not retrievable by PCR. Herbarium acronyms follow Thiers (continuously updated).

Family	Taxon ¹	Voucher, herbarium	Plastid gene		
			<i>accD</i>	<i>clpP</i>	<i>matK</i>
Acorales					
Acoraceae	<i>Acorus calamus</i> L.	R. Olmstead 97-149, DNA	KU127422	KU127420	KU127421
Alismatales					
Alismataceae	<i>Alisma plantago-aquatica</i> L.	S.Y. Smith 47, ALTA	KU127346	KU127345	KU127344

Family	Taxon ¹	Voucher, herbarium	Plastid gene		
			<i>accD</i>	<i>clpP</i>	<i>matK</i>
Araceae	<i>Gymnostachys anceps</i> R.Br.	M.W. Chase 3841, K	KU127277	KU127276	KU127275
Cymodoceaceae	<i>Halodule univervis</i> (Forssk.) Boiss	D.T. Dy <i>s.n.</i> , UBC	KU127294	KU127295	KU127293
Hydrocharitaceae	<i>Stratiotes aloides</i> L.	Bogner <i>s.n.</i> , ALTA	KU127278	KU127280	KU127279
	<i>Thalassia hemprichii</i> (Ehrenb. Ex Solms) Asch.	D.T. Dy <i>s.n.</i> , UBC	KU127391	KU127392	KU127393
Juncaginaceae	<i>Maundia triglochinooides</i> F.Muell.	L. Stanberg & G. Sainty LS 80, NSW	KU127351	KU127353	KU127352
Potamogetonaceae	<i>Groenlandia densa</i> (L.) Fourr.	Bogner <i>s.n.</i> , ALTA	KU127318	KU127319	KU127320
	<i>Potamogeton richardsonii</i> (A.Benn.) Rydb.	S.Y. Smith 50, ALTA	KU147406	KU147405	KU147404
Tofieldiaceae	<i>Zannichellia palustris</i> L.	J. Bruinsma <i>s.n.</i> , UBC	KU127488	KU127487	KU127486
	<i>Harperocallis flava</i> McDaniel	M.W. Chase 306, NCU	KU127409	KU127410	KU127411
	<i>Pleea tenuifolia</i> Michx.	W. Zomlefer 789, GA	KU127292	KU127291	KU127290
	<i>Tofieldia pusilla</i> (Michx.) Pers.	M.J. Waterway 2006-101, UBC	KU127447	KU127446	KU127445
Asparagales					
Amaryllidaceae	<i>Narcissus elegans</i> (Haw.) Spach	S.C.H. Barrett 1434, TRT	KU127379	KU127380	KU12738
Asparagaceae	<i>Asparagus</i> sp. Tourn. ex L.	V. Lam VL013, UBC	KU127362	KU127361	KU127360
	<i>Asphodelus albus</i> Mill.	L. Harder 1-000430, ALTA	KU147407	KU14740	KU147409

Family	Taxon ¹	Voucher, herbarium	Plastid gene		
			<i>accD</i>	<i>clpP</i>	<i>matK</i>
	<i>Chlorophytum comosum</i> (Thunb.) Jacques	M.A. McPherson 000321-1, ALTA	KU127258	KU127259	KU127257
	<i>Leopoldia comosa</i> (L.) Parl.	L. Harder 000419-1, ALTA	KU127479	KU127480	KU127481
Iridaceae	‡ <i>Geosiris aphylla</i> Baill.	Prance 30781, K	KU127298	KU127297	KU127297
	<i>Iris missouriensis</i> Nutt.	M.A. McPherson 000707-5a-7, ALTA	KU127373	KU127374	KU127375
	<i>Sisyrinchium montanum</i> Greene	M.A. McPherson 990704-71, ALTA	KU127389	KU127388	KU127390
Lanariaceae	<i>Lanaria lanata</i> (L.) T.Durand & Schinz	M.W. Chase 478, NCU	KU127322	KU127323	KU127321
Orchidaceae	<i>Coelogyne cristata</i> Lindl.	M.A. McPherson 010921-1, ALTA	KU127465	KU127464	KU127464
	‡ <i>Corallorhiza striata</i> var. <i>vreelandii</i> (Rydb.) L.O. Williams	see Barrett and Davis (2012)	JX087681.1³	JX087681.1³	JX087681.1³
	‡ <i>Neottia nidus-avis</i> (L.) Rich.	see Logacheva et al. (2011)	JF325876.1³	JF325876.1³	JF325876.1³
	‡ <i>Rhizanthella gardneri</i> R.S.Rogers	see Delannoy et al. (2011)	NC_014874³	NC_014874³	NC_014874³
Xanthorrhoeaceae	<i>Hemerocallis littorea</i> Makino	M.W. Chase 3833, K	KU127312	KU127313	KU127314
Arecaeae	<i>Roystonea princeps</i> (Becc.) Burret	E. Santiago #J-4, UPR	KU127448	KU127449	KU127450

Family	Taxon ¹	Voucher, herbarium	Plastid gene		
			<i>accD</i>	<i>clpP</i>	<i>matK</i>
Commelinales					
Pontederiaceae	<i>Hydrothrix gardneri</i> Hook.f.	Barrett 1414, TRT	KU127494	KU127495	KU127496
Dioscoreales					
Burmanniaceae	‡ <i>Apteria aphylla</i> (Nutt.) Barnhart ex Small	M.W. Chase 156, NCU	KU127387	– ²	– ²
	‡ <i>Apteria aphylla</i> (Nutt.) Barnhart ex Small	D.M. McNair 952, USMS	KU127260	Lost⁴	Lost⁴
	<i>Burmannia bicolor</i> Mart.	Maas et al. 9649, U	KU127430	KU127431	KU127432
	<i>Burmannia biflora</i> L.	M.W. Chase 157, NCU	KU127413	–	KU127412
	<i>Burmannia capitata</i> (Walter ex J.F.Gmel) Mart.	Maas et al. 9606, U	KU127340	–	KU127339
	<i>Burmannia coelestis</i> D.Don	K.Cameron <i>s.n.</i> , NCU	KU127394	KU127396	KU127395
	<i>Burmannia disticha</i> L.	Wilkin 1017, K	KU127427	KU127428	KU127429
	‡ <i>Burmannia hexaptera</i> Schltr.	T. Franke K/9, M	KU127369	KU127368	KU127367
	‡ <i>Burmannia itoana</i> Makino	Kun-Ping Lo 821, PPI	KU127493	KU127492	Lost⁴
	<i>Burmannia longifolia</i> Becc.	Johns 9157, K	KU127309	KU127310	KU127311
	‡ <i>Burmannia lutescens</i> Becc.	L.R. Caddick 352, K	KU127287	KU127288	KU127289
	<i>Burmannia madagascariensis</i> Baker	L.R. Caddick 312, K	KU127419	KU127418	KU127417
	‡ <i>Burmannia oblonga</i> Ridl.	P. Wilkin 866, K	KU127271	KU127270	KU127269
	‡ <i>Burmannia sphagnoides</i> Becc.	L.R. Caddick 348, K	KU127485	–	–
	<i>Burmannia stuebelii</i> Hieron. & Schltr.	M. Weigend 98/420, <i>s.n.</i>	KU127302	KU127301	KU127300

Family	Taxon ¹	Voucher, herbarium	Plastid gene		
			<i>accD</i>	<i>clpP</i>	<i>matK</i>
	‡ <i>Burmannia wallichii</i> (Miers) Hook.f.	Zhang <i>s.n.</i> , K	KU127365	KU127366	– ²
	‡ <i>Campylosiphon congestus</i> (C.H.Wright) Maas	T. Franke K/8, M	KU127382	KU147410	KU127383
	‡ <i>Dictyostega orobanchoides</i> subsp. <i>parviflora</i> (Benth.) Snelders & Maas	H. van der Werff et al. 18384, MO	KU127397	KU127399	KU127398
	‡ <i>Gymnosiphon aphyllus</i> Blume	L.R. Caddick 353, K	KU127299	–	– ²
	‡ <i>Gymnosiphon longistylus</i> (Benth.) Hutch.	Merckx et al. 132, LV	KU127364	KU127363	Lost⁴
Dioscoreaceae	<i>Dioscorea bulbifera</i> L.	R. Olmstead 97-151, DNA	KU127386	KU127385	KU127384
	<i>Dioscorea communis</i> (L.) Caddick & Wilkin	M. Merello et al. 2285, MO	KU127283	KU127282	KU127281
	<i>Dioscorea polystachya</i> Turcz.	V. Lam VL014, UBC	KU127286	KU127285	KU127284
	<i>Dioscorea guianensis</i> R.Knuth	O. Tellez 13081, <i>s.n.</i>	KU127461	KU127460	KU127462
Nartheciacae	<i>Aletris farinosa</i> L.	M.W. Chase 105, NCU	KU127325	KU127326	KU127324
	<i>Aletris foliosa</i> (Maxim) Bureau & Franch	K. Cameron <i>s.n.</i> , <i>s.n.</i>	KU127438	KU127437	KU127436
	<i>Lophiola aurea</i> Ker Gawl.	Whitten 95028, K	KU127256	KU127255	KU127254
	<i>Narthecium ossifragum</i> (L.) Huds.	R.A. Stockey & G.A. Rothwell 59, ALTA	KU127316	KU127315	KU127317
	<i>Nietneria paniculata</i> Steyerm.	O. Hokche & P.J. Maas 849, U	KU127405	KU127403	KU127404
Taccaceae	<i>Tacca leontopetaloides</i> (L.) Kuntze	P. Wilkin 817, K	KU127477	KU127478	KU127476

Family	Taxon ¹	Voucher, herbarium	Plastid gene		
			<i>accD</i>	<i>clpP</i>	<i>matK</i>
Thismiaceae	‡ <i>Geomitra clavigera</i> Becc.	L.R. Caddick 354, K	KU127426	–	– ²
	‡ <i>Thismia aseroe</i> Becc.	L.R. Caddick 349, K	KU127459	– ²	– ²
	‡ <i>Thismia</i> sp. Griff.	P. Bygrave 53, K	KU127457	–	– ²
Trichopodaceae	<i>Trichopus sempervirens</i> (H.Perrier) Caddick & Wilkin	Wilkin et al. 948, K	KU127477	KU127478	KU127476
Liliales					
Alstroemeriaceae	<i>Alstroemeria aurea</i> Graham	M.J. Crawley MJC 157, <i>s.n.</i>	KU127329	KU127328	KU127327
	<i>Luzuriaga radicans</i> Ruiz & Pav.	M.W. Chase 499, K	KU127414	KU127416	KU127415
Campynemataceae	<i>Campynema lineare</i> Labill.	M.F. Duretto 1842, HO	KU127400	KU127401	KU127402
Corsiaceae	‡ <i>Arachnitis uniflora</i> Phil.	A.A. Cocucci leg. 2157, CORD	KU127350	–	– ²
	‡ <i>Arachnitis uniflora</i> Phil.	R. Neyland 1928, MCN	KP462884.1	KP462884.1	Lost⁵
	‡ <i>Corsia</i> cf. <i>boridiensis</i> P.Royen	S. Lyon SPL470-2, PNG	KP462885.1³	KP462885.1³	KP462885.1³
Liliaceae	<i>Lilium superbum</i> L.	M.W. Chase 112, NCU	KU127330	KU127331	KU127332
Ripogonaceae	<i>Ripogonum elseyanum</i> F.Muell	M.W. Chase 187 NCU	KU127377	KU127378	KU127376
Pandanales					
Cyclanthaceae	<i>Asplundia moritziana</i> (Klotzsch) Harling	M.W. Chase 1236, <i>s.n.</i>	KU127348	KU127349	KU127347
	<i>Carludovica palmata</i> Ruiz & Pav.	M.W. Chase 14836, K	KU127473	KU127474	KU127475
	<i>Cyclanthus bipartitus</i> Poit. ex A.Rich.	M.W. Chase 1237, K	KU127444	KU127443	KU127444

Family	Taxon ¹	Voucher, herbarium	Plastid gene		
			<i>accD</i>	<i>clpP</i>	<i>matK</i>
	<i>Thoracocarpus bissectus</i> (Vell.) Harling	A. Fuentes & F. Torrico 5447, MO	KU127251	KU127252	KU127253
Pandanaceae	<i>Freycinetia graminea</i> Blume	S. Graham SWG-02-03-39 UBC	KU127440	KU127439	KU127441
	<i>Pandanus copelandii</i> Merr.	C. Sherman & K. Bynum 303 MO	KU127359	KU127357	KU127358
	<i>Pandanus odorifer</i> (Forssk.) Kuntze	M.W. Chase 19215, K	KU127467	KU127469	KU127468
Stemonaceae	<i>Croomia japonica</i> Miq.	Rothwell & Stockey 43, ALTA	KU127371	KU127372	KU127370
	<i>Pentastemona sumatrana</i> Steenis B.G.	Leiden 910375, KK	KU127304	KU127305	KU127303
	<i>Stemona tuberosa</i> Lour.	Rothwell & Stockey 46, ALTA	KU127491	KU127490	KU127489
	<i>Stichoneuron caudatum</i> Ridl.	Rothwell & Stockey 45, ALTA	KU127341	KU127342	KU127343
Triuridaceae	‡ <i>Sciaphila densiflora</i> Schltr.	Duangjai 029, BRUN	KU127466	–	–
	‡ <i>Sciaphila densiflora</i> Schltr.	Y. Pillon et al. 88, NOU, P	KU127274	KU127273	KU127272
	‡ <i>Sciaphila ledermannii</i> Engl.	T. Franke et al. #4/01, FPA	KU127264	KU127265	– ²
	‡ <i>Sciaphila</i> sp. Blume	J. Dransfield 7345, K	KU127458	–	–
Velloziaceae	<i>Acanthochlamys bracteata</i> P.C.Kao	P.C. Kao 1993, K	KU127336	KU127337	KU127338
	<i>Vellozia</i> sp. Vand.	Kubitzki & Feuerer 97-30, HBG	KU127335	KU127334	KU127333
	<i>Xerophyta [Talbotia] elegans</i> Balf.	Rothwell & Stockey 48, ALTA	KU127354	KU127355	KU127356
	<i>Xerophyta retinervis</i> Baker	G. Reeves 14, NBG	KU127268	KU127266	KU127267

Family	Taxon ¹	Voucher, herbarium	Plastid gene		
			<i>accD</i>	<i>clpP</i>	<i>matK</i>
Petrosaviales					
Petrosaviaceae	<i>Japonolirion osense</i> Nakai	M.W. Chase 3000, K	KU127408	KU127407	KU127406
	‡ <i>Petrosavia</i> aff. <i>sakurarii</i> (Makino) J.JSm. ex Steenis	Yukawa 09-25, TNS	KU127261	KU127262	KU127263
	‡ <i>Petrosavia sakurarii</i> (Makino) J.JSm. ex Steenis	Yukawa 09-47, TNS	KU127451	KU127452	KU127453
	‡ <i>Petrosavia stellaris</i> Becc.	Cameron 2154, NY	KU127307	KU127308	KU127306
	‡ <i>Petrosavia stellaris</i> Becc.	see Logacheva et al. (2014)	KF482381.1³	KF482381.1³	KF482381.1³
Poales					
Bromeliaceae	<i>Ananas comosus</i> (L.) Merr.	H.S. Rai 1003, ALTA	KU127483	KU127482	KU127484
Strelitziaceae	<i>Strelitzia reginae</i> Banks	H. O'Brien <i>s.n.</i> , ALTA	KU127434	KU127435	KU127433
Outgroups					
Magnoliaceae	<i>Magnolia grandiflora</i> L.	V.Lam & B. Zhuang VL025, UBC	KU127454	KU127456	KU127455
Cornaceae	<i>Davidia involucrata</i> Baill.	V.Lam & B. Zhuang VL024, UBC	KU127470	KU127471	KU127472

¹Previously published sequences presented or referenced in Barrett et al. (2013), except for *Amborella trichopoda* Baill. (Amborellaceae) NC_005086.1; *Buxus microphylla* Siebold & Zucc (Buxaceae) NC_009599.1; *Calycanthus floridus* var *glaucus* (Willd.) Torr. & A.Gray (Calycanthaceae) NC_004993.1; *Chloranthus spicatus* (Thunb.) Makino (Chloranthaceae) NC_009598.1; *Drimys grandensis* L.f. (Winteraceae) NC_008456.1; *Illicium oligandrum* Merr. & Chun (Schisandraceae) NC_009600.1; *Liriodendron tulipifera* L. (Magnoliaceae) NC_008326.1; *Nymphaea alba* L. (Nymphaeaceae) NC_006050.1; *Piper cenocladum* C.DC. (Piperaceae) DQ887677.1

² Possible contaminant identified for this gene (excluded from analysis).

³ GenBank sequences from previously published studies.

⁴ Gene lost based on unpublished whole plastomes (Chapter 4).

⁵ Gene lost based on published whole plastome data of *Arachnitis uniflora* (Mennes et al. 2015).

Additional references:

Thiers, B., continuously updated. Index Herbariorum: A global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium. <http://sweetgum.nybg.org/ih/>

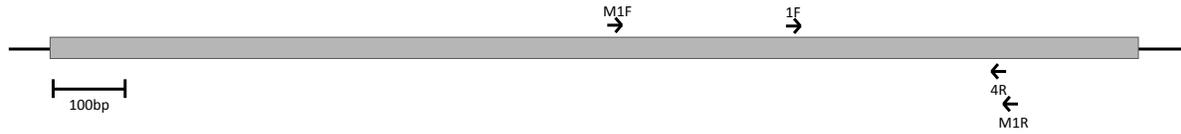
Table A.2 Data partitioning schemes inferred using PartitionFinder with the BIC criterion (see text for details; analyses c-m involve photosynthetic angiosperms, including several orchids, plus the mycoheterotrophic taxon noted). (a) all taxa included, (b) ‘green’ taxa only, i.e., only photosynthetic angiosperms (green orchids included, green Burmanniaceae excluded), (c) Burmanniaceae (green and fully mycoheterotrophic), (d) Corsiaceae, (e) Iridaceae, (f) Orchidaceae, (g) Petrosaviaceae, (h) Thismiaceae, (i) Triuridaceae, (j) green Burmanniaceae only, (k) *Geomitra clavigera* (Thismiaceae), (l) *Thismia aseroe* (Thismiaceae), (m) *Thismia* sp. Bygrave 53 (Thismiaceae). Plastid genes are indicated before the first underscore; ‘pos’ = codon position, ‘exon’ = exon number, and ‘intron’ = intron number.

Partition no.	Best model	Partition subsets
a)		
1	GTR + G + I	accD_pos1, clpP_exon2_pos3, clpP_exon3_pos1
2	GTR + G + I	accD_pos2, clpP_exon2_pos1, clpP_exon3_pos2
3	GTR + G + I	accD_pos3, matK_pos3
4	GTR + G	clpP_exon2_pos2, clpP_exon3_pos3, clpP_intron1, clpP_intron2, matK_pos1, matK_pos2
b)		
1	GTR + G	accD_pos1, clpP_exon2_pos1, clpP_exon3_pos1
2	GTR + G	accD_pos2, clpP_exon2_pos2, clpP_exon3_pos2
3	GTR + G	accD_pos3, matK_pos3
4	GTR + G	clpP_exon2_pos3, clpP_exon3_pos3, clpP_intron1, clpP_intron2, matK_pos1, matK_pos2
c)		
1	GTR + G	accD_pos1, clpP_exon2_pos1, clpP_exon3_pos1
2	GTR + G	accD_pos2, clpP_exon2_pos2, clpP_exon3_pos2
3	GTR + G + I	accD_pos3, matK_pos3
4	GTR + G	clpP_exon2_pos3, clpP_exon3_pos3; clpP_intron1, clpP_intron2, matK_pos1, matK_pos2
d)		
1	GTR + G	accD_pos1, clpP_exon2_pos1, clpP_exon3_pos1
2	GTR + G	accD_pos2, clpP_exon2_pos2, clpP_exon3_pos2
3	GTR + G	accD_pos3, matK_pos3
4	GTR + G	clpP_exon2_pos3, clpP_exon3_pos3, clpP_intron1, clpP_intron2, matK_pos1, matK_pos2
e)		
1	GTR + G	accD_pos1, clpP_exon2_pos1, clpP_exon3_pos1
2	GTR + G	accD_pos2, clpP_exon2_pos2, clpP_exon3_pos2
3	GTR + G	accD_pos3, matK_pos3
4	GTR + G	clpP_exon2_pos3, clpP_exon3_pos3, clpP_intron1, clpP_intron2, matK_pos1, matK_pos2
f)		
1	GTR + G	accD_pos1, clpP_exon2_pos1, clpP_exon3_pos1

Partition no.	Best model	Partition subsets
2	GTR + G	accD_pos2, clpP_exon2_pos2, clpP_exon3_pos2
3	GTR + G	accD_pos3, clpP_exon2_pos3, matK_pos3
4	GTR + G	clpP_exon3_pos3, clpP_intron1, clpP_intron2, matK_pos1, matK_pos2
g)		
1	GTR + G	accD_pos1, clpP_exon2_pos1, clpP_exon3_pos1
2	GTR + G	accD_pos2, clpP_exon2_pos2, clpP_exon3_pos2
3	GTR + G	accD_pos3, matK_pos3
4	GTR + G	clpP_exon2_pos3, clpP_exon3_pos3, clpP_intron1, clpP_intron2 matK_pos1, matK_pos2
h)		
1	GTR + G	accD_pos1, clpP_exon2_pos1, clpP_exon3_pos1
2	GTR + G	accD_pos2, clpP_exon2_pos2, clpP_exon3_pos2
3	GTR + G	accD_pos3, matK_pos3
4	GTR + G	clpP_exon2_pos3, clpP_exon3_pos3, clpP_intron1, clpP_intron, matK_pos1 matK_pos2
i)		
1	GTR + G	accD_pos1, clpP_exon2_pos1, clpP_exon3_pos1
2	GTR + G	accD_pos2, clpP_exon2_pos2, clpP_exon3_pos2
3	GTR + G	accD_pos3, clpP_exon2_pos3, matK_pos3
4	GTR + G	clpP_exon3_pos3, clpP_intron1, clpP_intron2, matK_pos1, matK_pos2
j)		
1	GTR + G	accD_pos1, clpP_exon2_pos1, clpP_exon3_pos1
2	GTR + G	accD_pos2, clpP_exon2_pos2, clpP_exon3_pos2
3	GTR + G	accD_pos3, clpP_exon2_pos3, matK_pos3
4	GTR + G	clpP_exon3_pos3, clpP_intron1, clpP_intron2, matK_pos1, matK_pos2
k)		
1	GTR + G	accD_pos1, clpP_exon2_pos1, clpP_exon3_pos1
2	GTR + G	accD_pos2, clpP_exon2_pos2, clpP_exon3_pos2
3	GTR + G	accD_pos3, matK_pos3
4	GTR + G	clpP_exon2_pos3, clpP_exon3_pos3, clpP_intron1, clpP_intron2 matK_pos1, matK_pos2
l)		
1	GTR + G	accD_pos1, clpP_exon2_pos1, clpP_exon3_pos1
2	GTR + G	accD_pos2, clpP_exon2_pos2, clpP_exon3_pos2
3	GTR + G	accD_pos3, matK_pos3
4	GTR + G	clpP_exon2_pos3, clpP_exon3_pos3, clpP_intron1, clpP_intron2 matK_pos1, matK_pos2
m)		
1	GTR + G	accD_pos1, clpP_exon2_pos1, clpP_exon3_pos1
2	GTR + G	accD_pos2, clpP_exon2_pos2, clpP_exon3_pos2
3	GTR + G	accD_pos3, matK_pos3
4	GTR + G	clpP_exon2_pos3, clpP_exon3_pos3, clpP_intron1, clpP_intron2 matK_pos1, matK_pos2

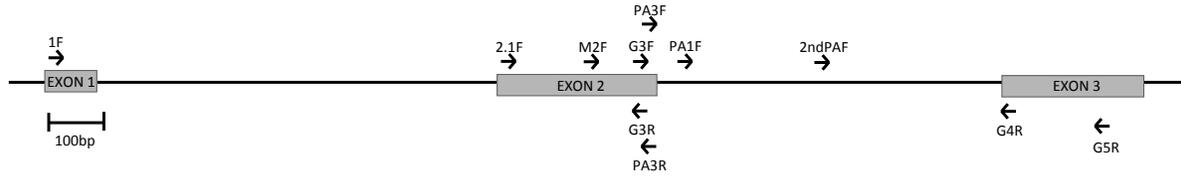
Figure A.1 Primer maps for *accD*, *clpP* and *matK*. Primers in bold are those typically used for amplification; grey box = coding region; black line = intron. ‘M’ indicates primers intended for use in monocots, ‘G’ those designed for use with angiosperms in general, ‘PA’ indicates those designed to deal with taxon-specific regions with homopolymer repeats. All primers are newly designed except for 1F and 4R (*accD*; Newmaster et al. 2007), and 2.1F, 5R, 3F_KIM, 1R_KIM, R(Equisetum) (*matK*; Fazekas et al. 2008; Ki-Joong Kim, unpublished). Primers are not drawn to scale.

a) *accD*



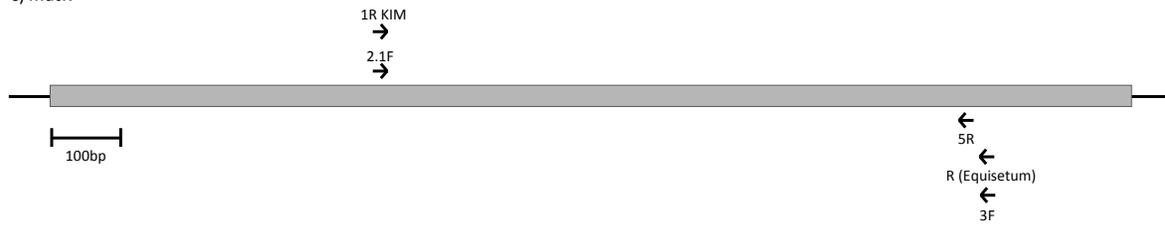
Forward primers		Reverse primers	
M1F:	ATGAGTAGTTCAGATAG	M1R:	GGTAAAAGAGTAATTGAACAAACA
1F:	AGTATGGGATCCGTAGTAGG	4R:	ATTGCATTTGCGGGTAAAAGA

b) *clpP*



Forward primers		Reverse primers	
1F:	ATGCCYRTTGGTGTTC	G3R:	GGAGGAGAAATTACCAARCG
2.1F:	CGAGAAAGATTACTTTRGG	PA3R:	CGTCTAGCATTCCCTCAC
M2F:	GCCATTTATGATACTATGC	G4R:	GGRTTATGRTYCACCAACCTGCT
G3F:	GGAGGAGAAATTACCAARCG	G5R:	CAGCAACAGAAGCCCAAG
PA3F:	CGTCTAGCATTCCCTCAC		
PA1F:	GAAGACTATGCCTTCGCCATATCG		
2ndPAF:	GAAACTTTGGGATTGC		

c) *matK*



Forward primers		Reverse primers	
1R_KIM:	ACCCAGTCCATCTGGAATCTTGGTTC	3F_KIM:	CTCGTAAACACAAAAGTACTGTACG
2.1F:	CCTATCCATCTGGAATCTTAG	5R:	CGACTTTTCTGTGCTAGAAC
		R (Equisetum):	GCTCGTAAACATAAAAAGTAC

Figure A.2 Phylogenetic placement of a putative contaminant *matK* sequence of *Thismia aseroe* (Thismiaceae; shown in bold) based on unpartitioned maximum likelihood analysis of this locus. The scale bar indicates the estimated number of substitutions per site.

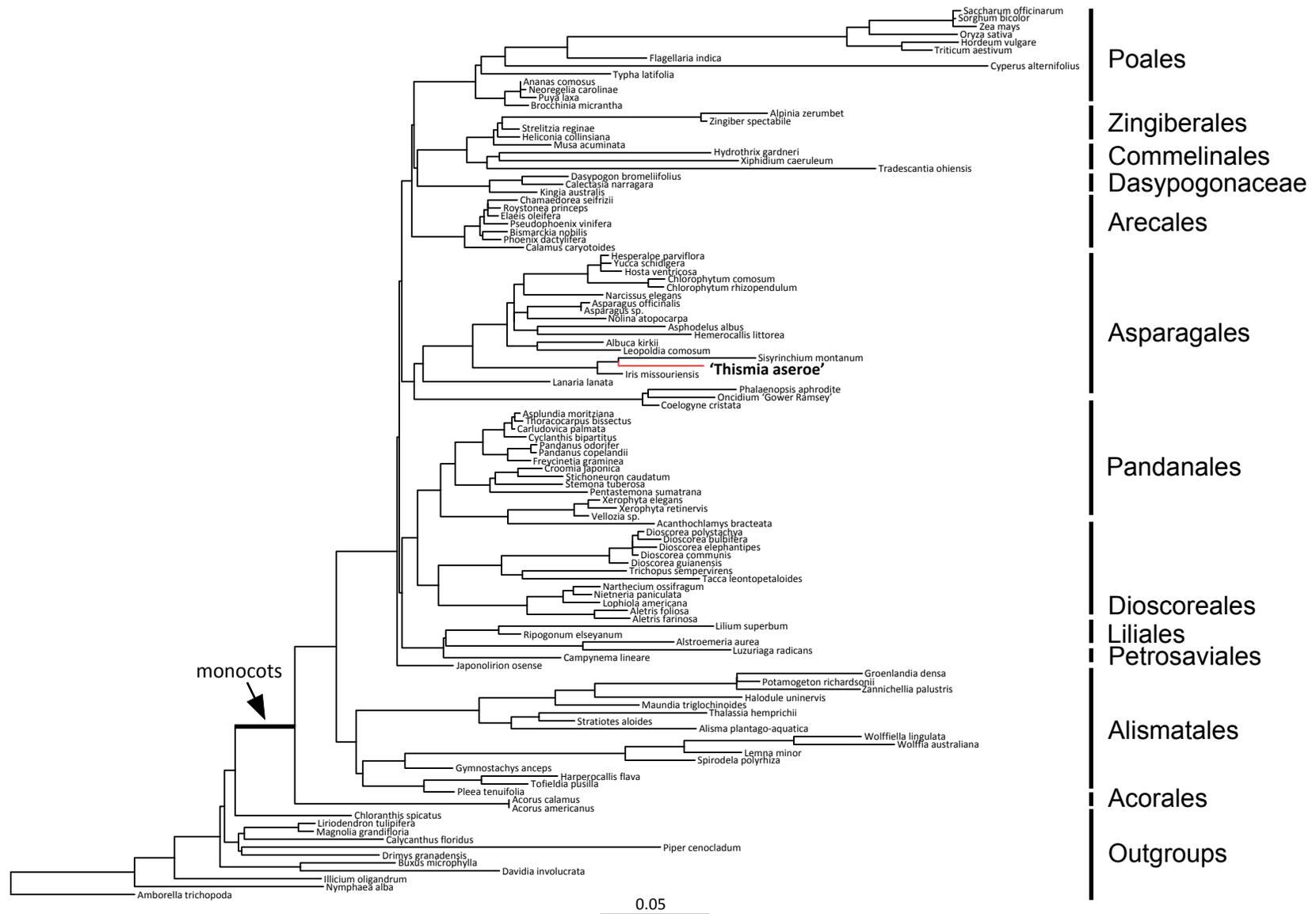


Figure A.3 Phylogenetic placement of a putative contaminant *matK* sequence of *Geomitra clavigera* (Thismiaceae; shown in bold) based on unpartitioned maximum likelihood analysis of this locus. The scale bar indicates the estimated number of substitutions per site.



Poales

Zingiberales

Commelinales

Dasygogonaceae

Arecales

Asparagales

Pandanales

Dioscoreales

Liliales

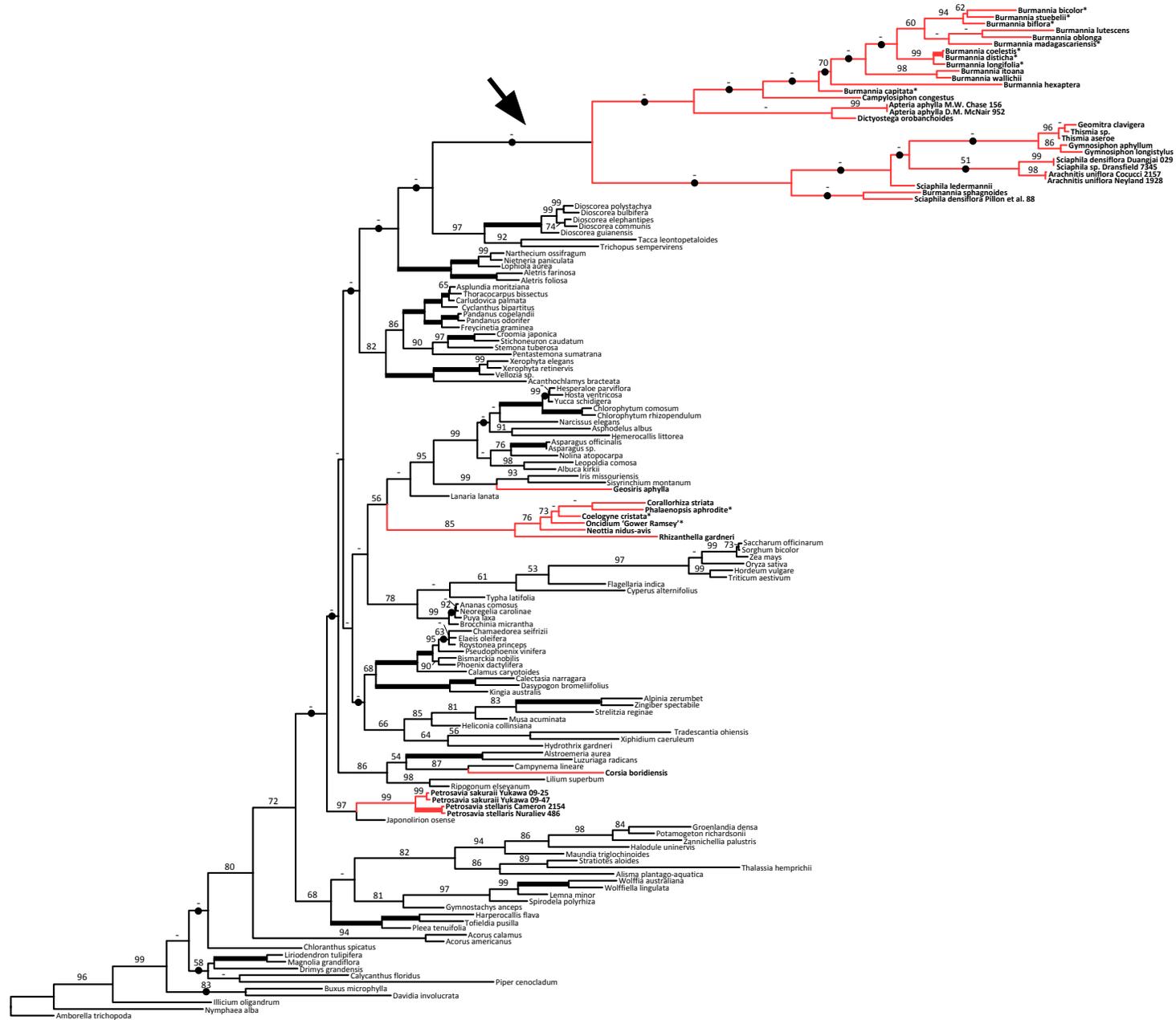
Petrosaviales

Alismatales

Acorales

Outgroups

Figure A.4 Phylogenetic inference of photosynthetic and fully mycoheterotrophic monocots (all lineages) based on parsimony analysis of a concatenated three-gene matrix (*accD*, *clpP* and *matK*). This is one of 324 most parsimonious trees (11,942 steps), shown as a phylogram. The large arrow points to a ‘fast clade’ of rapidly evolving lineages that may result from long-branch attraction. Lineages with mycoheterotrophs are indicated in red (asterisks indicate photosynthetic taxa in Burmanniaceae and Orchidaceae, the remainder are full mycoheterotrophs). Branches with 100% bootstrap support are shown as thick lines; bootstrap support values are otherwise indicated beside branches (<50% support is indicated with a dash, '-'). The scale bar indicates estimated number of changes.



100 changes

Figure A.5 Phylogeny of photosynthetic and fully mycoheterotrophic monocots (all lineages) based on an unpartitioned likelihood analysis of a concatenated three-gene matrix (*accD*, *clpP* and *matK*). Lineages with mycoheterotrophs are indicated in red (asterisks indicate photosynthetic taxa in Burmanniaceae and Orchidaceae, the remainder are full mycoheterotrophs). Branches with 100% bootstrap support are shown as thick lines; bootstrap support values are otherwise indicated beside branches (<50% support is indicated with a dash, '-'). The scale bar indicates estimated number of substitutions per site.

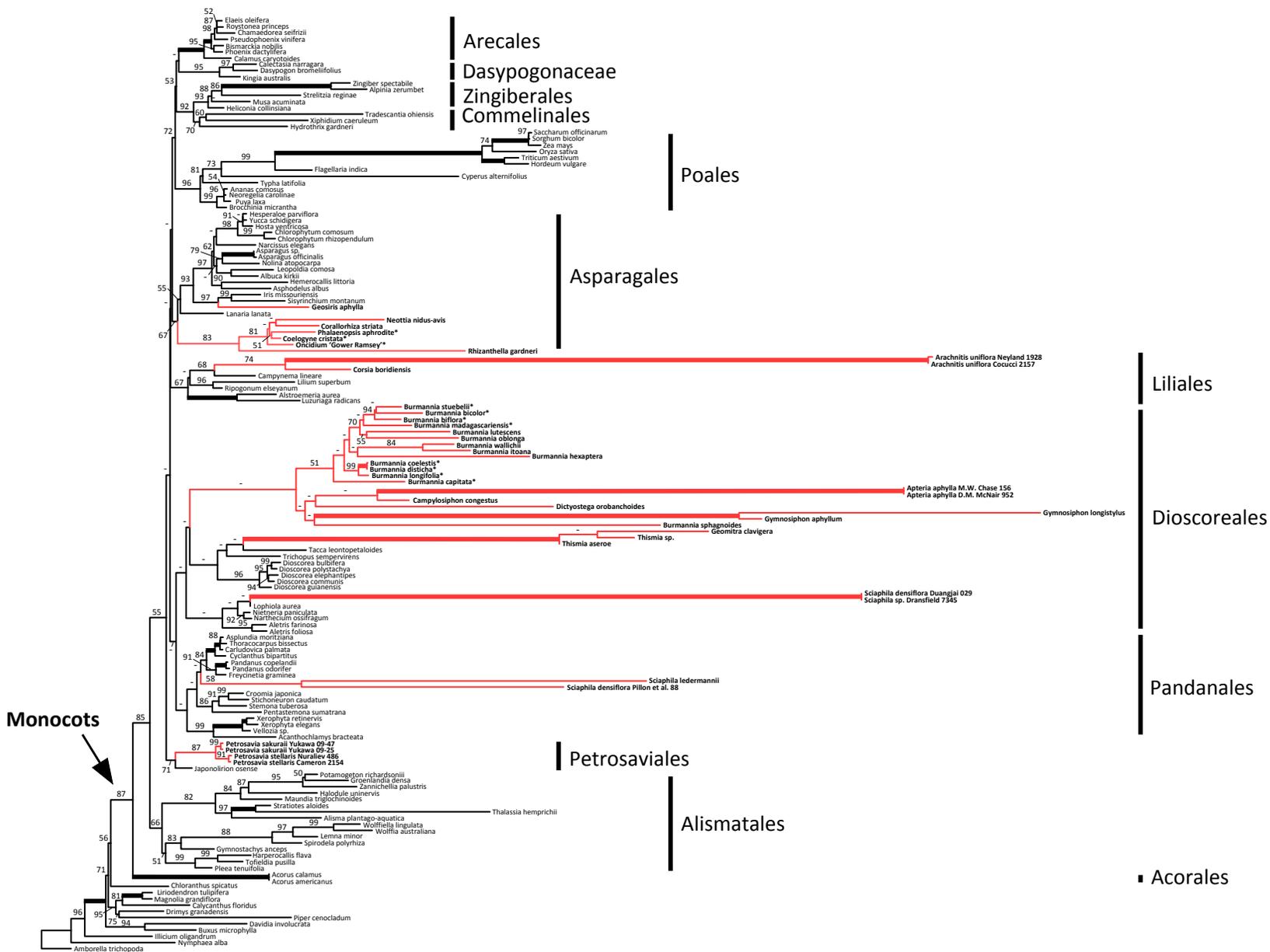


Figure A.6 Phylogenetic placement of *Geosiris* based on partitioned ML analysis of three plastid genes (*accD*, *clpP* and *matK*). *Geosiris* is shown in red. The scale bar indicates the estimated number of substitutions per site.

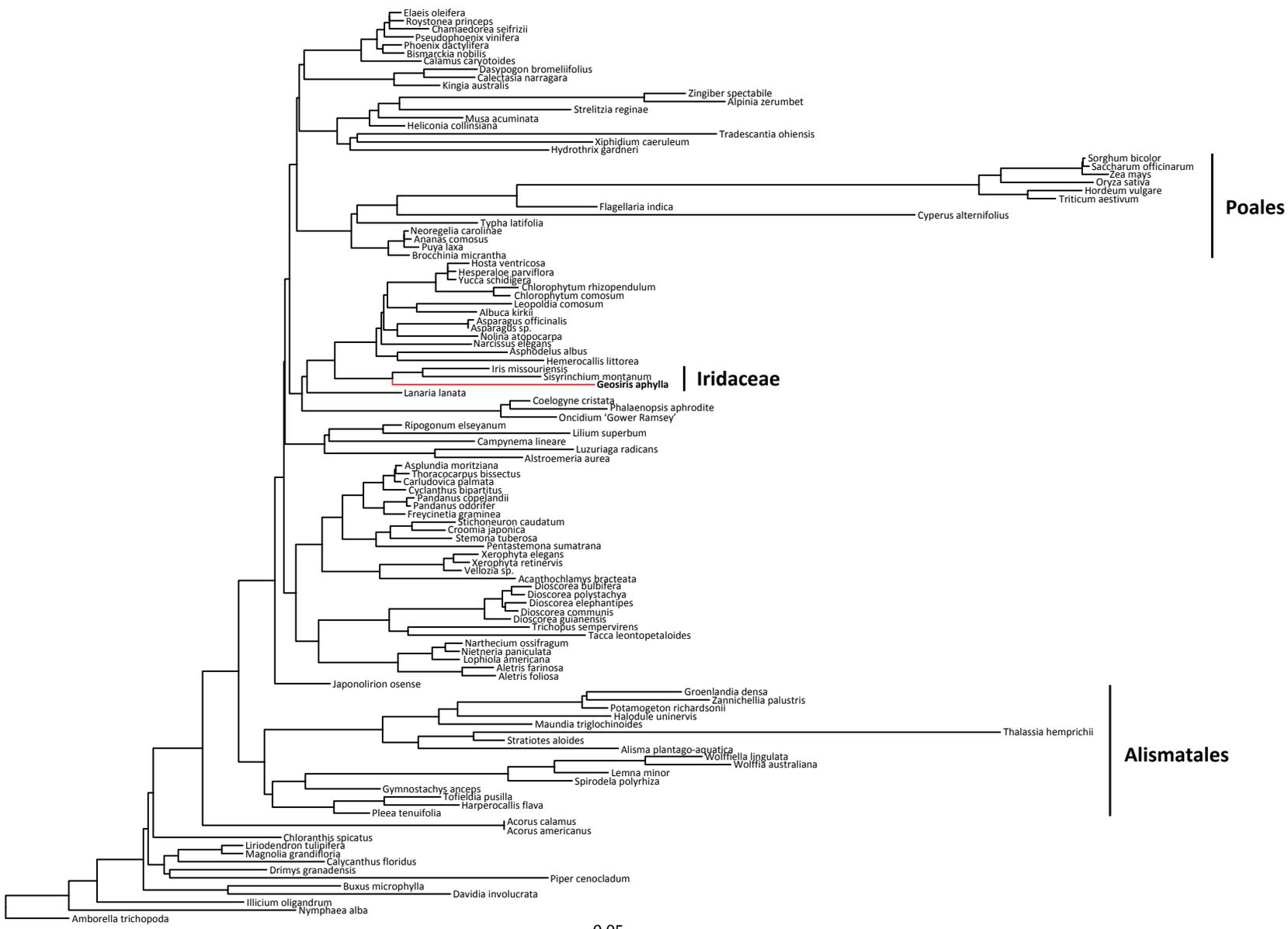


Figure A.7 Phylogenetic placement of mycoheterotrophic Orchidaceae based on partitioned ML analysis of three plastid genes (*accD*, *clpP* and *matK*). Orchidaceae are shown in red; photosynthetic species are indicated with an asterisk (*), the remaining species are fully mycoheterotrophs. The scale bar indicates the estimated number of substitutions per site.

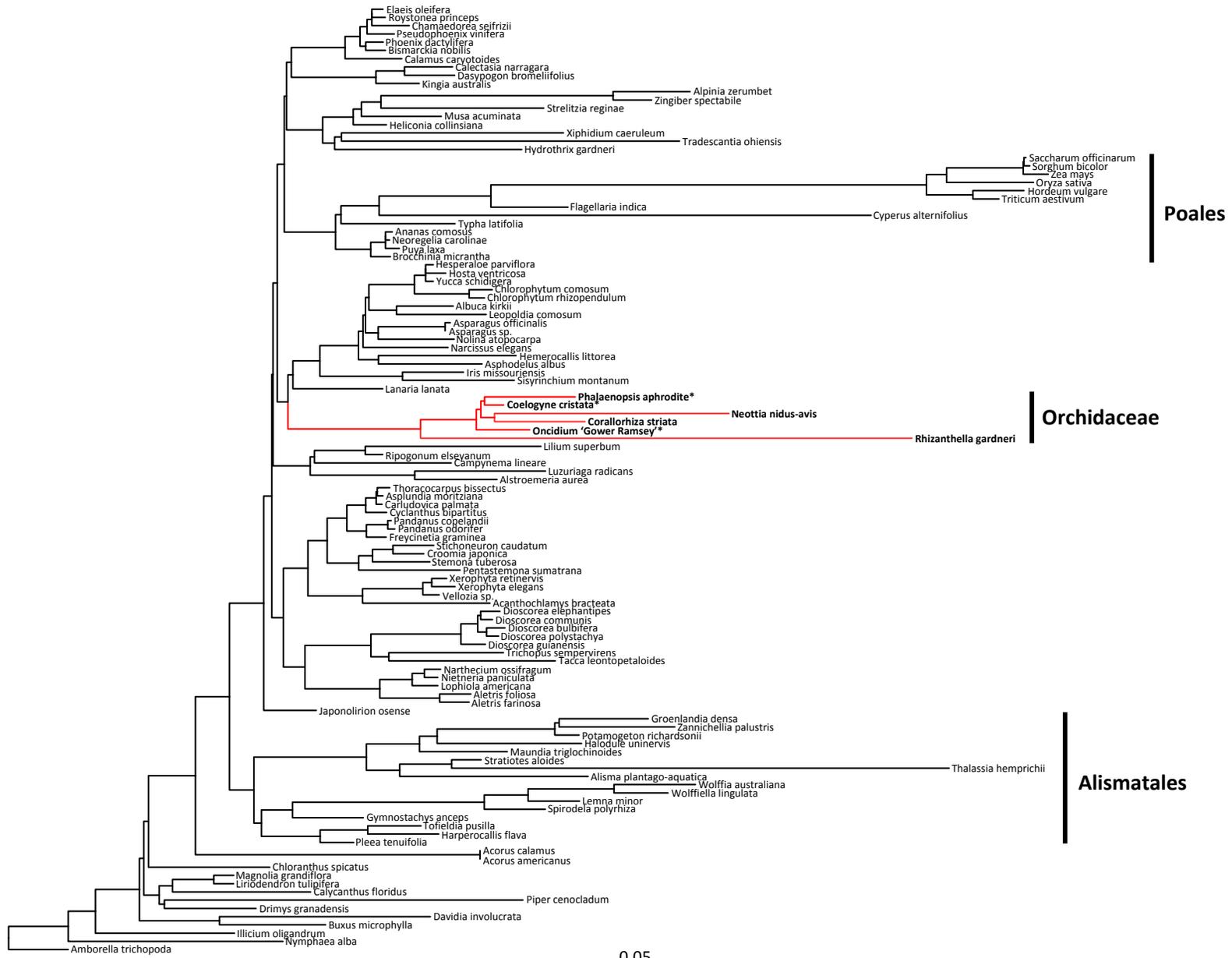


Figure A.8 Phylogenetic placement of Burmanniaceae based on partitioned ML analysis of three plastid genes (*accD*, *clpP* and *matK*). Burmanniaceae are shown in red; photosynthetic species (following Merckx et al. 2006) are indicated with an asterisk (*), the remaining species are fully mycoheterotrophs. The scale bar indicates the estimated number of substitutions per site.

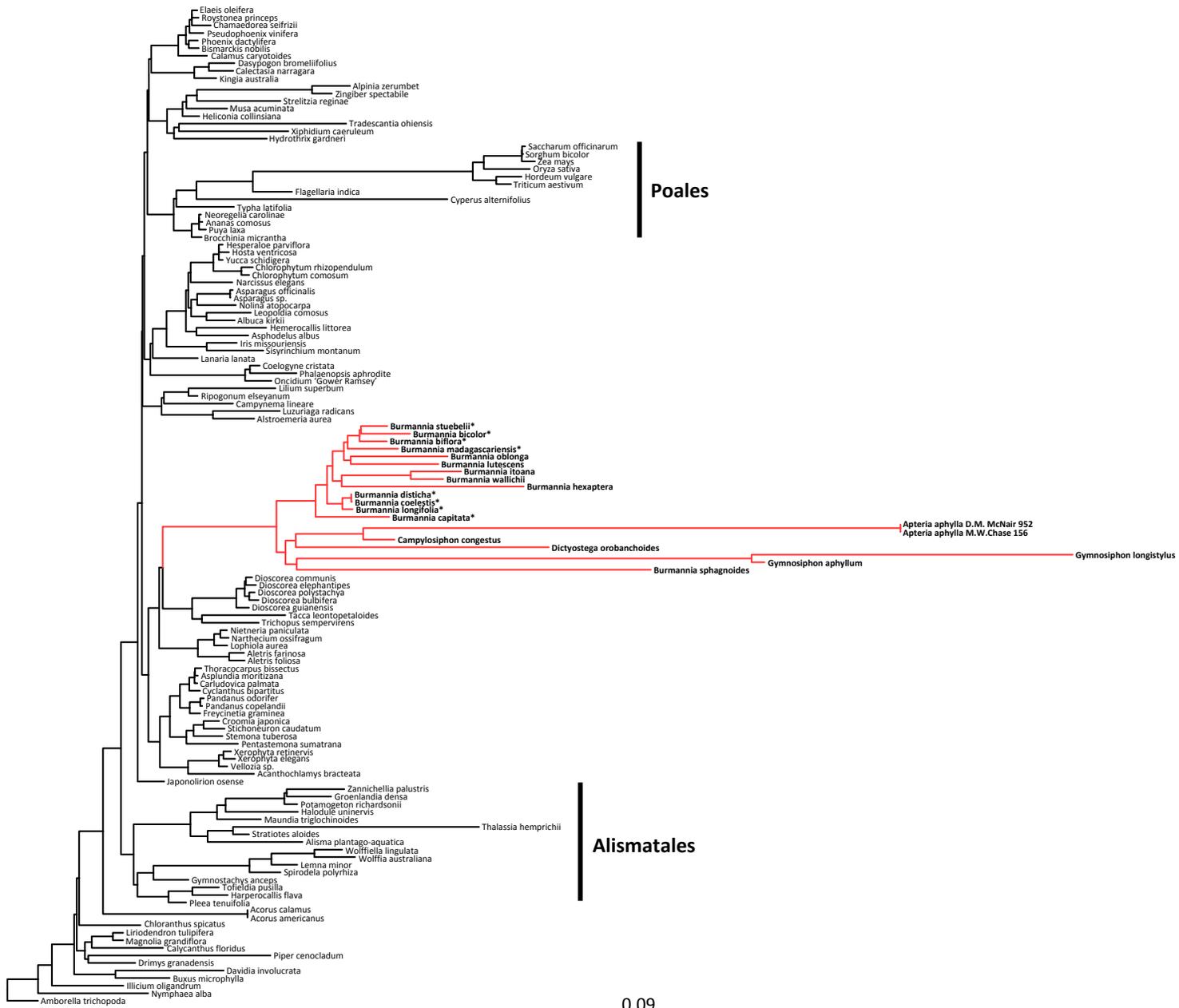


Figure A.9 Phylogenetic placement of Corsiaceae based on partitioned ML analysis of three plastid genes (*accD*, *clpP* and *matK*). Corsiaceae shown in red. The scale bar indicates the estimated number of substitutions per site.

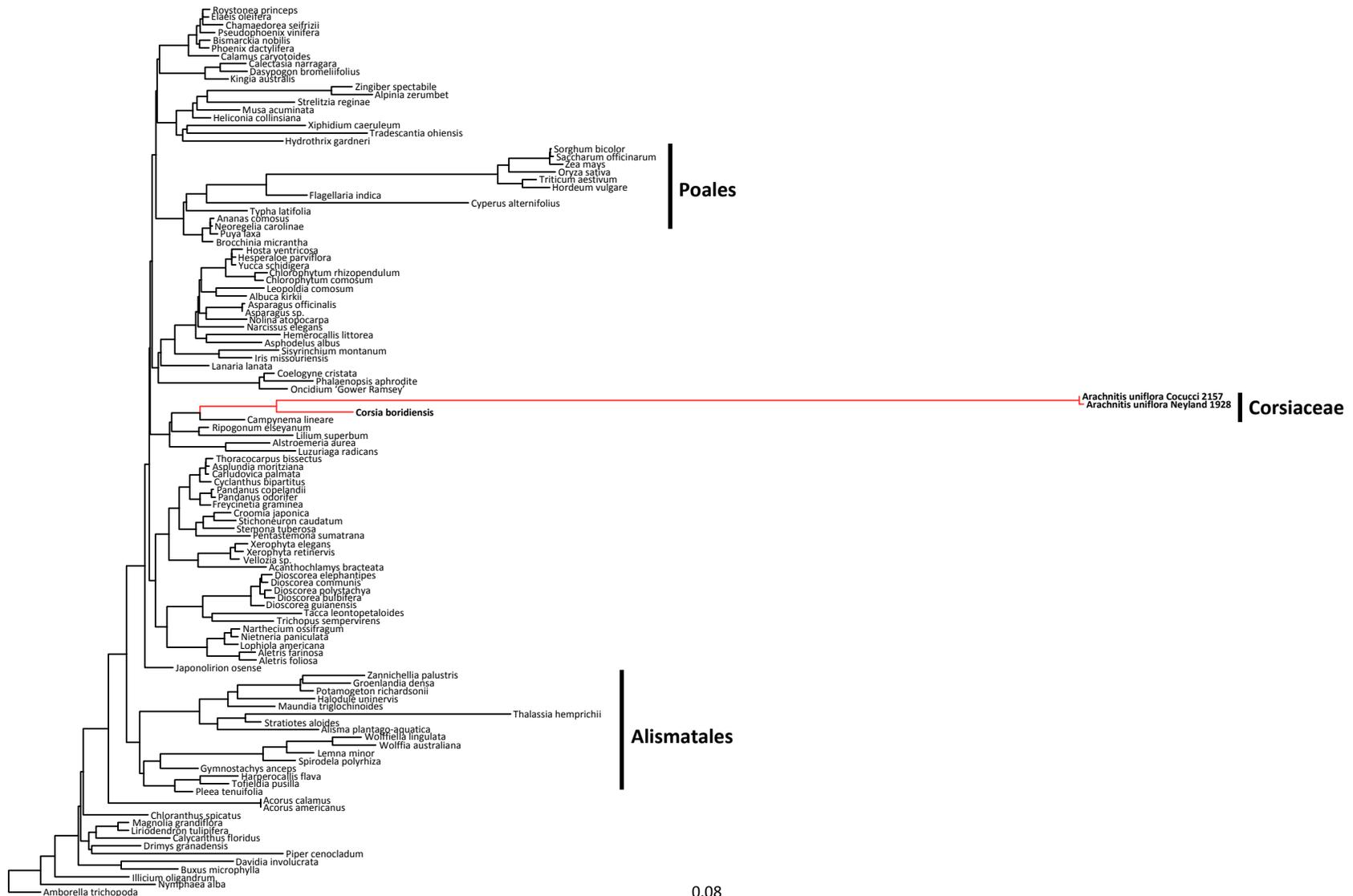


Figure A.10 Phylogenetic placement of Petrosaviaceae based on partitioned ML analysis of three plastid genes (*accD*, *clpP* and *matK*). Species of *Petrosavia* are shown in red. The scale bar indicates the estimated number of substitutions per site.

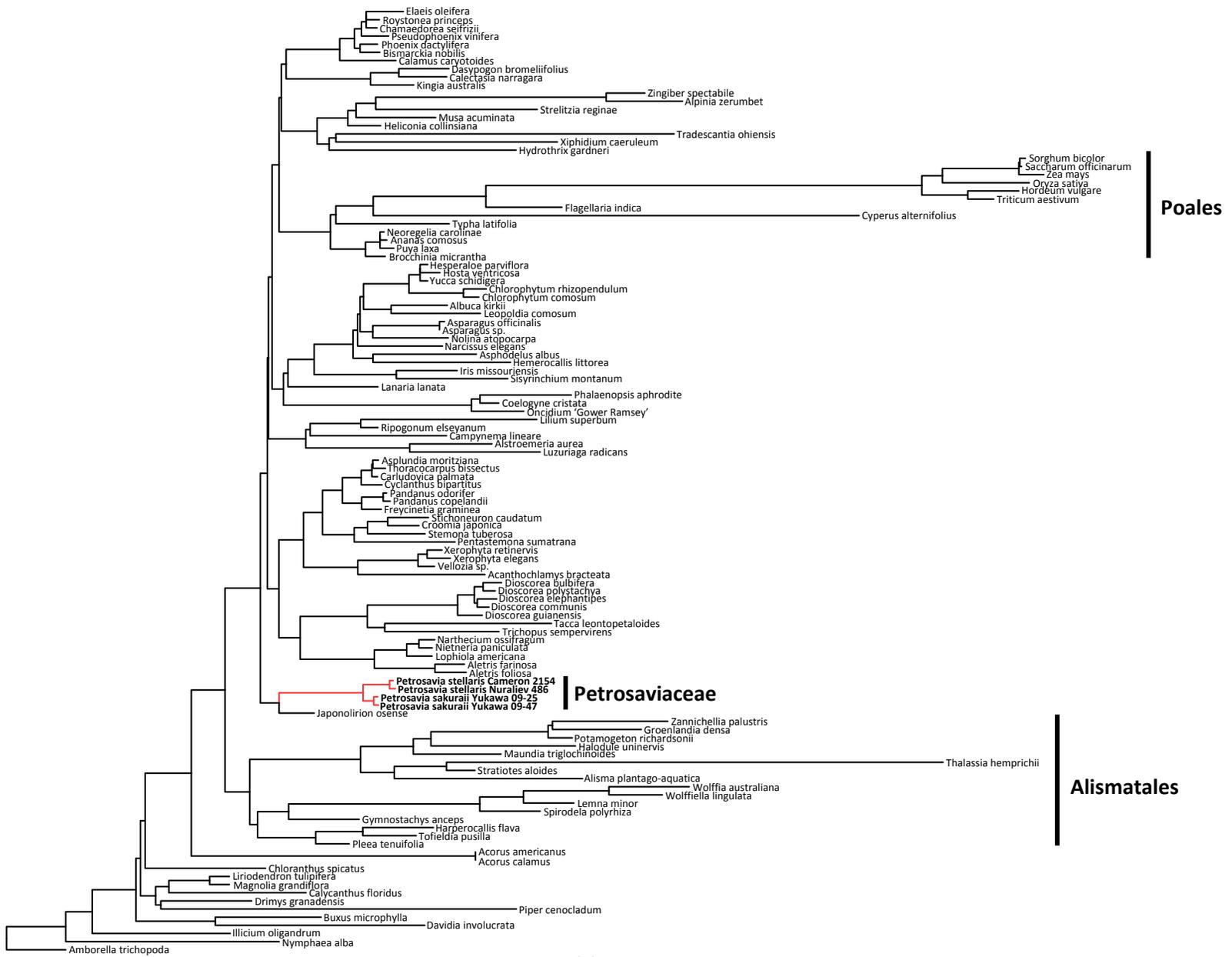


Figure A.11 Phylogenetic placement of Triuridaceae based on partitioned ML analysis of three plastid genes (*accD*, *clpP* and *matK*). Triuridaceae shown in red. The scale bar indicates the estimated number of substitutions per site.

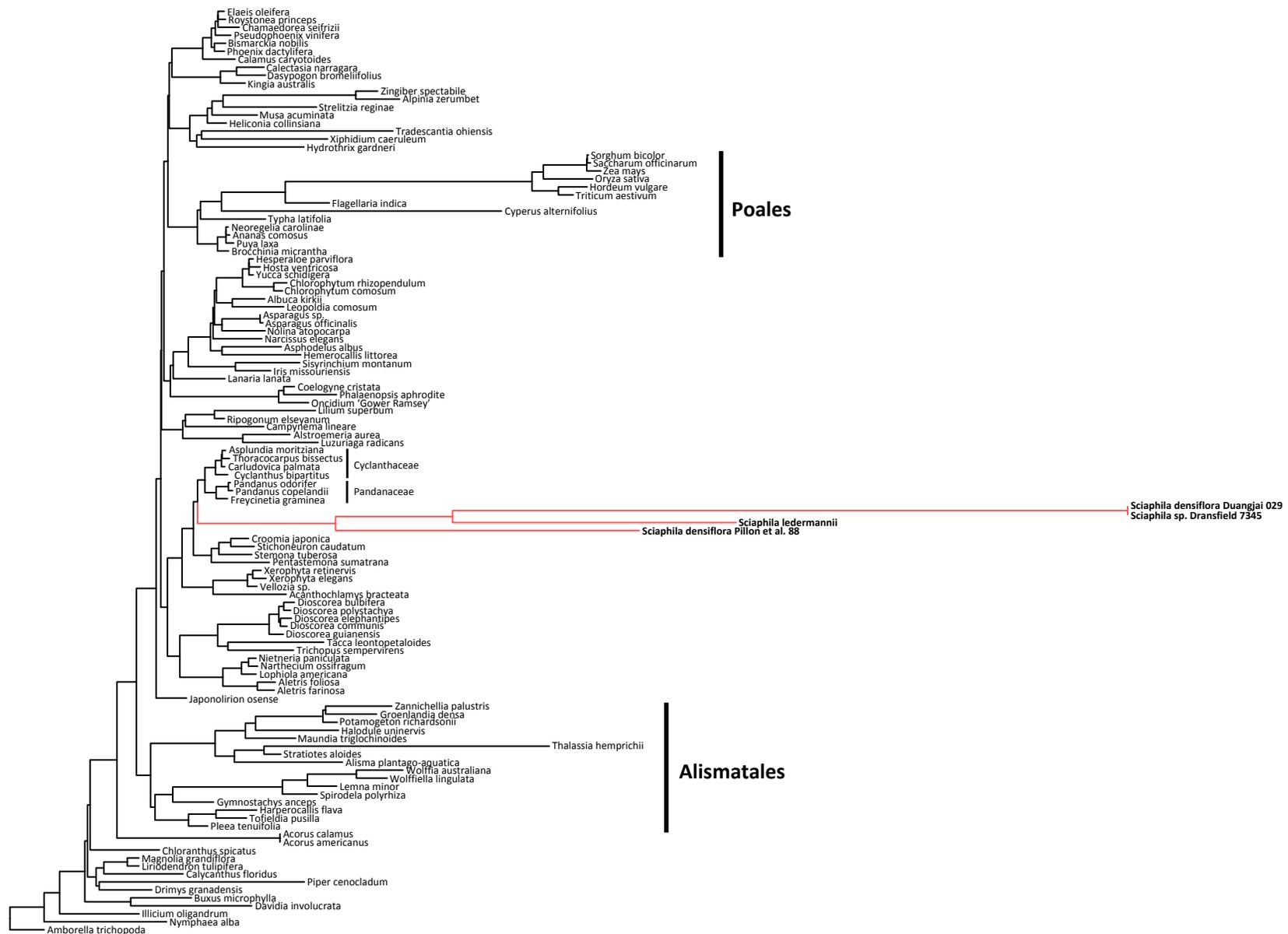


Figure A.12 Phylogenetic placement of Triuridaceae based on unpartitioned ML analysis of three plastid genes (*accD*, *clpP* and *matK*). Triuridaceae shown in red. The scale bar indicates the estimated number of substitutions per site.



0.07

Figure A.13 Phylogenetic placement of Thismiaceae based on partitioned ML analysis of three plastid genes (*accD*, *clpP* and *matK*), with Thismiaceae represented only by *accD*. Thismiaceae shown in red. The scale bar indicates the estimated number of substitutions per site.

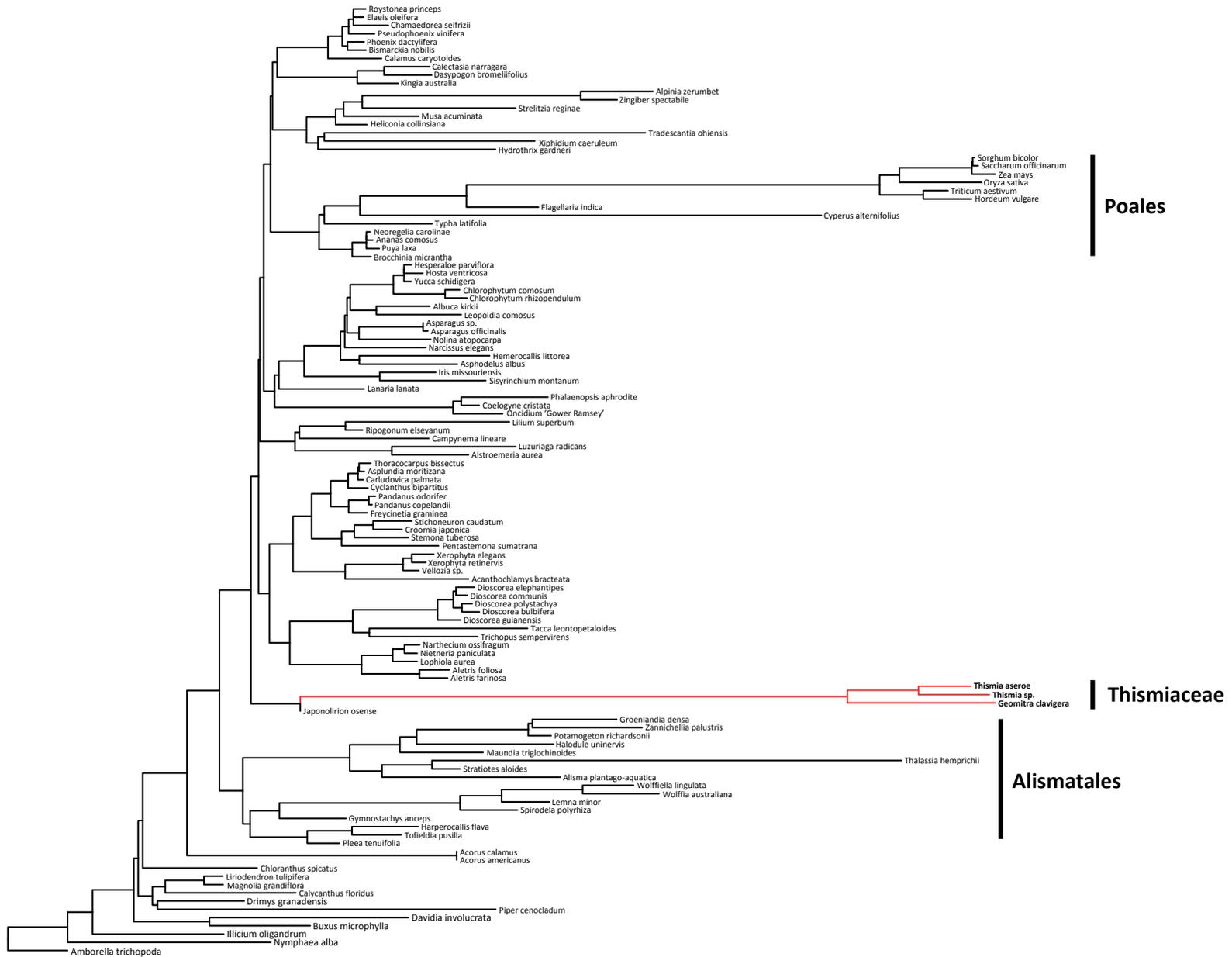
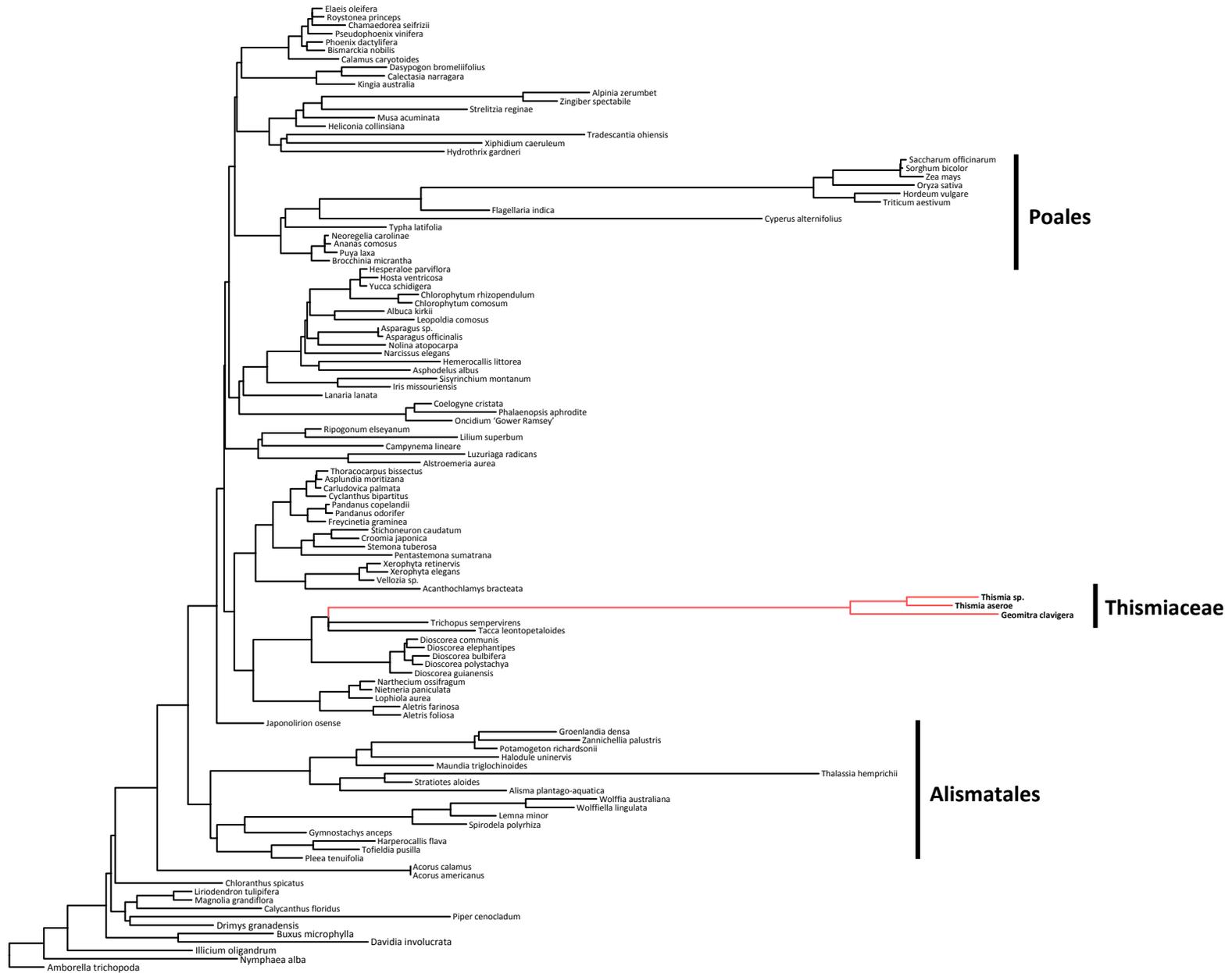
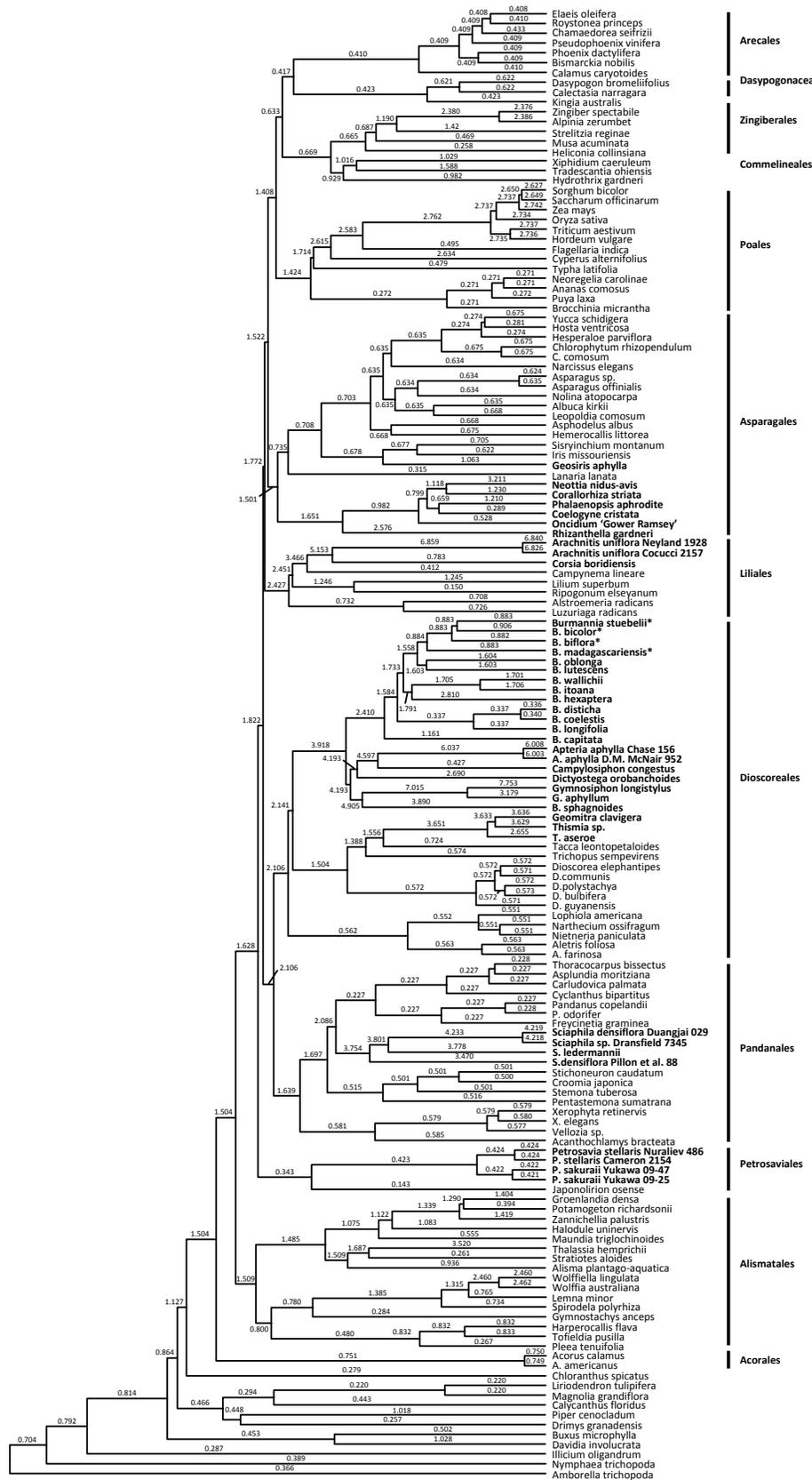


Figure A.14 Phylogenetic placement of Thismiaceae based on unpartitioned ML analysis of three plastid genes (*accD*, *clpP* and *matK*), with Thismiaceae represented only by *accD*. Thismiaceae shown in red. The scale bar indicates the estimated number of substitutions per site.



0.05

Figure A.15 Relative substitution rates among green and fully mycoheterotrophic monocot lineages based on a Bayesian analysis of a three-gene plastid (*accD*, *clpP* and *matK*) dataset, under a random-local-clock model, with a constrained topology for all lineages (see text for details). Numbers above branches indicate the relative rates (substitutions per site) for the specified branch.



Appendix B: Supplementary tables and figures for Chapter 3

Table B.1: Specimen source information; herbarium abbreviations follow Thiers (continuously updated).

Species ¹	Family	Voucher number [Collector number (herbarium)]	GenBank accession
Pandanales			
<i>Carludovica palmata</i> Ruiz & Pav.	Cyclanthaceae	M.W. Chase 14836, K	KP462882.1
<i>Cyclanthus bipartitus</i> Poit. ex A. Rich	Cyclanthaceae	M.W. Chase 1237, K	KT205192 - KT205273
<i>Freycinetia banksii</i> A. Cunn	Pandanaceae	S.W. Graham 02-03-14, UBC	KT205110 - KT205191
<i>Saranga sinuosa</i> Hemsl.	Pandanaceae	Gallaher 461, BISH, HAW	KT204539 - KT204619
<i>Croomia japonica</i> Miq.	Stemonaceae	Rothwell & Stockey 43, ALTA	KT204620 - KT204701
<i>Stemona tuberosa</i> Lour.	Stemonaceae	Rothwell & Stockey 46, ALTA	KT204702 - KT204783
<i>Stichoneuron caudatum</i> Ridl.	Stemonaceae	Rothwell & Stockey 45, ALTA	KT204946 - KT205027
<i>Pentastemona sumatrana</i> Steenis	Stemonaceae	B.G. Leiden 910375, K	KT205028 - KT205109
<i>Sciaphila densiflora</i> Schltr.	Triuridaceae	Pillon Y. <i>et al.</i> 88, NOU, P	KR902497.1
<i>Xerophyta retinervis</i> Baker	Velloziaceae	B.G. Reeves 14, NBG	KT204784 - KT204865
Dioscoreales			
<i>Lophiola aurea</i> Ker Gawl.	Nartheciaceae	Whitten 95028, K	KT204866 - KT204945

¹ Additional sequences: *Acorus calamus* L. (NC_007407), *Alstroemeria aurea* Graham (KC968976), *Amborella trichopoda* Baill. (NC_005086), *Buxus microphylla* Siebold & Zucc. (NC_009599), *Calycanthus floridus* var. *glaucus* (Willd.) Torr. & A.Gray (NC_004993), *Dioscorea elephantipes* (NC_009601), *Drimys granadensis* L.f. (NC_008456), *Elaeis oleifera* (Kunth) Cortés (EU016883–EU016962), *Fritillaria taipaiensis* P.Y. Li (NC_023247), *Hordeum vulgare* L. (NC_008590), *Illicium oligandrum* Merr. & Chun (NC_009600), *Japonolirion osense* Nakai (JQ068951- JQ069028), *Lemna minor* L. (NC_010109), *Lilium longiflorum* Thunb. (KC968977), *Liriodendron tulipifera* L. (NC_008326), *Musa acuminata* Colla (EU016983–EU017063), *Nandina domestica* Thunb. (NC_008336), *Nuphar advena* (Aiton) W.T.Aiton (NC_008788), *Oncidium* Sw. Gower Ramsey (NC_014056), *Orontium aquaticum* L. (NC_010109), *Oryza sativa* L. (NC_001320), *Phalaenopsis aphrodite* Rchb.f. subsp. *formosana* (NC_007499.1), *Phoenix dactylifera* L. (NC_013991), *Piper cenocladum* C.DC. (NC_008457), *Plantanus occidentalis* L. (NC_008335), *Saccharum officinarum* L. (NC_006084), *Smilax china* L. (HM536959), *Sorghum bicolor* (L.) Moench (NC_008602), *Spirodela polyrhiza* (L.) Schleid. (NC_015891), *Triticum aestivum* L. (NC_002762), *Typha latifolia* L. (NC_013823), *Veratrum patulum* Loes. (NC_022715), *Vitis vinifera* L. (NC_007957), *Wolffia australiana* (Benth.) Hartog & Plas (NC_015899), *Wolffiella lingulata* (Hegelm.) Hegelm. (NC_015894), *Yucca schidigera* Ortgies (DQ069337–DQ069702, EU016681–EU016700), *Zea mays* L. (NC_001666). For additional taxa, see Givnish *et al.* (2010) (Arecales, Asparagales, Commelinales, Dasyogonales, Poales), Barrett *et al.* (2013) (Arecales, Commelinales, Dasyogonales, Zingiberales), and Mennes *et al.* (2015) (Liliales and Pandanales)

Table B.2 Data partitioning schemes inferred using PartitionFinder or PartitionFinderProtein with the BIC criterion. (a) “Codon” partitioning scheme for the nucleotide matrix; (b) “GxC” (gene x codon) partitioning scheme; (c) Amino-acid partitioning scheme; and (d) GxC (gene x codon) partitioning scheme for matrix with re-alignments for *accD*, *rpl20* and *rps18* (see text for details). Plastid genes are indicated before the underscore; the ‘pos’ term after an understore indicates the codon position (not applicable for *rrn* genes).

Partition no.	Best Model	Partition subsets
a)		
1	GTR + G	codon_pos1
2	GTR + G	codon_pos2
3	GTR + G	codon_pos3
4	GTR + G	rrn16, rrn23, rrn4.5, rrn5
b)		
1	GTR + G	<i>accD_pos1</i> , <i>accD_pos2</i> , <i>clpP_pos3</i> , <i>lhbA_pos3</i> , <i>matK_pos2</i> , <i>ndhF_pos1</i> , <i>ndhF_pos2</i> , <i>petG_pos3</i> , <i>psbE_pos3</i> , <i>psbH_pos2</i> , <i>psbJ_pos2</i> , <i>psbJ_pos3</i> , <i>psbL_pos2</i> , <i>rpl16_pos1</i> , <i>rpl22_pos1</i> , <i>rpl22_pos2</i> , <i>rpl32_pos1</i> , <i>rps14_pos3</i> , <i>rps16_pos2</i> , <i>rps4_pos3</i> , <i>ycf3_pos2</i> , <i>ycf3_pos3</i>
2	GTR + G	<i>accD_pos3</i> , <i>atpB_pos3</i> , <i>atpE_pos3</i> , <i>atpF_pos3</i> , <i>atpI_pos3</i> , <i>matK_pos1</i> , <i>ndhC_pos3</i> , <i>ndhJ_pos3</i> , <i>ndhK_pos3</i> , <i>petN_pos3</i> , <i>psaI_pos2</i> , <i>psbA_pos3</i> , <i>psbD_pos3</i> , <i>psbH_pos3</i> , <i>psbK_pos3</i> , <i>rpl14_pos3</i> , <i>rpl20_pos3</i> , <i>rpl32_pos2</i> , <i>rpl33_pos3</i> , <i>rpoA_pos3</i> , <i>rpoB_pos3</i> , <i>rpoC1_pos3</i> , <i>rps2_pos3</i> , <i>rps8_pos3</i>
3	GTR + G	<i>5rps12_pos3</i> , <i>atpA_pos1</i> , <i>atpA_pos2</i> , <i>atpF_pos1</i> , <i>clpP_pos1</i> , <i>ndhA_pos1</i> , <i>psaC_pos2</i> , <i>psaI_pos3</i> , <i>psbH_pos1</i> , <i>psbM_pos2</i> , <i>rbcL_pos1</i> , <i>rbcL_pos2</i> , <i>rpl20_pos2</i> , <i>rpoC2_pos1</i> , <i>rps15_pos1</i> , <i>rps16_pos1</i> , <i>rps18_pos3</i> , <i>rps3_pos1</i> , <i>rps8_pos1</i> , <i>ycf4_pos1</i>
4	GTR + G	<i>atpA_pos3</i> , <i>matK_pos3</i> , <i>ndhE_pos3</i> , <i>petD_pos3</i> , <i>psaC_pos3</i> , <i>rpl16_pos3</i> , <i>rpl36_pos3</i> , <i>rps3_pos3</i>
5	GTR + G	<i>3rps12_pos3</i> , <i>5rps12_pos2</i> , <i>atpB_pos1</i> , <i>atpI_pos1</i> , <i>clpP_pos2</i> , <i>infA_pos2</i> , <i>lhbA_pos2</i> , <i>ndhB_pos3</i> , <i>ndhI_pos1</i> , <i>ndhJ_pos1</i> , <i>ndhJ_pos2</i> , <i>petB_pos1</i> , <i>petB_pos2</i> , <i>petD_pos2</i> , <i>psaB_pos1</i> , <i>psbA_pos1</i> , <i>psbA_pos2</i> , <i>psbB_pos2</i> , <i>psbC_pos2</i> , <i>psbD_pos2</i> , <i>psbK_pos2</i> , <i>psbL_pos1</i> , <i>psbM_pos1</i> , <i>psbN_pos1</i> , <i>psbT_pos2</i> , <i>rpl23_pos3</i> , <i>rpl36_pos2</i> , <i>rpoB_pos1</i> , <i>rpoB_pos2</i> , <i>rpoC1_pos2</i> , <i>rps11_pos1</i> , <i>rps14_pos2</i> , <i>rps15_pos2</i> , <i>rps2_pos2</i> , <i>rps3_pos2</i> , <i>rps4_pos2</i> , <i>rps7_pos3</i> , <i>ycf2_pos1</i> , <i>ycf2_pos2</i> , <i>ycf2_pos3</i> , <i>ycf3_pos1</i>
6	GTR + G	<i>atpB_pos2</i> , <i>atpF_pos2</i> , <i>ccsA_pos1</i> , <i>ccsA_pos2</i> , <i>lhbA_pos1</i> , <i>ndhA_pos2</i> , <i>ndhD_pos1</i> , <i>ndhD_pos2</i> , <i>ndhG_pos1</i> , <i>ndhK_pos1</i> , <i>petD_pos1</i> , <i>petL_pos1</i> , <i>petL_pos3</i> , <i>psaB_pos2</i> , <i>psaI_pos1</i> , <i>psbB_pos1</i> , <i>psbD_pos1</i> , <i>psbE_pos2</i> , <i>psbF_pos2</i> , <i>psbF_pos3</i> , <i>psbJ_pos1</i> , <i>psbN_pos3</i> , <i>rpl16_pos2</i> , <i>rpl20_pos1</i> , <i>rpl36_pos1</i> , <i>rpoC2_pos2</i> , <i>rps8_pos2</i>
7	GTR + G	<i>5rps12_pos1</i> , <i>atpE_pos1</i> , <i>atpE_pos2</i> , <i>atpI_pos2</i> , <i>cemA_pos1</i> , <i>cemA_pos2</i> , <i>infA_pos1</i> , <i>ndhC_pos1</i> , <i>ndhE_pos1</i> , <i>ndhG_pos2</i> , <i>ndhH_pos1</i> , <i>ndhH_pos2</i> , <i>ndhI_pos2</i> , <i>ndhK_pos2</i> , <i>petA_pos1</i> , <i>petA_pos2</i> , <i>petL_pos2</i> , <i>psaA_pos1</i> , <i>psaA_pos2</i> , <i>psaC_pos1</i> , <i>psaJ_pos1</i> , <i>psaJ_pos2</i> ,

Partition no.	Best Model	Partition subsets
8	GTR + G	psbC_pos1, psbL_pos3, psbN_pos2, rpl14_pos1, rpl2_pos3, rpl33_pos1, rpl33_pos2, rpoA_pos1, rpoA_pos2, rpoC1_pos1, rps11_pos2, rps14_pos1, rps18_pos1, rps18_pos2, rps19_pos1, rps19_pos2, rps2_pos1, rps4_pos1, ycf4_pos2
9	GTR + G	3rps12_pos1, 3rps12_pos2, atpH_pos1, atpH_pos2, ndhB_pos1, ndhB_pos2, ndhC_pos2, ndhE_pos2, petG_pos1, petG_pos2, petN_pos1, petN_pos2, psbE_pos1, psbF_pos1, psbI_pos1, psbI_pos2, psbT_pos1, rpl14_pos2, rpl23_pos1, rpl23_pos2, rpl2_pos1, rpl2_pos2, rps7_pos1, rps7_pos2, rrn16, rrn23, rrn4_5, rrn5
10	GTR + G	atpH_pos3, cemA_pos3, infA_pos3, ndhG_pos3, ndhI_pos3, petA_pos3, petB_pos3, psaA_pos3, psaB_pos3, psbB_pos3, psbC_pos3, psbI_pos3, psbM_pos3, psbT_pos3, rbcL_pos3, rpoC2_pos3, rps15_pos3, rps19_pos3, ycf4_pos3
11	GTR + G	ccsA_pos3, ndhD_pos3, ndhH_pos3, psbK_pos1, rpl22_pos3, rps16_pos3
12	GTR + I + G	ndhA_pos3, psaJ_pos3, rpl32_pos3, rps11_pos3 ndhF_pos3
c)		
1	JTT + G + F	accD
2	CPREV + I + G	atpA, psbC
3	JTT + G	atpB
4	JTT + G	atpE, clpP, ndhI
5	JTT + G	atpF
6	CPREV+G	atpH
7	JTT+G	5rps12, atpI
8	JTT+G+F	ccsA
9	JTT+G+F	cemA, ndhA
10	JTT+G	infA
11	CPREV+G	lhbA
12	JTT+G+F	matK
13	JTT+G+F	ndhB
14	CPREV + G	ndhC
15	CPREV + G	ndhD
16	MTMAM+G	ndhE
17	JTT + G + F	ndhF
18	JTT + G	ndhG
19	JTT + I + G	ndhH
20	JTT + G	ndhJ
21	JTT + G	ndhK
22	JTT + G	petA, rpoC1
23	JTT + G	petB
24	JTT + G	petD, psbH
25	CPREV + G	petG
26	MTMAM+G	petL
27	MTMAM+G	petN
28	CPREV + I + G	psaA
29	CPREV + I + G	psaB

Partition no.	Best Model	Partition subsets
30	JTT + I + G	psaC
31	CPREV + G	psaI
32	CPREV + G	psaJ
33	CPREV + I + G	psbA
34	JTT + I + G	psbB
35	CPREV + I + G	psbD
36	JTT + G	psbE
37	JTT + G	psbF, rps11
38	CPREV + G	psbI
39	MTMAM+G	psbJ
40	CPREV + G	psbK
41	MTMAM+G	psbL
42	CPREV + G	psbM
43	MTMAM+G	psbN
44	JTT + G	psbT
45	LG + G	rbcL
46	JTT + G	rpl2
47	JTT + G	rpl14
48	CPREV + G	rpl16
49	JTT + G	rpl20
50	JTT + G	rpl22
51	JTT + G	rpl23
52	JTT + G	rpl32
53	JTT + G	rpl33
54	CPREV + G	rpl36
55	JTT + G	rpoA
56	JTT + G	rpoB
57	JTT + G + F	rpoC2
58	JTT + G	rps2
59	JTT + G	rps3
60	JTT + G	rps4
61	JTT + G	rps7
62	CPREV + G	rps8
63	CPREV + G	3rps12
64	JTT + G	rps14
65	JTT + G	rps15
66	JTT + G	rps16
67	JTT + G + F	rps18
68	JTT + G	rps19
69	JTT + G + F	ycf2
70	JTT + G	ycf3
71	CPREV + G	ycf4
d)		
1	GTR + G	5rps12_pos3, accD_pos1, accD_pos2, clpP_pos3, ndhA_pos1, psaI_pos3, psbE_pos3, psbH_pos2, psbJ_pos1, psbJ_pos2, psbJ_pos3, rbcL_pos1, rbcL_pos2, rpl16_pos1, rpl22_pos1, rpl22_pos2, rpl33_pos1, rpl33_pos2, rpoC2_pos1, rps14_pos3, rps4_pos3, ycf3_pos2

Partition no.	Best Model	Partition subsets
2	GTR + G	accD_pos3, atpB_pos3, atpE_pos3, atpF_pos3, atpI_pos3, matK_pos1, ndhC_pos3, ndhJ_pos3, ndhK_pos3, petN_pos3, psaI_pos2, psbA_pos3, psbD_pos3, psbH_pos3, psbK_pos3, rpl14_pos3, rpl32_pos2, rpl33_pos3, rpoA_pos3, rpoB_pos3, rpoC1_pos3, rps2_pos3, rps8_pos3
3	GTR + G	atpA_pos1, atpB_pos2, atpF_pos1, ccsA_pos1, ndhG_pos1, ndhH_pos1, psaC_pos2, psaI_pos1, psbB_pos1, psbH_pos1, psbM_pos2, rpl16_pos2, rpl20_pos1, rpoA_pos1, rps11_pos1, rps15_pos1, rps3_pos1, rps8_pos1, ycf4_pos1
4	GTR + G	5rps12_pos1, atpA_pos2, atpE_pos1, atpE_pos2, atpF_pos2, atpI_pos2, ccsA_pos2, cemA_pos1, cemA_pos2, clpP_pos1, clpP_pos2, infA_pos1, infA_pos2, lhbA_pos1, ndhA_pos2, ndhC_pos1, ndhD_pos1, ndhD_pos2, ndhE_pos1, ndhG_pos2, ndhH_pos2, ndhI_pos2, ndhK_pos1, ndhK_pos2, petA_pos1, petA_pos2, petD_pos1, petL_pos1, petL_pos2, petL_pos3, psaA_pos1, psaB_pos2, psaC_pos1, psaJ_pos1, psbA_pos2, psbC_pos1, psbD_pos1, psbE_pos2, psbF_pos2, psbF_pos3, psbL_pos3, psbN_pos3, rpl20_pos2, rpl2_pos3, rpl36_pos1, rpoA_pos2, rpoC1_pos1, rpoC2_pos2, rps11_pos2, rps14_pos1, rps18_pos1, rps18_pos2, rps19_pos1, rps19_pos2, rps2_pos1, rps2_pos2, rps3_pos2, rps4_pos1, rps4_pos2, rps8_pos2, ycf4_pos2
5	GTR + G	atpA_pos3, matK_pos3, ndhE_pos3, petD_pos3, psaC_pos3, rpl16_pos3, rpl36_pos3, rps3_pos3
6	GTR + G	3rps12_pos3, 5rps12_pos2, atpB_pos1, atpI_pos1, lhbA_pos2, ndhB_pos3, ndhE_pos2, ndhI_pos1, ndhJ_pos1, ndhJ_pos2, petB_pos1, petB_pos2, petD_pos2, petG_pos1, psaB_pos1, psaJ_pos2, psbA_pos1, psbC_pos2, psbD_pos2, psbK_pos2, psbL_pos1, psbM_pos1, psbN_pos1, psbT_pos1, psbT_pos2, rpl14_pos1, rpl23_pos3, rpl2_pos1, rpl36_pos2, rpoB_pos1, rpoB_pos2, rpoC1_pos2, rps14_pos2, rps15_pos2, rps7_pos1, rps7_pos3, ycf2_pos1, ycf2_pos2, ycf2_pos3, ycf3_pos1
7	GTR + G	3rps12_pos1, 3rps12_pos2, atpH_pos1, atpH_pos2, ndhB_pos1, ndhB_pos2, ndhC_pos2, petG_pos2, petN_pos1, petN_pos2, psaA_pos2, psbB_pos2, psbE_pos1, psbF_pos1, psbI_pos1, psbI_pos2, psbN_pos2, rpl14_pos2, rpl23_pos1, rpl23_pos2, rpl2_pos2, rps16_pos1, rps7_pos2, rrn16, rrn23, rrn4_5, rrn5
8	GTR + G	atpH_pos3, cemA_pos3, infA_pos3, ndhG_pos3, ndhI_pos3, petA_pos3, petB_pos3, psaA_pos3, psaB_pos3, psbB_pos3, psbC_pos3, psbI_pos3, psbM_pos3, psbT_pos3, rbcL_pos3, rpoC2_pos3, rps15_pos3, rps19_pos3, ycf4_pos3
9	GTR + G	ccsA_pos3, ndhA_pos3, ndhD_pos3, ndhH_pos3, psaJ_pos3, psbK_pos1, rpl22_pos3, rpl32_pos3, rps11_pos3, rps16_pos3
10	GTR + G	lhbA_pos3, matK_pos2, ndhF_pos1, ndhF_pos2, petG_pos3, psbL_pos2, rpl20_pos3, rpl32_pos1, rps16_pos2, rps18_pos3, ycf3_pos3
11	GTR + I + G	ndhF_pos3

Table B.3 Log likelihoods of branch models for the 18 genes retained in the *Sciaphila* plastome. Note that the trans-spliced exons of *rps12* are treated operationally as two genes below; see text for further details.

Gene	Model	lnL ^a	LRT ^b	<i>P/P</i> -corrected ^c
3'- <i>rps12</i>	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.149$	-610.263		
	M1: $\omega_{\text{MHT}} = 0.202, \omega_{\text{green}} = 0.115$	-609.869	0.789	--/--
5'- <i>rps12</i>	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.150$	-506.460		
	M1: $\omega_{\text{MHT}} = 0.312, \omega_{\text{green}} = 0.115$	-506.460	2.672	--/--
<i>accD</i>	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.309$	-8167.290		
	M1: $\omega_{\text{MHT}} = 0.255, \omega_{\text{green}} = 0.316$	-8166.681	1.218	--/--
<i>clpP</i>	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.174$	-2868.960		
	M1: $\omega_{\text{MHT}} = 0.288, \omega_{\text{green}} = 0.154$	-2866.248	5.423	* / *
<i>matK</i>	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.367$	-12771.087		
	M1: $\omega_{\text{MHT}} = 0.491, \omega_{\text{green}} = 0.358$	-12769.319	3.536	--/--
<i>rpl2</i>	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.252$	-2303.569		
	M1: $\omega_{\text{MHT}} = 0.280, \omega_{\text{green}} = 0.235$	-2303.389	0.360	--/--
<i>rpl14</i>	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.111$	-1630.594		
	M1: $\omega_{\text{MHT}} = 0.240, \omega_{\text{green}} = 0.096$	-1627.974	5.240	* / *
<i>rpl16</i>	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.169$	-2310.496		
	M1: $\omega_{\text{MHT}} = 0.093, \omega_{\text{green}} = 0.185$	-2308.459	4.075	* / --
<i>rpl20</i>	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.353$	-1930.838		
	M1: $\omega_{\text{MHT}} = 0.313, \omega_{\text{green}} = 0.360$	-1930.759	0.158	--/--
<i>rpl36</i>	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.103$	-522.304		
	M1: $\omega_{\text{MHT}} = 0.100, \omega_{\text{green}} = 0.103$	-522.303	0.001	--/--
<i>rps2</i>	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.183$	-3730.279		
	M1: $\omega_{\text{MHT}} = 0.258, \omega_{\text{green}} = 0.171$	-3728.948	2.663	--/--
<i>rps3</i>	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.178$	-3847.162		

Gene	Model	lnL ^a	LRT ^b	<i>P/P</i> -corrected ^c
<i>rps4</i>	M1: $\omega_{\text{MHT}} = 0.224, \omega_{\text{green}} = 0.173$	-3846.696	0.931	--/--
	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.174$	-3056.894		
<i>rps7</i>	M1: $\omega_{\text{MHT}} = 0.211, \omega_{\text{green}} = 0.168$	-3056.556	0.677	--/--
	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.330$	-1147.401		
<i>rps8</i>	M1: $\omega_{\text{MHT}} = 0.611, \omega_{\text{green}} = 0.203$	-1144.219	6.364	* / *
	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.259$	-2262.744		
<i>rps11</i>	M1: $\omega_{\text{MHT}} = 0.403, \omega_{\text{green}} = 0.244$	-2261.666	2.156	--/--
	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.097$	-2310.140		
<i>rps14</i>	M1: $\omega_{\text{MHT}} = 0.129, \omega_{\text{green}} = 0.093$	-2309.673	0.933	--/--
	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.218$	-1449.355		
<i>rps18</i>	M1: $\omega_{\text{MHT}} = 0.138, \omega_{\text{green}} = 0.233$	-1448.617	1.400	--/--
	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.214$	-1416.231		
<i>rps19</i>	M1: $\omega_{\text{MHT}} = 0.192, \omega_{\text{green}} = 0.217$	-1416.198	0.065	--/--
	M0: $\omega_{\text{MHT}} = \omega_{\text{green}} = 0.160$	-1249.963		
	M1: $\omega_{\text{MHT}} = 0.111, \omega_{\text{green}} = 0.171$	-1249.468	0.320	--/--

Abbreviations for ω ratios: ω_{MHT} = *Sciaphila* and ω_{green} = green outgroups.

^a Log likelihood of the data for the model.

^b Log likelihood ratio test statistic $-2(\ln L \text{ M0} - \ln L \text{ M1})$ to evaluate differences in model fit.

^c *P*-values for uncorrected/Bonferroni corrected χ^2 tests, where * = $P < 0.05$ and dashes (--) indicates not significant.

Table B.4 Log likelihoods of branch-site models for the 18 genes retained in the *Sciaphila* plastome. “Staggered” refers to analyses performed on a realigned matrix for *accD*, *rpl20* and *rps18* (note that the trans-spliced exons of *rps12* are treated operationally as two genes below; see text for further details).

Gene	LRT ^a	<i>P/P</i> corrected ^b	LRT (staggered)	<i>P/P</i> -corrected (staggered)
3'- <i>rps12</i>	0	--/--		
5'- <i>rps12</i>	1.259	--/--		
<i>accD</i>	18.740	*** / ***	2.679	--/--
<i>clpP</i>	0	--/--		
<i>matK</i>	2.095	--/--		
<i>rpl2</i>	0	--/--		
<i>rpl14</i>	0.004	--/--		
<i>rpl16</i>	0	--/--		
<i>rpl20</i>	10.480	** / **	0	--/--
<i>rpl36</i>	0	--/--		
<i>rps2</i>	0	--/--		
<i>rps3</i>	0	--/--		
<i>rps4</i>	0.207	--/--		
<i>rps7</i>	0.124	--/--		
<i>rps8</i>	0	--/--		
<i>rps11</i>	0	--/--		
<i>rps14</i>	0	--/--		
<i>rps18</i>	3.887	*/--	0	--/--
<i>rps19</i>	0	--/--		

^a Log likelihood ratio test statistic $-2(\ln L H_0 - \ln L H_1)$ to evaluate differences in model fit.

^b *P*-values for uncorrected/Bonferroni corrected χ^2 tests, where * = $P < 0.05$, ** = $P < 0.01$,

*** = $P < 0.001$ and dashes (--) indicates not significant.

Figure B.1 Angiosperm phylogeny inferred from a likelihood analysis of 82 plastid coding regions (22 in *Sciaphila*), analyzed using a nucleotide-based substitution model (GTR+G) and an unpartitioned data set ($-\ln L = 660,751.389$). Bootstrap support values are indicated beside branches where less than 100%. The scale bar indicates the estimated substitutions per site.



Figure B.2 Angiosperm phylogeny inferred from a likelihood analysis of 82 plastid coding regions (22 in *Sciaphila*), analyzed using nucleotide-based substitution models with a “codon” data partitioning scheme (see text and table S2a for details) ($-\ln L = 651,480.509$). Bootstrap support values are indicated beside branches where less than 100%. The scale bar indicates the estimated substitutions per site.

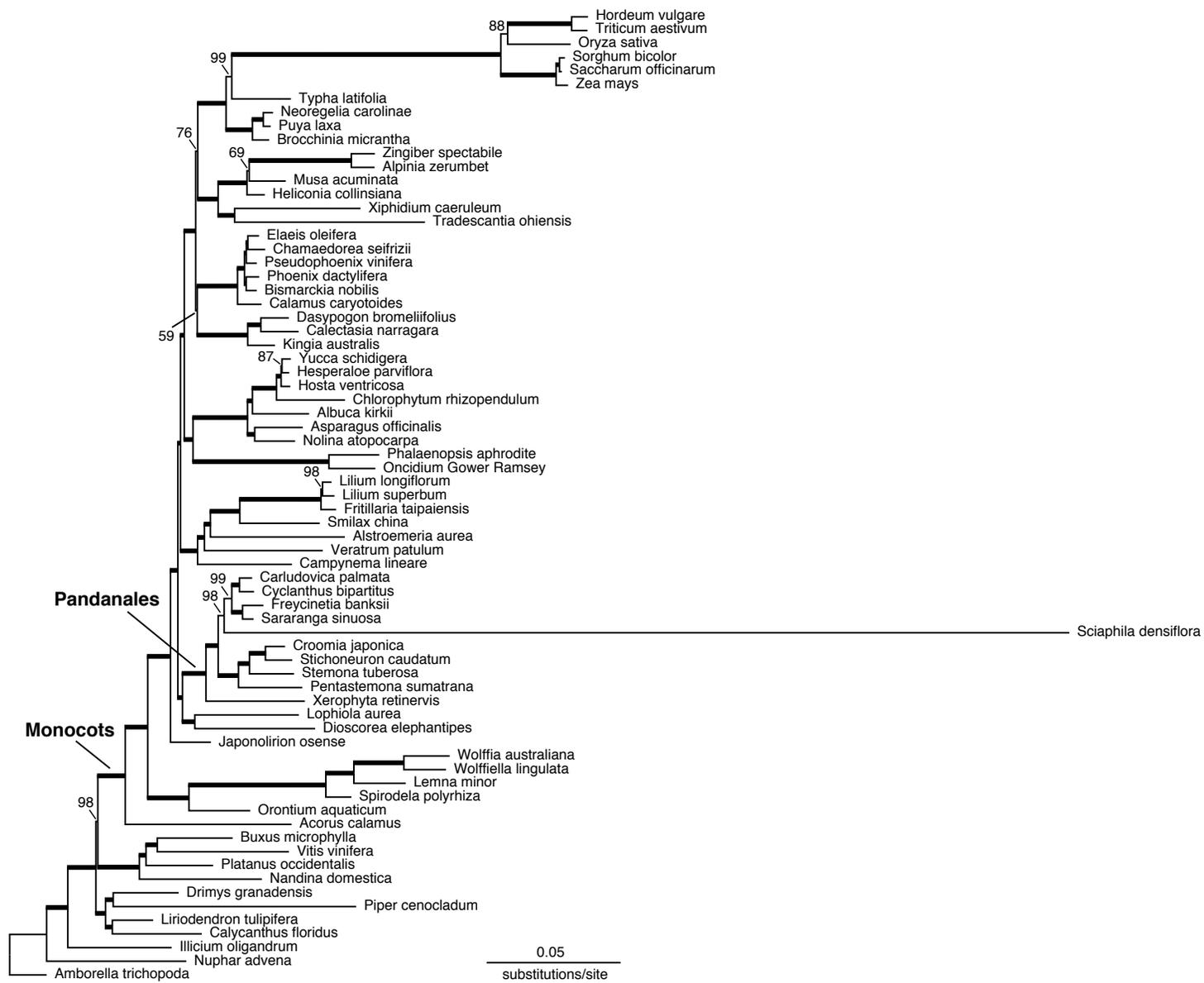


Figure B.3 Angiosperm phylogeny inferred from a likelihood analysis of 82 plastid coding regions (22 in *Sciaphila*), analyzed using nucleotide-based substitution models with a “G x C” (gene by codon) data partitioning scheme (see text and table S2b for details) (-lnL = 649,042.208). Bootstrap support values are indicated beside branches where less than 100%. The scale bar indicates the estimated substitutions per site.

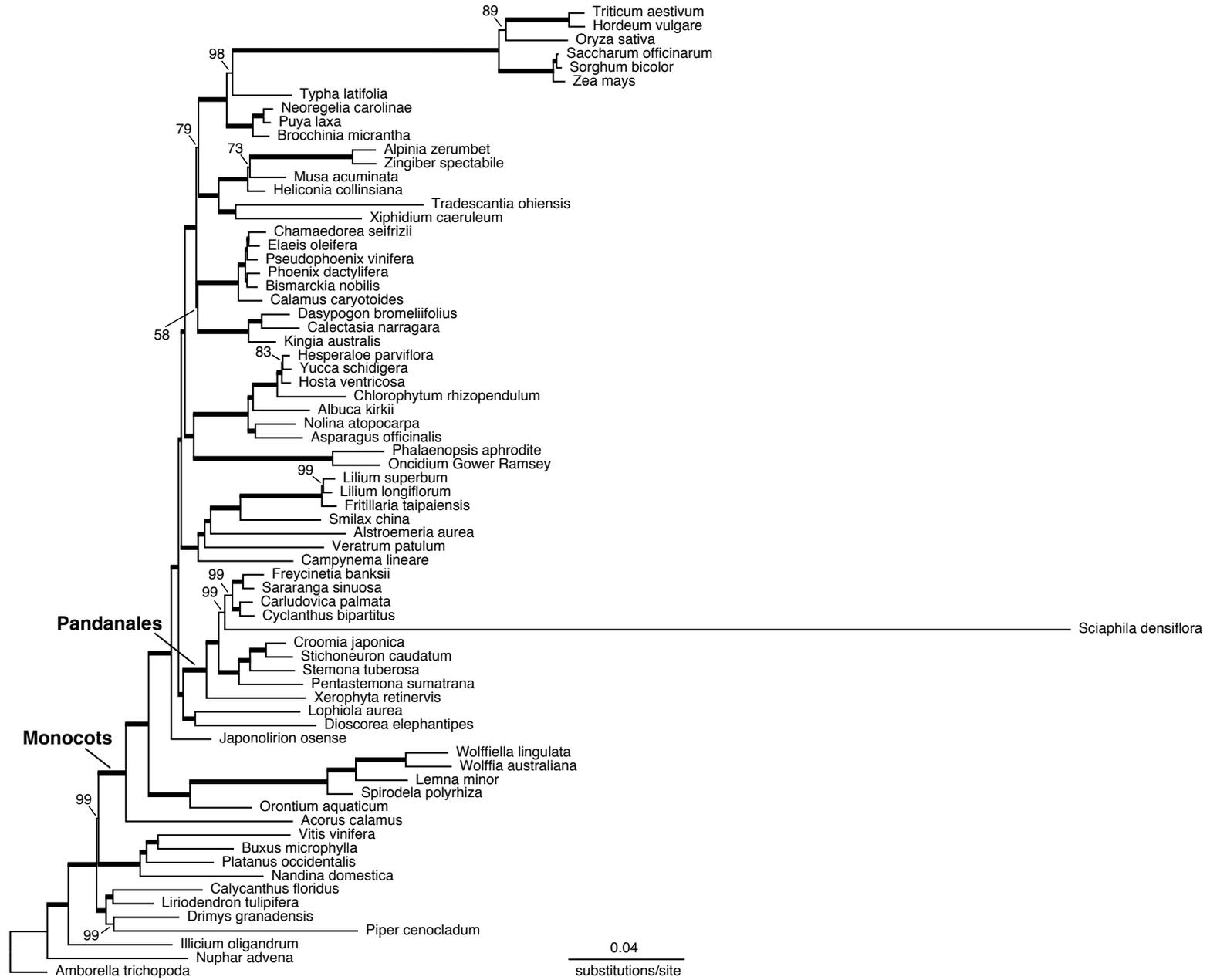


Figure B.4 Angiosperm phylogeny inferred from a likelihood analysis of 78 plastid protein coding regions (18 in *Sciaphila*), analyzed using an amino-acid model (JTT+G) and an unpartitioned data set ($-\ln L = 317,841.968$). Bootstrap support values are indicated besides branches where less than 100%. The scale bar indicates the estimated substitutions per residue.

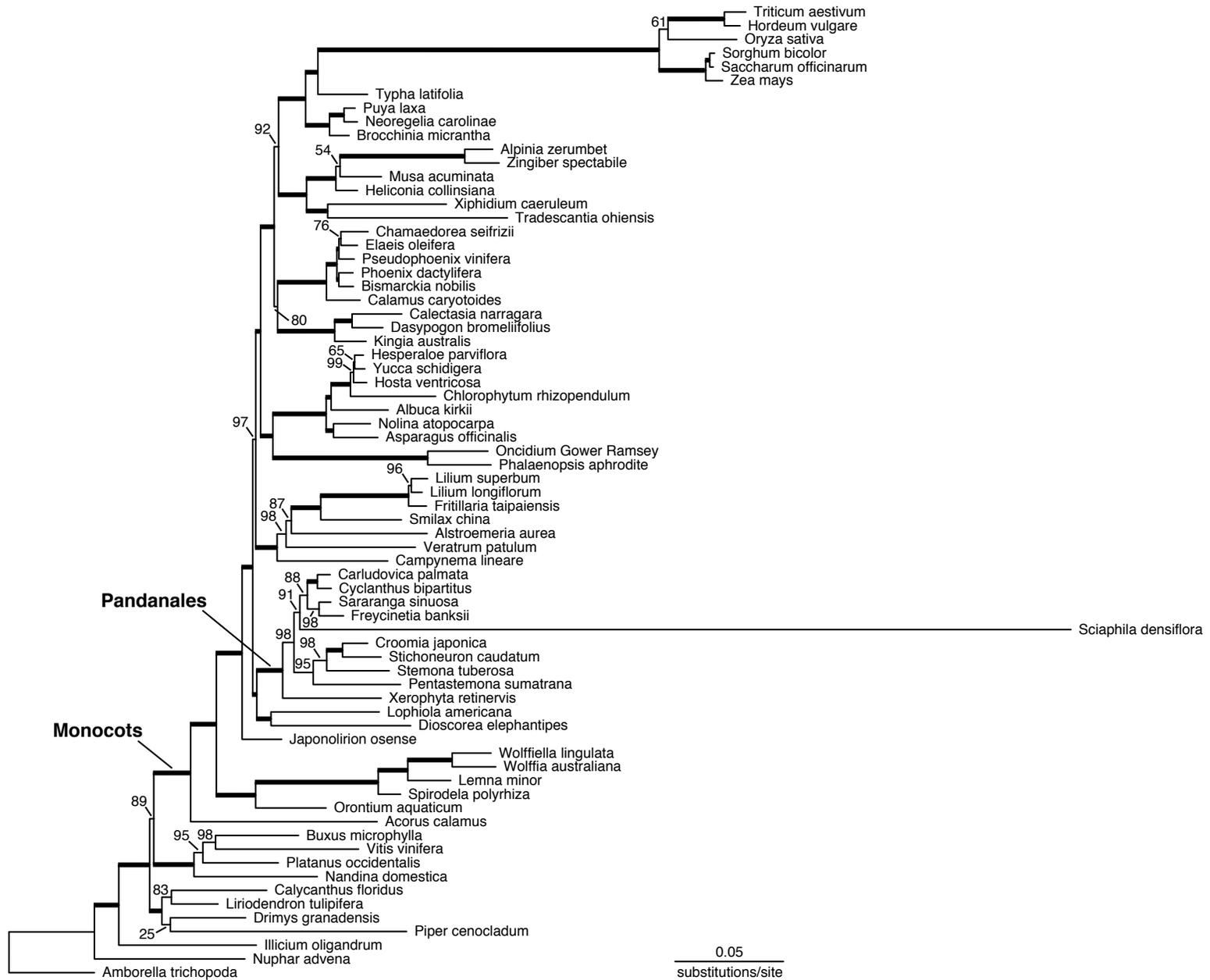


Figure B.5. Angiosperm phylogeny inferred from a likelihood analysis of 78 plastid protein coding regions (18 in *Sciaphila*), analyzed using amino-based substitution models with a gene-based data partitioning scheme (see text and table S2c for details) ($-\ln L = 314559.998$). Bootstrap support values are indicated beside branches where less than 100%. The scale bar indicates the estimated substitutions per residue.

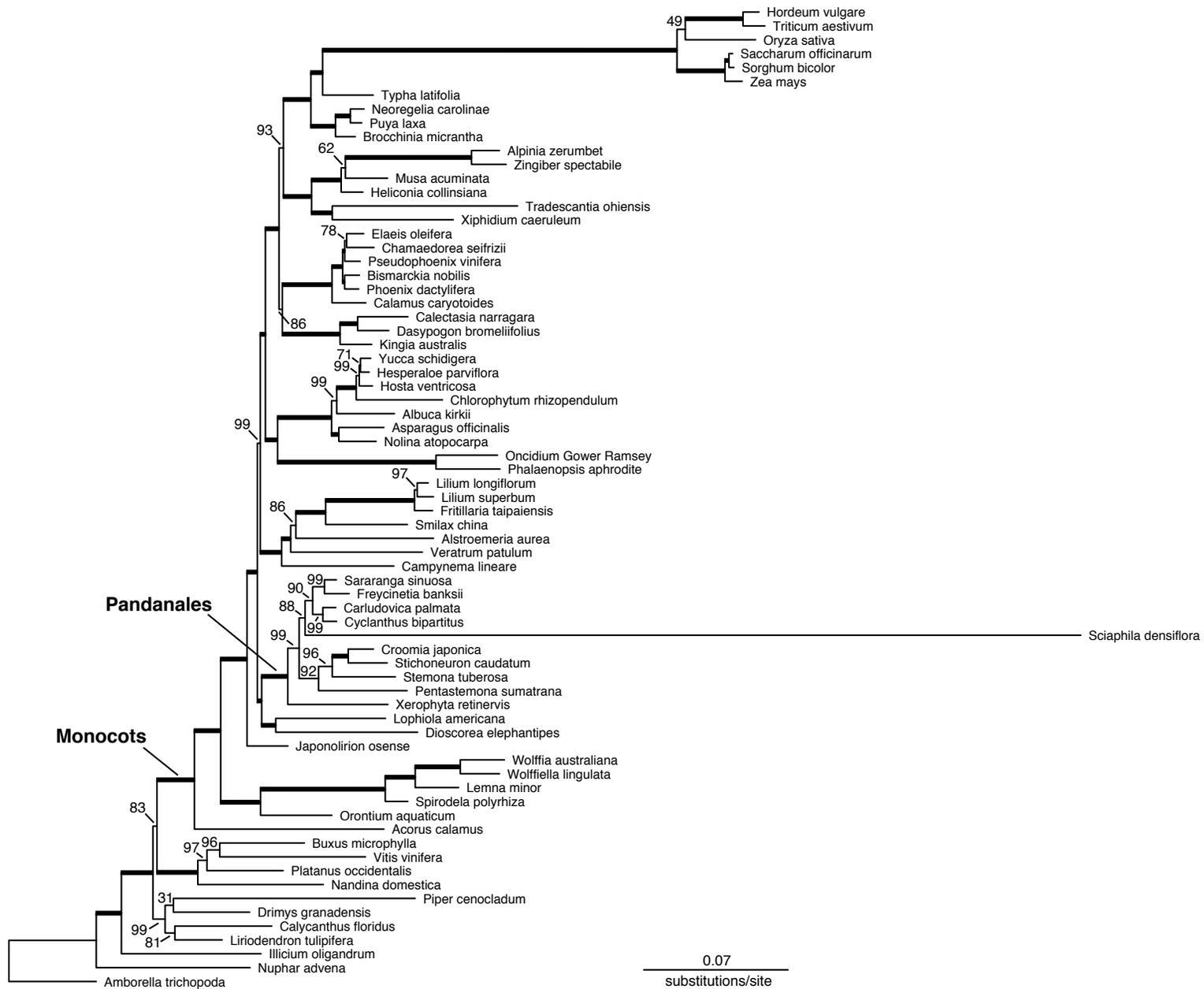


Figure B.6 Angiosperm phylogeny inferred from a likelihood analysis of 82 plastid coding regions (22 in *Sciaphila*), analyzed using a codon-based substitution model and an unpartitioned nucleotide data set ($-\ln L = 665,708.28$). Bootstrap support values were estimated for two subsets of the taxa (shown in bold): for branches with two bootstrap values, the first number indicates results from analyses that included only taxa from Dioscoreales and Pandanales, and the second number indicates results from analysis with additional selected taxa selected across monocots. A dot indicates bootstrap support values of 100 for both subsets; the value in parentheses corresponds to the unrooted branch in the analysis that included only Dioscoreales and Pandanales. The scale bar indicates the estimated substitutions per site.

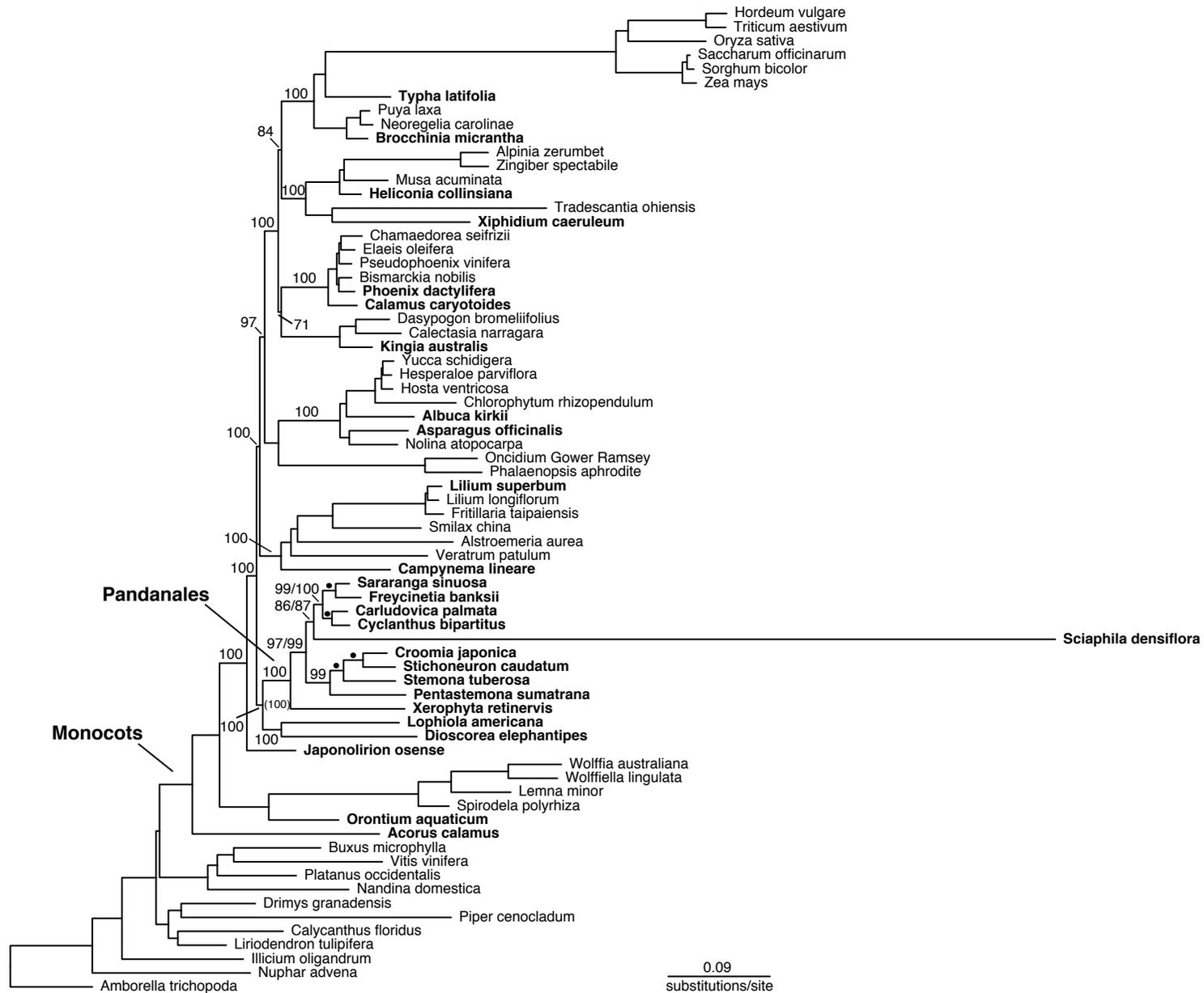


Figure B.7 Angiosperm phylogeny inferred from two parsimony analyses, the first of 82 plastid coding regions (22 in *Sciaphila*; shortest tree found, tree score 107,969), and the second of 82 plastid coding regions excluding *Sciaphila* (shortest tree found, tree score 106,099 steps). Thick lines indicate 100% bootstrap support values in both analyses. For branches with two bootstrap values, numbers above branches indicate results from the analysis that included *Sciaphila* and numbers below branches show results from the analysis excluding *Sciaphila*. Branches with only one number had the same bootstrap support in both analyses, except for the branch leading to *Sciaphila* and *Xerophyta*, for which the bootstrap value (shown in brackets) was only calculated for the first analysis. The scale bar indicates the inferred number of changes.

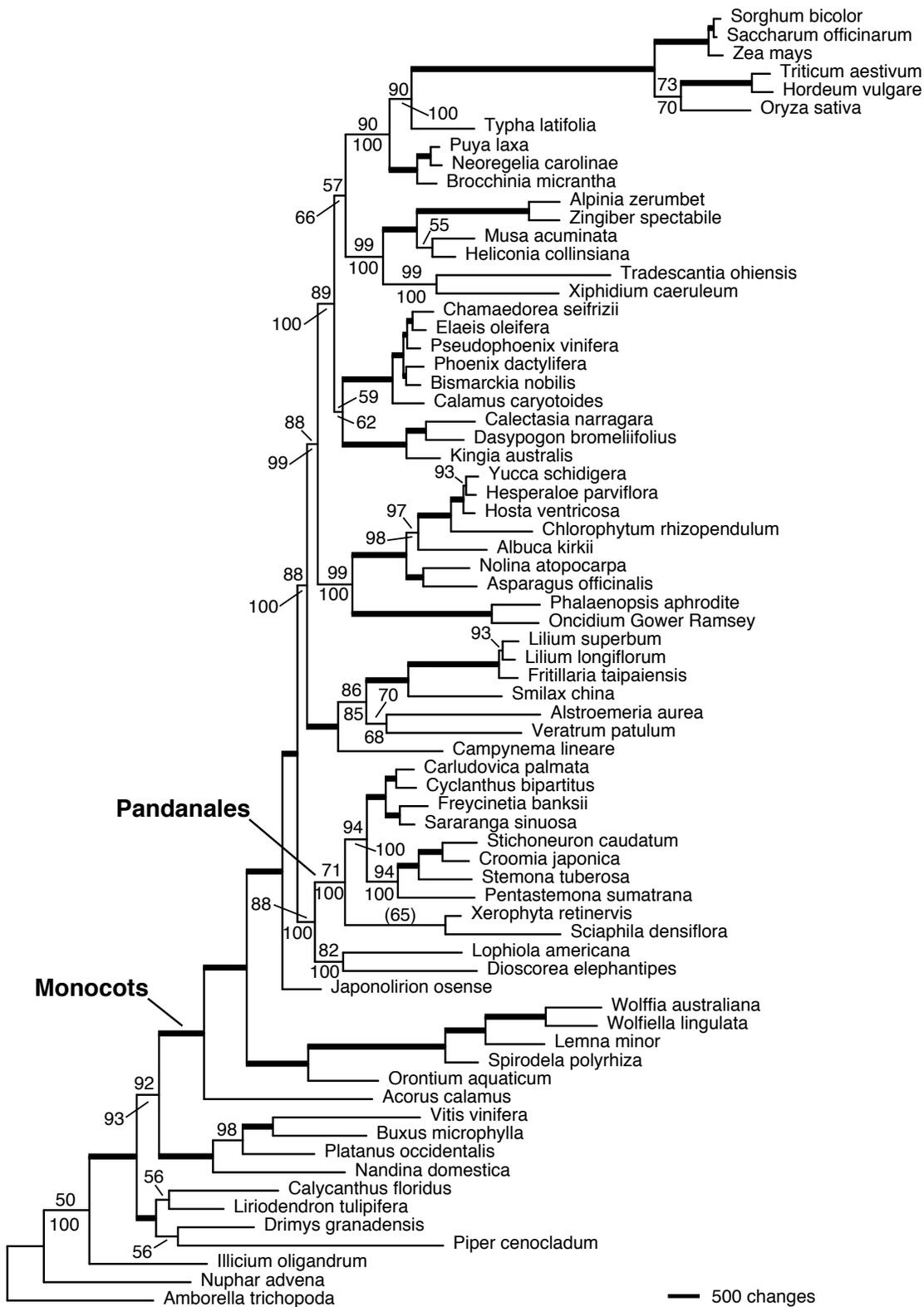


Figure B.8 Angiosperm phylogeny inferred from a likelihood analysis of 82 plastid coding regions (22 in *Sciaphila*), analyzed using a nucleotide-based substitution model (GTR+G) and an unpartitioned data set ($-\ln L = 660,544.84$), with realignments for *accD*, *rpl20* and *rps18* (see text for details). Bootstrap support values are indicated beside branches where less than 100%. The scale bar indicates the estimated substitutions per site.

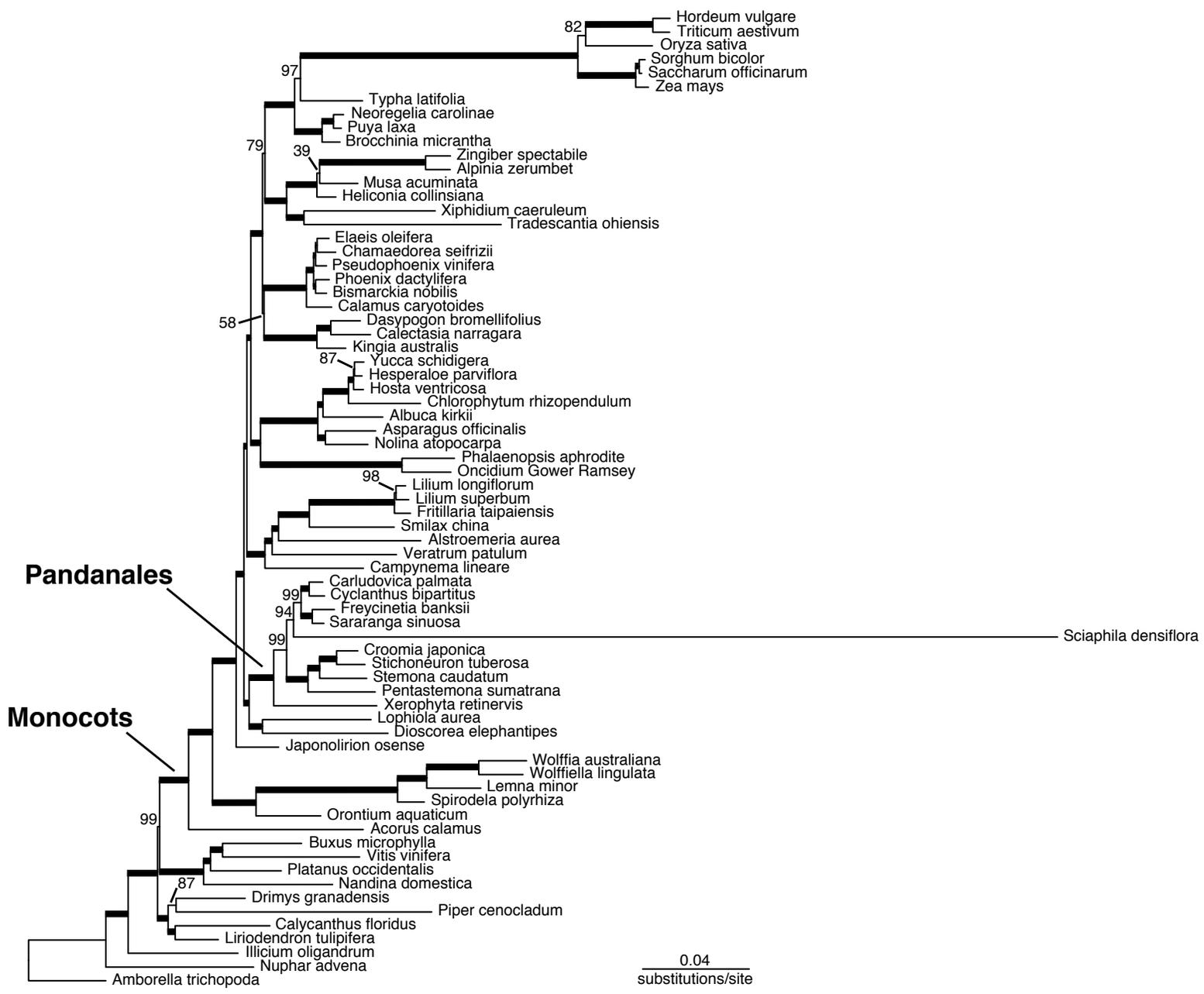


Figure B.9 Angiosperm phylogeny inferred from a likelihood analysis of 82 plastid coding regions (22 in *Sciaphila*), analyzed using nucleotide-based substitution models with a “G x C” (gene by codon) data partitioning scheme (see text and table S2d for details) ($-\ln L = 648,801.64$), with realignments for *accD*, *rpl20* and *rps18* (see text for details). Bootstrap support values are indicated beside branches where less than 100%. The scale bar indicates the estimated substitutions per site.



Appendix C: Supplementary tables and figures for Chapter 4

Table C.1 Specimen source information for fully mycoheterotrophic lineages, and one or more photosynthetic taxa from the same lineage; herbarium abbreviations follow Thiers (continuously updated). Fully mycoheterotrophic species indicated with an asterisk (*).

Species ¹	Family	Voucher number	GenBank accession
Asparagales			
<i>Iris missouriensis</i> Nutt.	Iridaceae	M.A. McPherson 000707-5a-7, ALTA	NC_XXXXXX
* <i>Geosiris aphylla</i> Baill.	Iridaceae	Prance 30781, K	NC_XXXXXX
Dioscoreales			
* <i>Apteria aphylla</i> (Nutt.) Barnhart ex. Small	Burmanniaceae	D.M. McNair 952, USMS	NC_XXXXXX
<i>Burmannia bicolor</i> Mart.	Burmanniaceae	Maas et al. 9649, U	NC_XXXXXX
<i>Burmannia capitata</i> (Walter ex. J.F. Gmel)	Burmanniaceae	Maas et al. 9606, U	NC_XXXXXX
* <i>Burmannia itoana</i> Makino	Burmanniaceae	Kun-Ping Lo 821, PPI	NC_XXXXXX
* <i>Campylosiphon congestus</i> (C.H. Wright)	Burmanniaceae	T. Franke K/8, M	NC_XXXXXX
* <i>Gymnosiphon longistylus</i> (Benth.) Hutch.	Burmanniaceae	Merckx et al. 132, LV	NC_XXXXXX
Liliales²			
* <i>Arachnitis uniflora</i> Phil.	Corsiaceae	R. Neyland 1928, MCN	KP_462884.1
* <i>Corsia</i> cf. <i>boridiensis</i> P.Royen	Corsiaceae	S. Lyon SPL470-2, PNG	KP_462885.1
<i>Campynema lineare</i> Labill.	Campynemataceae	M.F. Duretto 1842, HO	KP_462881.1

Species ¹	Family	Voucher number	GenBank accession
<i>Lilium superbum</i> L.	Liliaceae	M.W. Chase 112, NCU	KP_462883.1
Pandanales³			
<i>Carludovica palmata</i> Ruiz & Pav.	Cyclanthaceae	M.W. Chase 14836, K	NC_0267856.1
* <i>Sciaphila densiflora</i> Schltr.	Triuridaceae	Y.Pillon et al. 88, HOU, P	KR_902497.1
Petrosaviales			
<i>Japonolirion osense</i> Nakai	Petrosaviaceae	M.W. Chase 3000, K	NC_XXXXXX
* <i>Petrosavia sakurarii</i> (Makino) J.Jsm. ex Steenis	Petrosaviaceae	Yukawa 09-47, TNS	NC_XXXXXX
* <i>Petrosavia</i> aff. <i>sakurarii</i> (Makino) J.Jsm. ex Steenis	Petrosaviaceae	Yukawa 09-25, TNS	NC_XXXXXX

¹Additional published species included here: *Acorus calamus* L. (NC_007407), *Amborella trichopoda* Baill. (NC_005086), *Buxus microphylla* Siebold & Zucc. (NC_009599), *Calycanthus floridus* var. *glaucus* (Willd.) Torr. & A.Gray (NC_004993), *Dioscorea elephantipes* (NC_009601), *Drimys granadensis* L.f. (NC_008456), *Elaeis oleifera* (Kunth) Cortés (EU016883-EU016962), *Fritillaria taipaiensis* P.Y.Li (NC_023247), *Hordeum vulgare* L. (NC_008590), *Illicium oligandrum* Merr. & Chun (NC_009600), *Musa acuminata* Colla (EU016983-EU017063), *Nandina domestica* Thunb. (NC_008336), *Nuphar advena* (Aiton) W.T. Aiton (NC_008788), *Oncidium* Sw. Gower Ramsey (NC_014056), *Petrosavia stellaris* Becc. (KF482381.1), *Phoenix dactylifera* L. (NC_013991), *Piper cenocladum* C.DC. (NC_008457), *Platanus occidentalis* L. (NC_008335), *Saccharum officinarum* L. (NC_006084), *Smilax china* L. (HM536959), *Sorghum bicolor* (L.) Moench (NC_008602), *Spirodela polyrhiza* (L.) Schleid (NC_015891), *Triticum aestivum* L. (NC_002762), *Typha latifolia* L. (NC_013823), *Veratrum patulum* Loes. (NC_022715), *Vitis vinifera* L. (NC_007957), *Wolffia australiana* (Benth.) Hartog & Plas (NC_015899), *Wolffiella lingulata* (Hegelm.) Hegelm. (NC_015894), *Yucca schidigera* Roezli ex Ortgies (DQ069337–DQ069702, EU016681–EU016700), *Zea mays* L. (NC_001666). For additional taxa,

see Givnish et al. (2010) (Arecales, Asparagales, Commelinales, Dasypogonales, Poales), Givnish et al. 2015 (Liliales), Barrett et al. (2013) (Arecales, Commelinales, Dasypogonales, Zingiberales).

²Gene sets published previously in Mennes et al. (2015).

³Plastid genome for *Sciaphila* and gene sets for several Dioscoreales and Pandanales published in Mennes et al. (2015) and Chapter 3.

Table C.2 Data partitioning scheme inferred using PartitionFinder with the BIC criterion (see text for details). Plastid genes are indicated before the first underscore; ‘pos’ = codon position, ‘exon’ = exon number, and ‘intron’ = intron number.

Partition no.	Best model	Partition subsets
1	GTR + G	atpA_pos1, atpE_pos2, atpF_pos2, ndhA_pos1, ndhA_pos2, ndhD_pos1, ndhD_pos2, ndhH_pos1, ndhH_pos2, petA_pos2, petB_pos1, psaA_pos1, psaB_pos1, psbC_pos1, psbC_pos2, psbF_pos3, psbH_pos1, lhbA_pos1, rbcL_pos2, rpl16_pos2, rpl2_pos3, rpl33_pos1, rpoA_pos1, rpoA_pos2, rpoC2_pos2, rps11_pos1, rps3_pos2, rps8_pos1, ycf4_pos1
2	GTR + G	3rps12_pos3, 5rps12_pos1, atpA_pos2, atpE_pos1, atpI_pos2, cemA_pos2, clpP_pos2, infA_pos1, infA_pos2, ndhC_pos1, ndhE_pos1, ndhJ_pos1, ndhJ_pos2, ndhK_pos1, ndhK_pos2, petA_pos1, petD_pos1, petD_pos2, petL_pos1, petL_pos2, psaJ_pos1, psaJ_pos2, psbA_pos2, psbE_pos2, psbT_pos2, rpl14_pos1, rpl23_pos3, rpoB_pos1, rpoB_pos2, rpoC1_pos1, rpoC1_pos2, rps11_pos2, rps14_pos1, rps14_pos2, rps16_pos1, rps19_pos1, rps19_pos2, rps2_pos1, rps2_pos2, rps4_pos2, ycf2_pos3, ycf4_pos2
3	GTR + G + I	atpA_pos3, ndhI_pos3, petD_pos3, psaC_pos3, psbB_pos3, psbH_pos2, psbM_pos3, psbT_pos3
4	GTR + G	5rps12_pos2, atpB_pos1, atpB_pos2, atpI_pos1, clpP_pos1, ndhG_pos2, ndhI_pos1, ndhI_pos2, psaA_pos2, psaB_pos2, psaC_pos1, psaC_pos2, psbB_pos2, psbD_pos1, psbD_pos2, psbE_pos1, psbL_pos1, psbL_pos3, psbM_pos1, psbN_pos1, psbN_pos2, rpl2_pos1, rpl36_pos1, rpl36_pos2, rps18_pos1, rps18_pos2, rps3_pos1, rps4_pos1, rps7_pos1, rps8_pos2, ycf2_pos1, ycf2_pos2
5	GTR + G	atpB_pos3, ndhE_pos3, petA_pos3, petB_pos3, psaA_pos3, rbcL_pos3, rpl16_pos3, rpl32_pos3, rpoC2_pos3, rps19_pos3
6	GTR + G	atpE_pos3, infA_pos3, ndhG_pos3, ndhJ_pos3, psaB_pos3, psbA_pos3, psbC_pos3, psbI_pos3, psbK_pos3, rpoA_pos3, rps3_pos3, ycf4_pos3
7	GTR + G	5rps12_pos3, accD_pos1, accD_pos2, atpF_pos1, cemA_pos1, ndhG_pos1, psaI_pos1, psaI_pos3, psbB_pos1, psbF_pos2, psbJ_pos1, psbJ_pos3, psbK_pos2, psbN_pos3, psbZ_pos3, rpl16_pos1, rpl20_pos1,

Partition no.	Best model	Partition subsets
		rpl20_pos2, rpl32_pos2, rpl33_pos2, rpoC2_pos1, rps15_pos2, rps16_pos2, rps18_pos3, rps4_pos3, ycf3_pos2
8	GTR + G	accD_pos3, atpF_pos3, atpH_pos3, atpI_pos3, cemA_pos3, matK_pos1, ndhC_pos3, ndhK_pos3, petN_pos3, psaI_pos2, psbD_pos3, psbE_pos3, psbH_pos3, psbJ_pos2, rpl14_pos3, rpl20_pos3, rpl33_pos3, rpoB_pos3, rpoC1_pos3, rps15_pos3, rps2_pos3, rps8_pos3
9	GTR + G	3rps12_pos2, atpH_pos1, ndhB_pos1, ndhB_pos2, ndhB_pos3, ndhC_pos2, ndhE_pos2, petB_pos2, petG_pos1, petN_pos1, psbA_pos1, psbF_pos1, psbI_pos1, psbM_pos2, psbT_pos1, psbZ_pos2, rpl14_pos2, rpl23_pos1, rpl23_pos2, rpl2_pos2, rps7_pos2, rps7_pos3, rrn16, rrn23, rrn4, rrn5, ycf3_pos1
10	GTR + G	3rps12_pos1, atpH_pos2, petG_pos2, petN_pos2, psbI_pos2
11	GTR + G	ccsA_pos1, clpP_pos3, ndhF_pos1, ndhF_pos2, petG_pos3, petL_pos3, psbL_pos2, rpl22_pos1, rpl32_pos1, rps15_pos1, ycf3_pos3
12	GTR + G	ccsA_pos2, matK_pos2, rbcL_pos1, rpl22_pos2, rps14_pos3
13	GTR + G + I	ccsA_pos3, ndhD_pos3, ndhH_pos3
14	GTR + G	matK_pos3, rpl36_pos3
15	GTR + G + I	ndhA_pos3, psbK_pos1
16	GTR + G + I	ndhF_pos3
17	GTR + G	psaJ_pos3, rpl22_pos3, rps11_pos3
18	GTR + G	rps16_pos3

Table C.3. Status of *matK* and group IIA introns in retained plastid genes of full mycoheterotrophs taxa. ORF = open reading frame.

Taxon		<i>matK</i> status	Retained genes ¹ and their group IIA intron status
Burmanniaceae	<i>Apteria aphylla</i>	Lost	3'- <i>rps12</i> ; intron absent <i>rpl2</i> ; intron absent
	<i>Burmannia itoana</i>	Lost	3'- <i>rps12</i> ; intron absent <i>rpl2</i> ; intron present <i>clpP</i> ; intron present
	<i>Campylosiphon congestus</i>	ORF	<i>trnA</i> -UGC, intron present 3'- <i>rps12</i> ; intron present <i>rpl2</i> ; intron present <i>clpP</i> ; intron 2 present
	<i>Gymnosiphon longistylus</i>	Lost	3'- <i>rps12</i> ; intron present <i>rpl2</i> ; intron present <i>clpP</i> ; intron 2 present
Corsiaceae	<i>Arachnitis uniflora</i>	Lost	3'- <i>rps12</i> ; intron present <i>rpl2</i> ; intron present <i>clpP</i> ; intron 2 absent
	<i>Corsia</i> cf. <i>boridiensis</i>	ORF	3'- <i>rps12</i> ; intron present <i>rpl2</i> ; intron present <i>clpP</i> ; intron 2 present
Iridaceae	<i>Geosiris aphylla</i>	ORF	<i>trnA</i> -UGC; intron present <i>trnI</i> -GAU; intron present <i>trnK</i> -UUU; intron present <i>trnV</i> -UAC; intron present <i>rpl2</i> ; intron present 3'- <i>rps12</i> ; intron present <i>clpP</i> ; intron 2 present
Orchidaceae ²	<i>Corallorhiza striata</i>	ORF	<i>trnA</i> -UGC; intron present <i>trnI</i> -GAU; intron present

Taxon	<i>matK</i> status	Retained genes ¹ and their group IIA intron status
		<i>trnK</i> -UUU; intron present <i>trnV</i> -UAC; intron present <i>rpl2</i> ; intron present 3'- <i>rps12</i> ; intron present <i>clpP</i> ; intron 2 present
<i>Epipogium aphyllum</i>	Lost	<i>rpl2</i> ; intron present <i>clpP</i> ; intron 2 present ⁴
<i>Epipogium roseum</i>	Lost	<i>rpl2</i> ; intron present <i>clpP</i> ; intron 2 present
<i>Neottia nidus-avis</i>	ORF? ³	<i>trnA</i> -UGC; intron present <i>trnK</i> -UUU; intron present <i>rpl2</i> , intron present 3'- <i>rps12</i> ; intron present <i>clpP</i> ; intron 2 present
<i>Rhizanthella gardneri</i>	Lost	<i>rpl2</i> ; intron present <i>clpP</i> ; intron present
Petrosaviaceae ⁵		
<i>Petrosavia sakuraii</i>	ORF	<i>trnA</i> -UGC; intron present <i>trnI</i> -GAU; intron present <i>trnK</i> -UUU; intron present <i>trnV</i> -UAC; intron present <i>atpF</i> ; intron present <i>rpl2</i> ; intron present 3'- <i>rps12</i> ; intron present <i>clpP</i> ; intron 2 present
<i>Petrosavia</i> aff. <i>sakuraii</i>	ORF	<i>trnA</i> -UGC; intron present <i>trnI</i> -GAU; intron present <i>trnK</i> -UUU; intron present <i>trnV</i> -UAC; intron present <i>atpF</i> ; intron present <i>rpl2</i> ; intron present 3'- <i>rps12</i> ; intron present

Taxon	<i>matK</i> status	Retained genes ¹ and their group IIA intron status	
<i>Petrosavia stellaris</i>	ORF	<i>clpP</i> ; intron 2 present <i>trnA</i> -UGC; intron present <i>trnI</i> -GAU; intron present <i>trnK</i> -UUU; intron present <i>trnV</i> -UAC; intron present <i>atpF</i> ; intron present <i>rpl2</i> ; intron present 3'- <i>rps12</i> ; intron present <i>clpP</i> ; intron 2 present	
Triuridaceae	<i>Sciaphila densiflora</i>	ORF	3'- <i>rps12</i> ; intron present <i>rpl2</i> ; intron present <i>clpP</i> ; intron 2 present

¹Retained plastid genes that typically have group IIA introns

²Published plastome sequences: *Corallorhiza*, Barrett and Davis (2012); *Epipogium* spp, Schelkunov et al. (2015); *Neottia*, Logacheva et al. (2011); *Rhizanthella*, Delannoy et al. (2011)

³Truncated *matK* open reading frame present (reported as a possible pseudogene in Logacheva et al. 2011)

⁴*Epipogium aphyllum* has a novel intron present (see Schelkunov et al. 2015 for details).

⁵Published plastome sequence: *Petrosavia stellaris*, Logacheva et al. (2014)

Figure C.1 Circular plastome map of autotrophic *Campynema lineare* (Campynemataceae). Genes located inside the circle are transcribed clockwise, those outside 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 (*). Inferred pseudogenes are marked with 'ψ'.

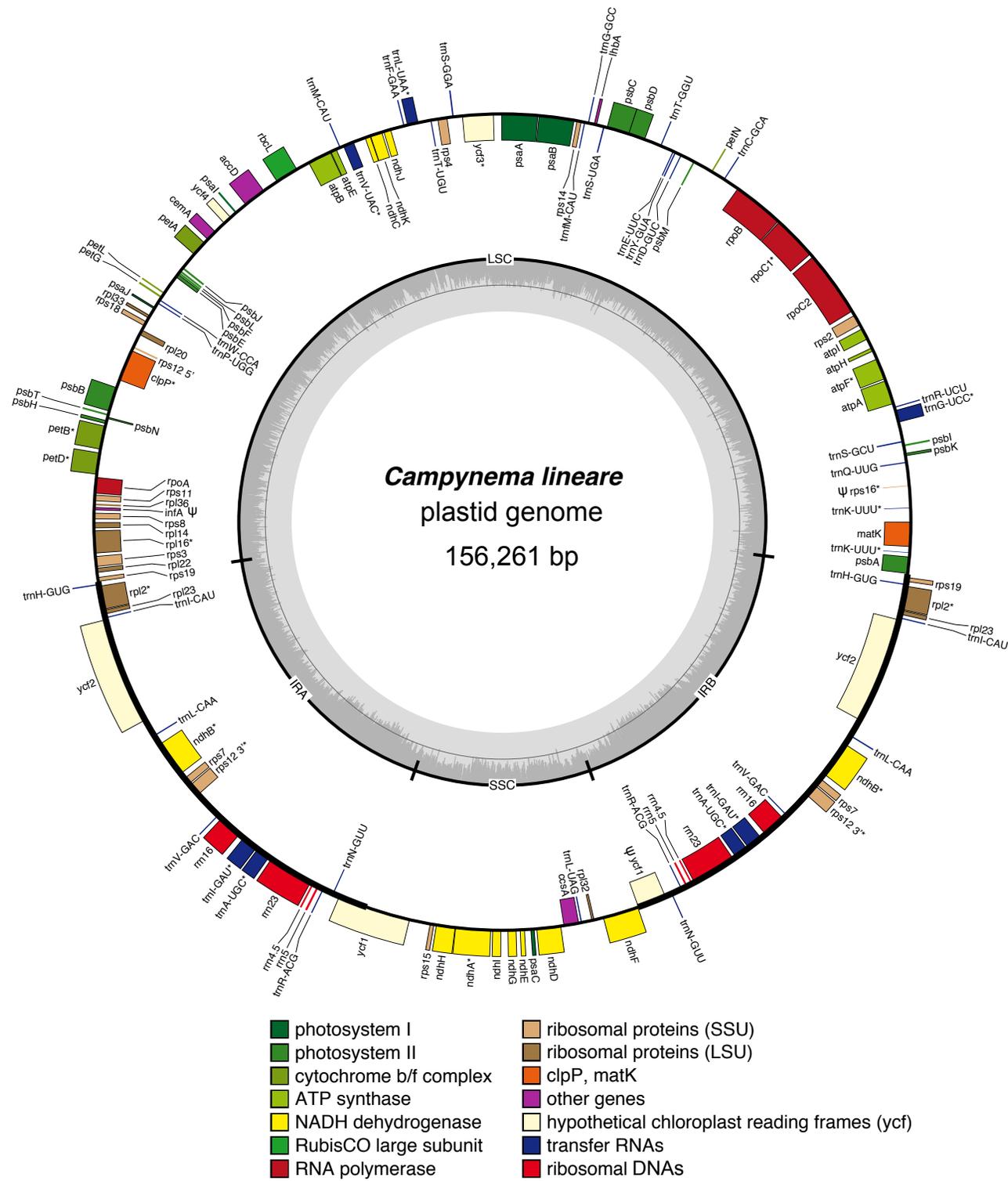
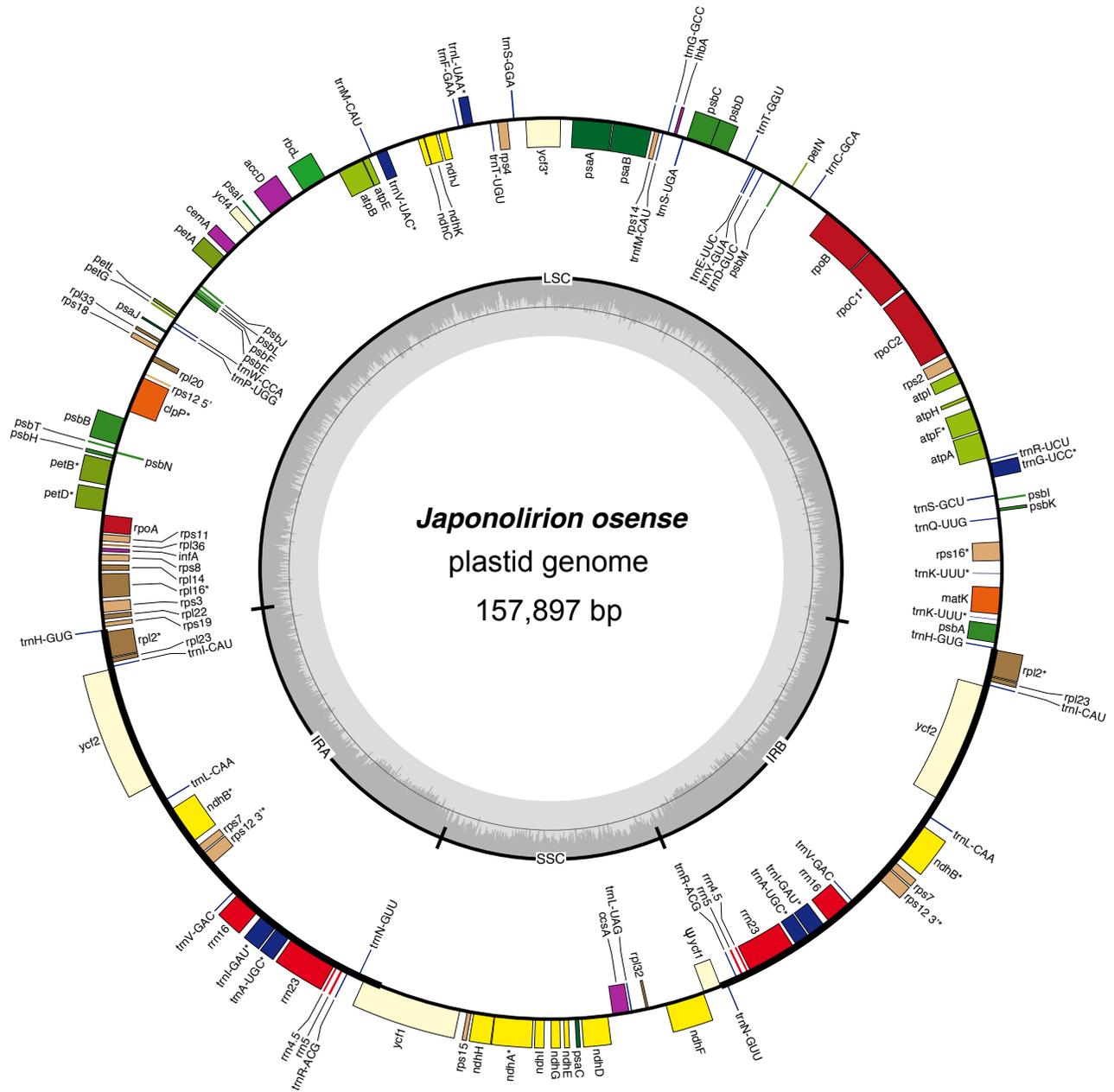


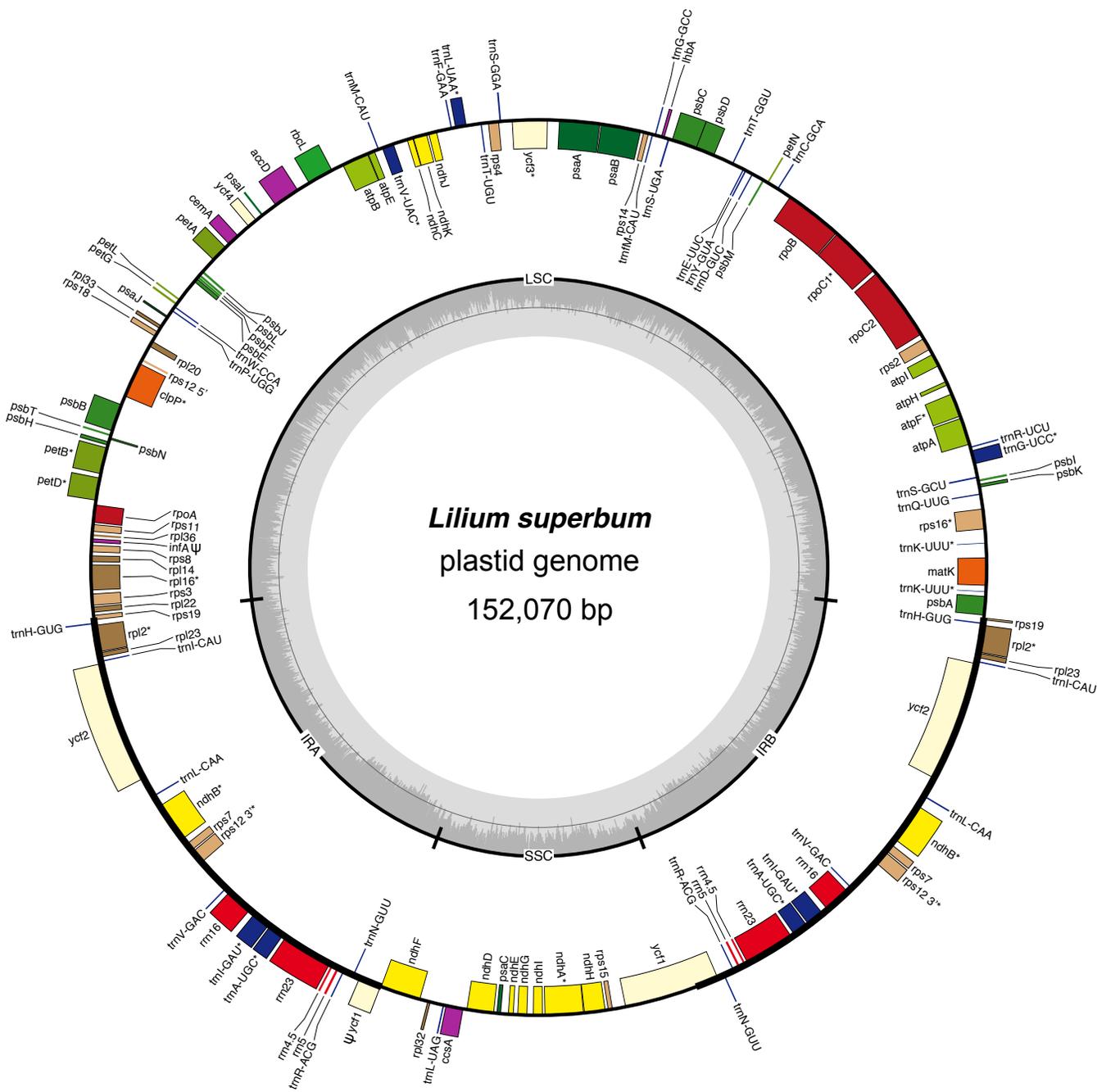
Figure C.2 Circular plastome map of autotrophic *Iris missouriensis* (Iridaceae). Genes located inside the circle are transcribed clockwise, those outside counterclockwise. The grey circle marks the GC content; the inner circle marks a 50% threshold. Thick black lines indicate inverted repeat (IR) copies. Genes with introns are indicated with asterisks (*). Inferred pseudogenes are marked with ‘ ψ ’.

Figure C.3 Circular plastome map of autotrophic *Japonolirion osense* (Petrosaviaceae). Genes located inside the circle are transcribed clockwise, those outside 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 (*). Inferred pseudogenes are marked with 'ψ'.



- | | |
|---|--|
| ■ photosystem I | ■ ribosomal proteins (SSU) |
| ■ photosystem II | ■ ribosomal proteins (LSU) |
| ■ cytochrome b/f complex | ■ clpP, matK |
| ■ ATP synthase | ■ other genes |
| ■ NADH dehydrogenase | ■ hypothetical chloroplast reading frames (ycf) |
| ■ RubisCO large subunit | ■ transfer RNAs |
| ■ RNA polymerase | ■ ribosomal DNAs |

Figure C.4 Circular plastome map of autotrophic *Lilium superbum* (Liliaceae). Genes located inside the circle are transcribed clockwise, those outside 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 (*). Inferred pseudogenes are marked with ‘ ψ ’.



Lilium superbium
 plastid genome
 152,070 bp

- | | |
|---|--|
| photosystem I | ribosomal proteins (SSU) |
| photosystem II | ribosomal proteins (LSU) |
| cytochrome b/f complex | clpP, matK |
| ATP synthase | other genes |
| NADH dehydrogenase | hypothetical chloroplast reading frames (ycf) |
| RubisCO large subunit | transfer RNAs |
| RNA polymerase | ribosomal DNAs |

Figure C.5 Mauve-based alignment comparing mycoheterotrophic *Campylosiphon congestus* and autotrophic *Burmannia bicolor* (Burmanniaceae) (a linear-map of *B. bicolor* appears first for reference see Fig. 4.1). A single copy of the inverted repeat region was included in this comparison. The location of a second copy of a duplicated IR copy is shown by the red arrow. Coloured blocks have shared gene order between genomes, referred to as ‘locally colinear blocks’ (LCBs). LCBs appearing above the central line for the mycoheterotroph are colinear and in the same orientation as the reference sequence; those below align in reverse complement. Coloured lines link LCBs shared between taxa.

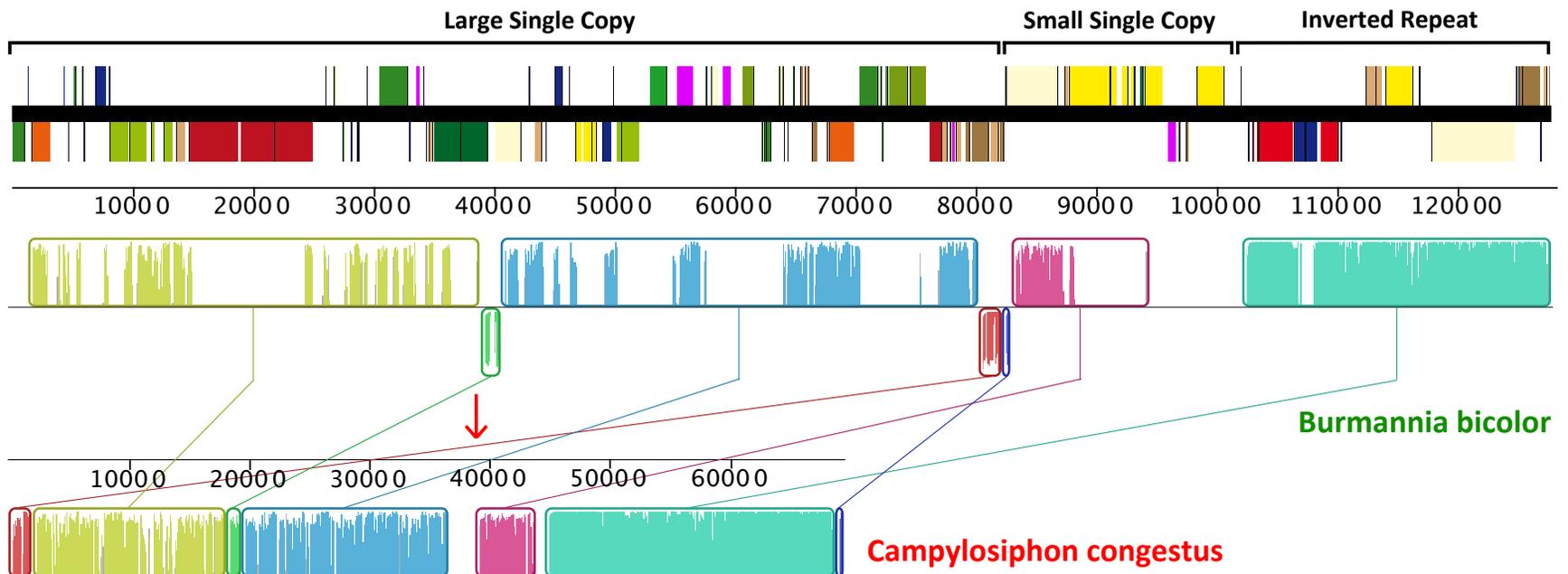


Figure C.6 Mauve-based alignment comparing mycoheterotrophic *Burmannia itoana* and autotrophic *B. bicolor* (Burmanniaceae) (a linear-map of *B. bicolor* appears first for reference, see Fig. 4.1). A single copy of the inverted repeat region was included in this comparison. Coloured blocks have shared gene order between genomes, referred to as ‘locally colinear blocks’ (LCBs). LCBs appearing above the central line for the mycoheterotroph are colinear and in the same orientation as the reference sequence; those below align in reverse complement. Coloured lines link LCBs shared between taxa.

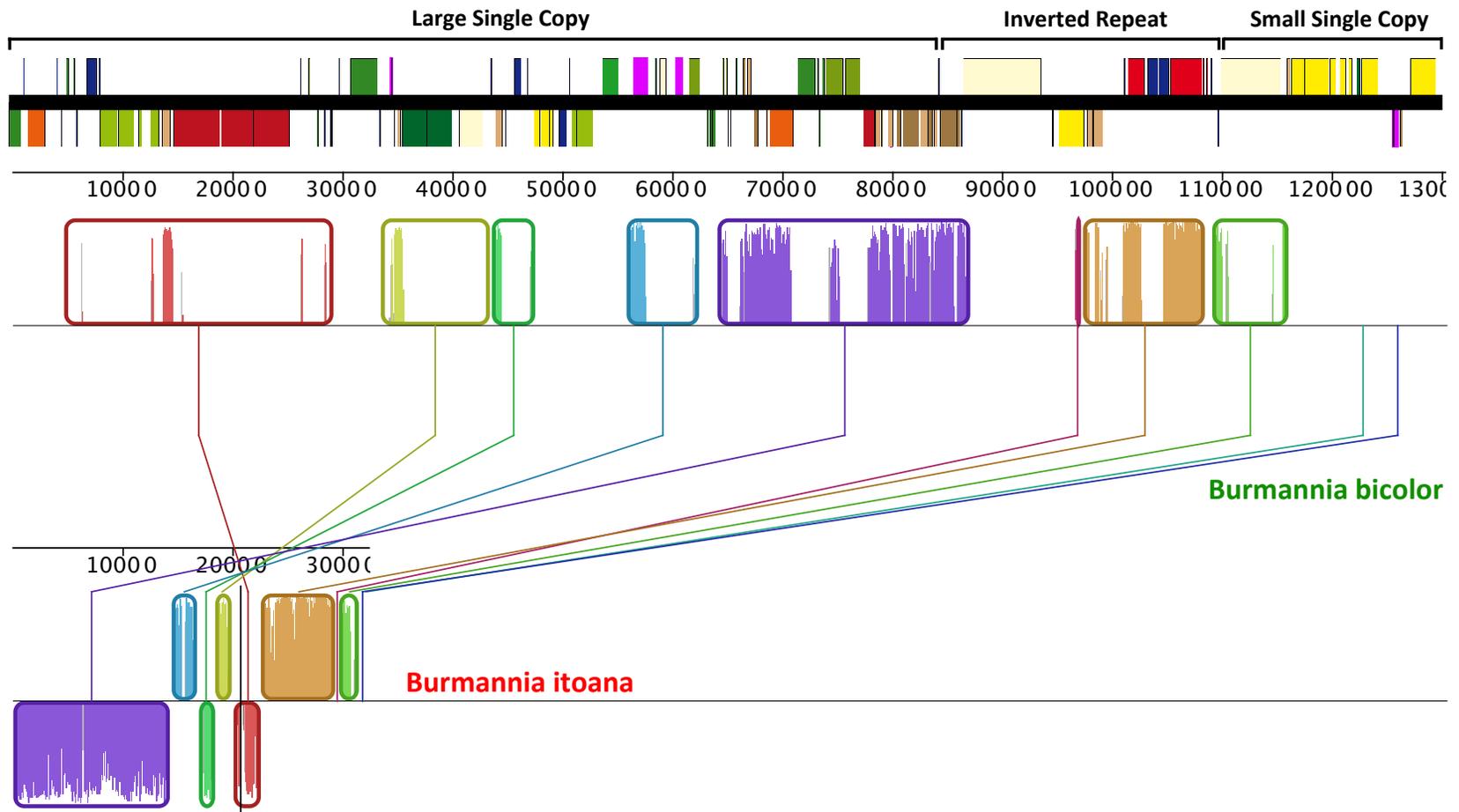


Figure C.7 Mauve-based alignment comparing mycoheterotrophic *Gymnosiphon longistylus* and autotrophic *Burmannia bicolor* (Burmanniaceae) (a linear-map of *B. bicolor* appears first for reference, see Fig. 4.1). A single copy of the inverted repeat region was included in this comparison. Coloured blocks have shared gene order between genomes, referred to as ‘locally colinear blocks’ (LCBs). LCBs appearing above the central line for the mycoheterotroph are colinear and in the same orientation as the reference sequence; those below align in reverse complement. Coloured lines link LCBs shared between taxa.

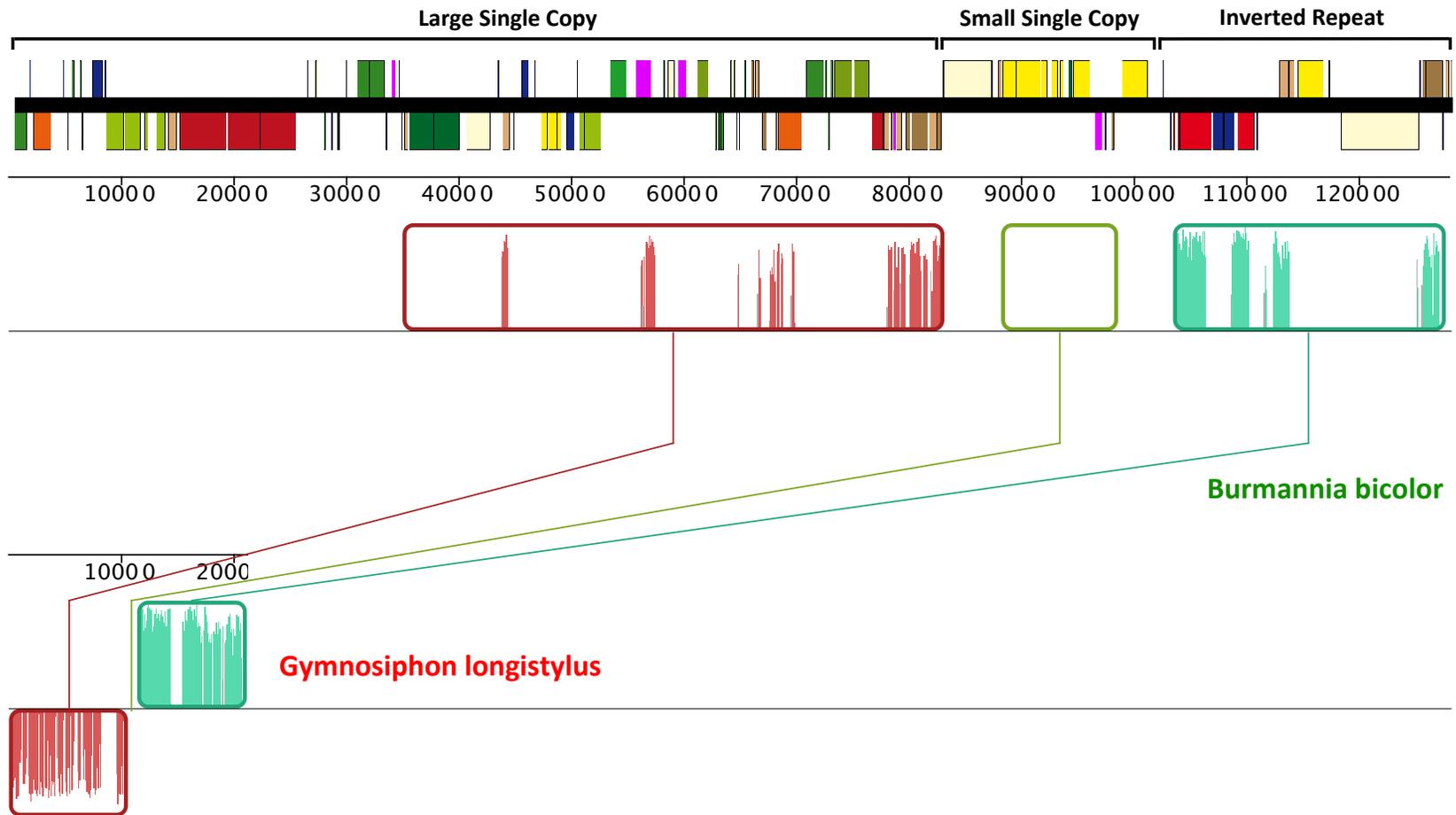


Figure C.8 Mauve-based alignment comparing mycoheterotrophic *Apteria aphylla* and autotrophic *Burmannia bicolor* (Burmanniaceae) (a linear-map of *B. bicolor* appears first for reference, see Fig. 4.1). A single copy of the inverted repeat region was included in this comparison. Coloured blocks have shared gene order between genomes, referred to as ‘locally colinear blocks’ (LCBs). LCBs appearing above the central line for the mycoheterotroph are colinear and in the same orientation as the reference sequence, those below align in reverse complement. Coloured lines link LCBs shared between taxa.

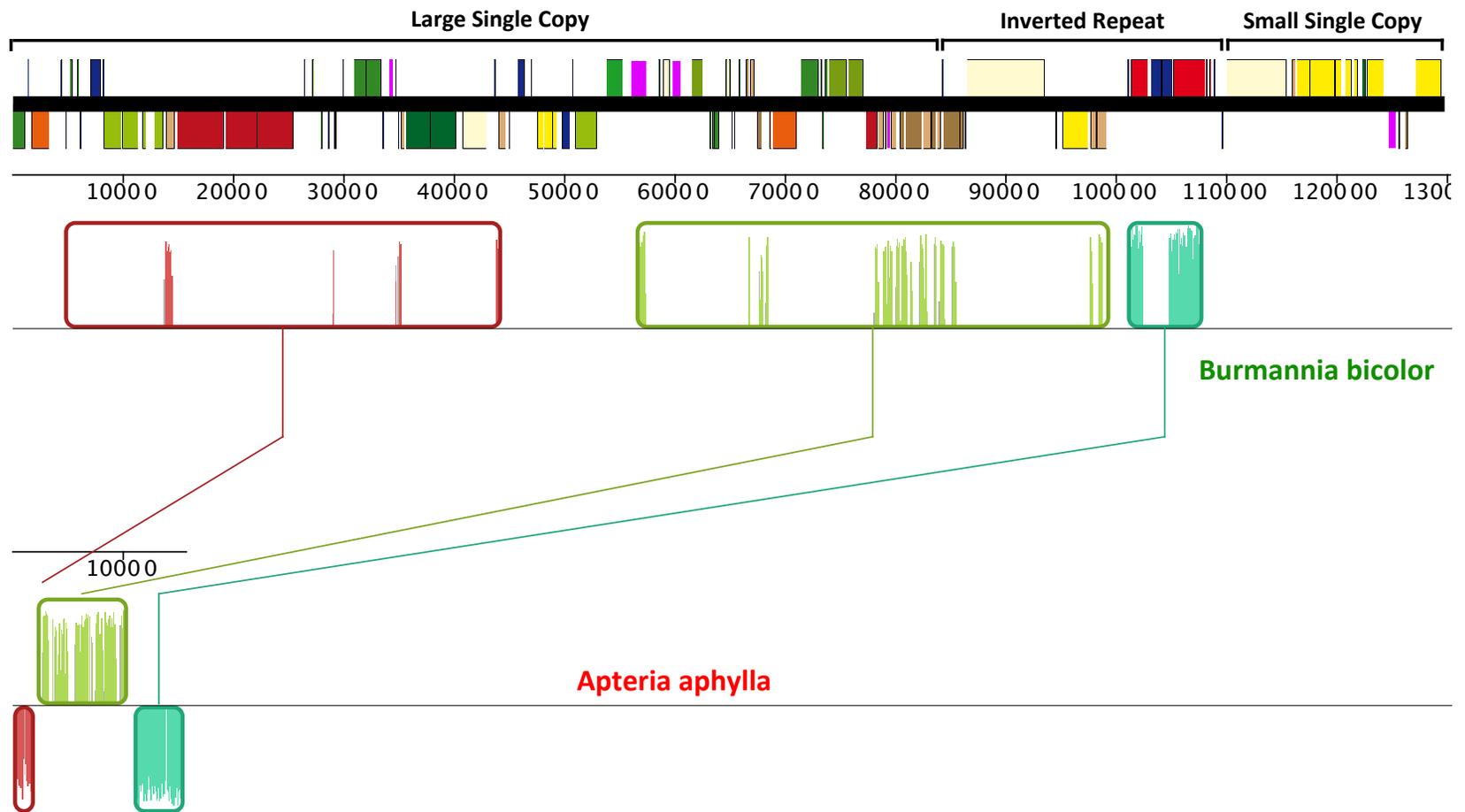


Figure C.9 Mauve-based alignment comparing mycoheterotrophic *Corsia cf. boridiensis* (Corsiaceae) and autotrophic *Campynema lineare* (Campynemataceae) (a linear-map of *C. lineare* appears first for reference, see Fig. C.1). A single copy of the inverted repeat region was included in this comparison. Coloured blocks have shared gene order between genomes, referred to as ‘locally colinear blocks’ (LCBs). LCBs appearing above the central line for the mycoheterotroph are colinear and in the same orientation as the reference sequence, those below align in reverse complement. Coloured lines link LCBs shared between taxa.

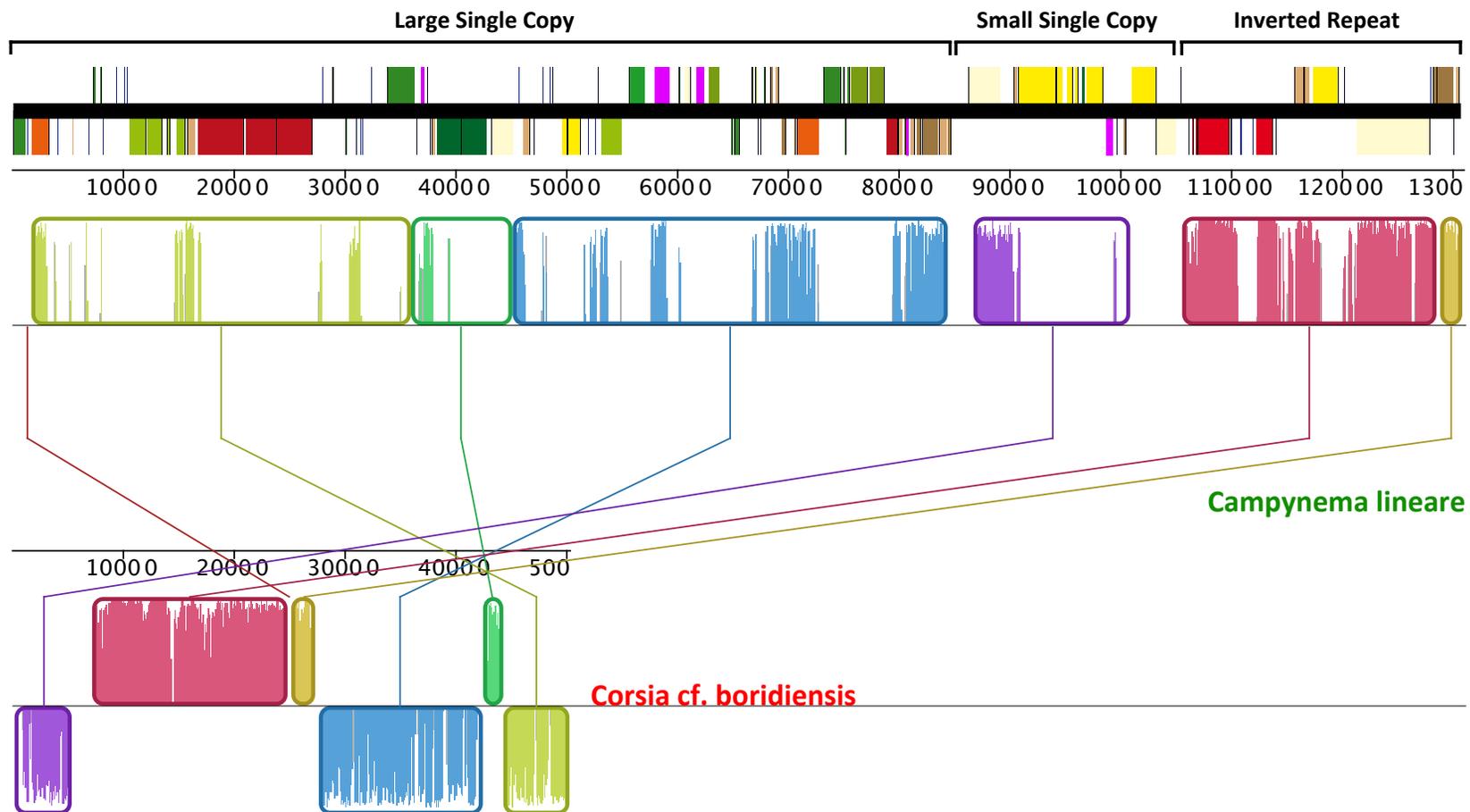


Figure C.10 Mauve-based alignment comparing mycoheterotrophic *Arachnitis uniflora* (Corsiaceae) and autotrophic *Campynema lineare* (Campynemataceae) (a linear-map of *C. lineare* appears first for reference, see Fig. C.1). A single copy of the inverted repeat region was included in this comparison. Coloured blocks have shared gene order between genomes, referred to as ‘locally colinear blocks’ (LCBs). LCBs appearing above the central line for the mycoheterotroph are colinear and in the same orientation as the reference sequence, those below align in reverse complement. Coloured lines link LCBs shared between taxa.

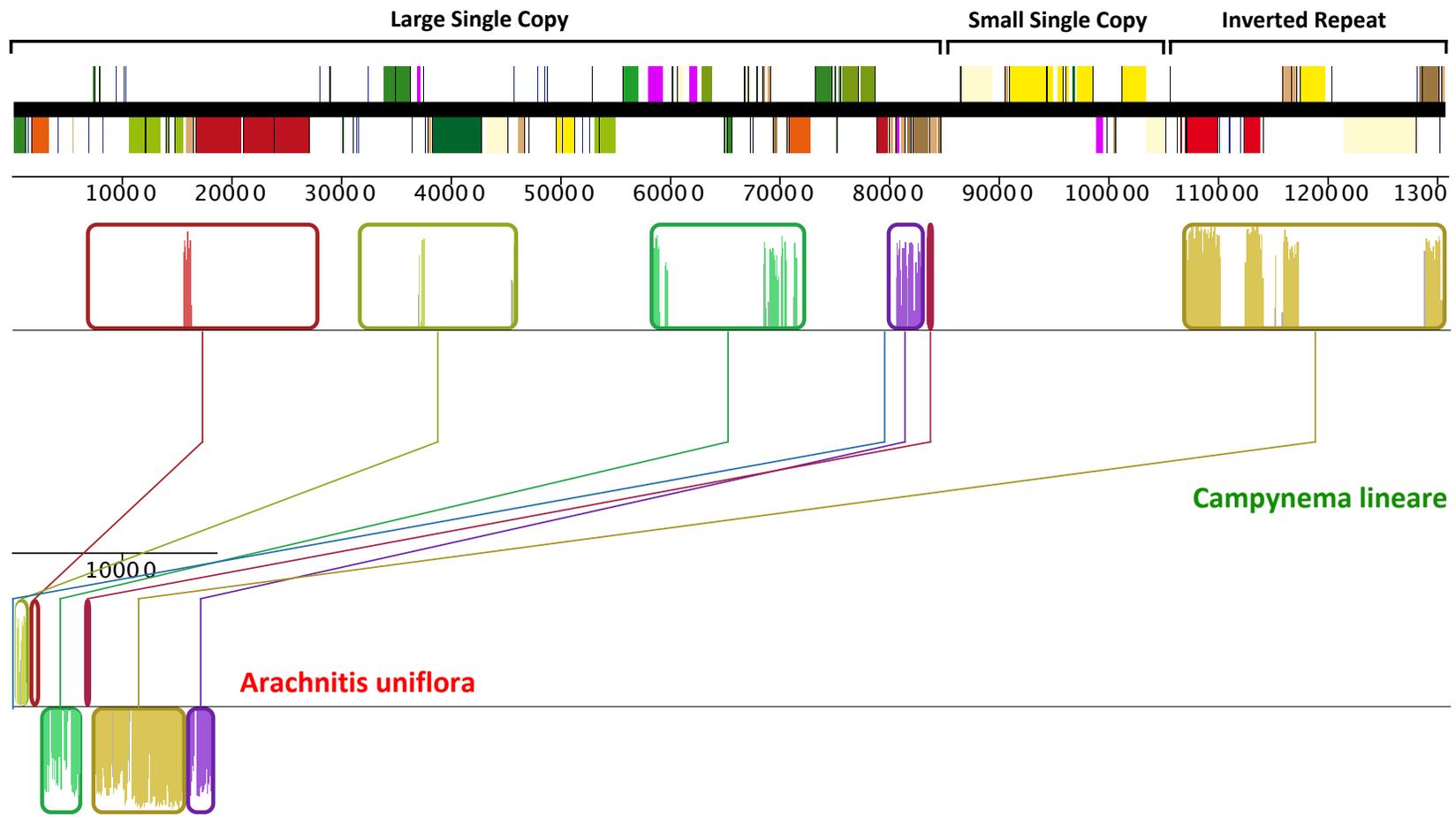


Figure C.11 Mauve-based alignment comparing mycoheterotrophic *Geosiris aphylla* and autotrophic *Iris missouriensis* (Iridaceae) (a linear-map of *I. missouriensis* appears first for reference, see Fig. C.2). A single copy of the inverted repeat region was included in this comparison. Coloured blocks have shared gene order between genomes, referred to as ‘locally colinear blocks’ (LCBs). LCBs appearing above the central line for the mycoheterotroph are colinear and in the same orientation as the reference sequence; those below align in reverse complement. Coloured lines link LCBs shared between taxa.

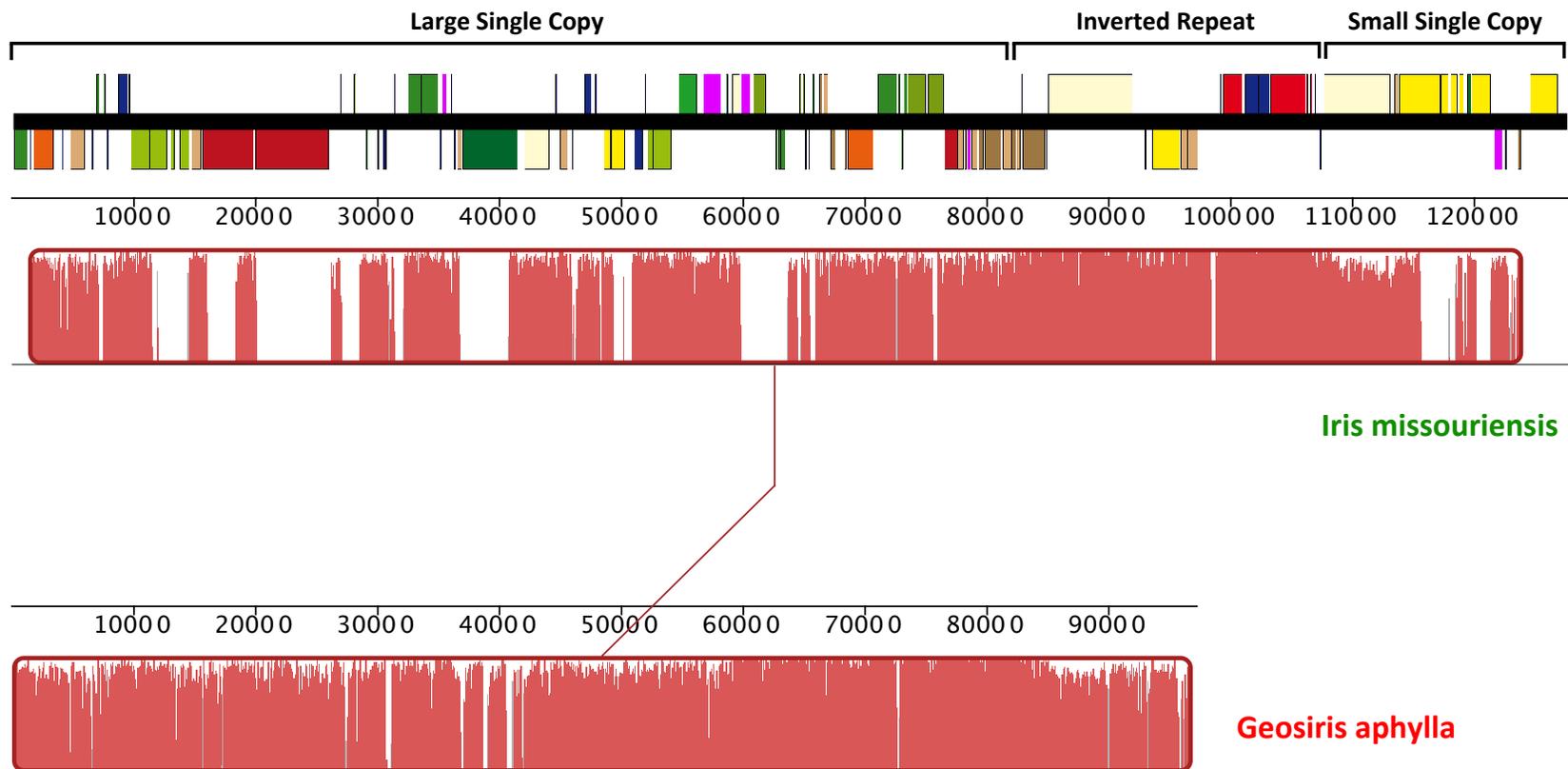


Figure C.12 Mauve-based alignment comparing mycoheterotrophic *Petrosavia* spp. and autotrophic *Japonolirion osense* (Petrosaviaceae) (a linear-map of *J. osense* appears first for reference, see Fig. C.3). A single copy of the inverted repeat region was included in this comparison. Coloured blocks have shared gene order between genomes, referred to as ‘locally colinear blocks’ (LCBs). LCBs appearing above the central line for the mycoheterotroph are colinear and in the same orientation as the reference sequence; those below align in reverse complement. Coloured lines link LCBs shared between taxa.

Petrosaviaceae

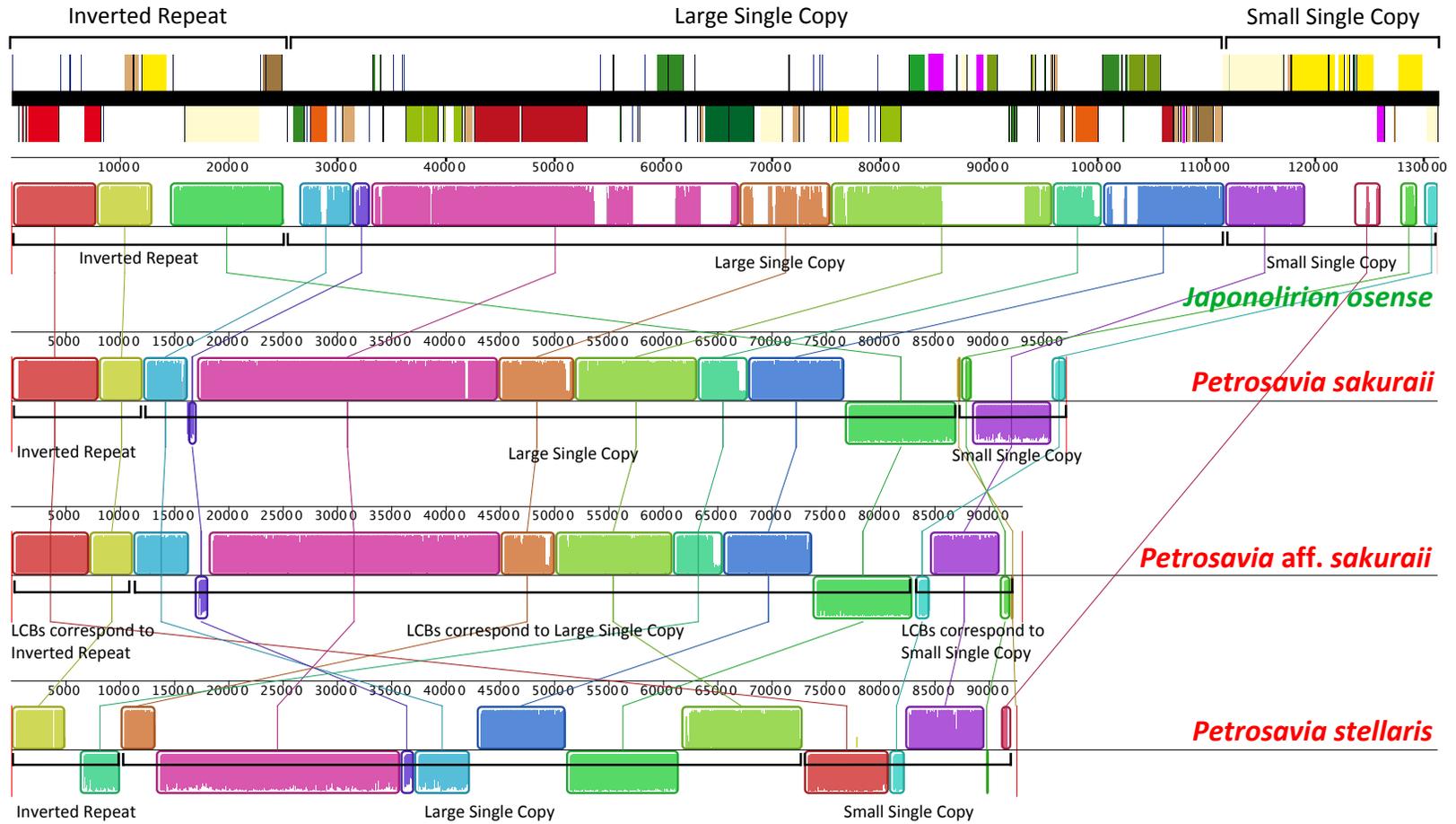


Figure C.13 Mauve-based alignment comparing autotrophic *Burmannia capitata* and *Burmannia bicolor* (Burmanniaceae) (a linear-map of *B. bicolor* appears first for reference, see also Fig. 4.1). A single copy of the inverted repeat region was included in this comparison. Both genomes share the same gene order.

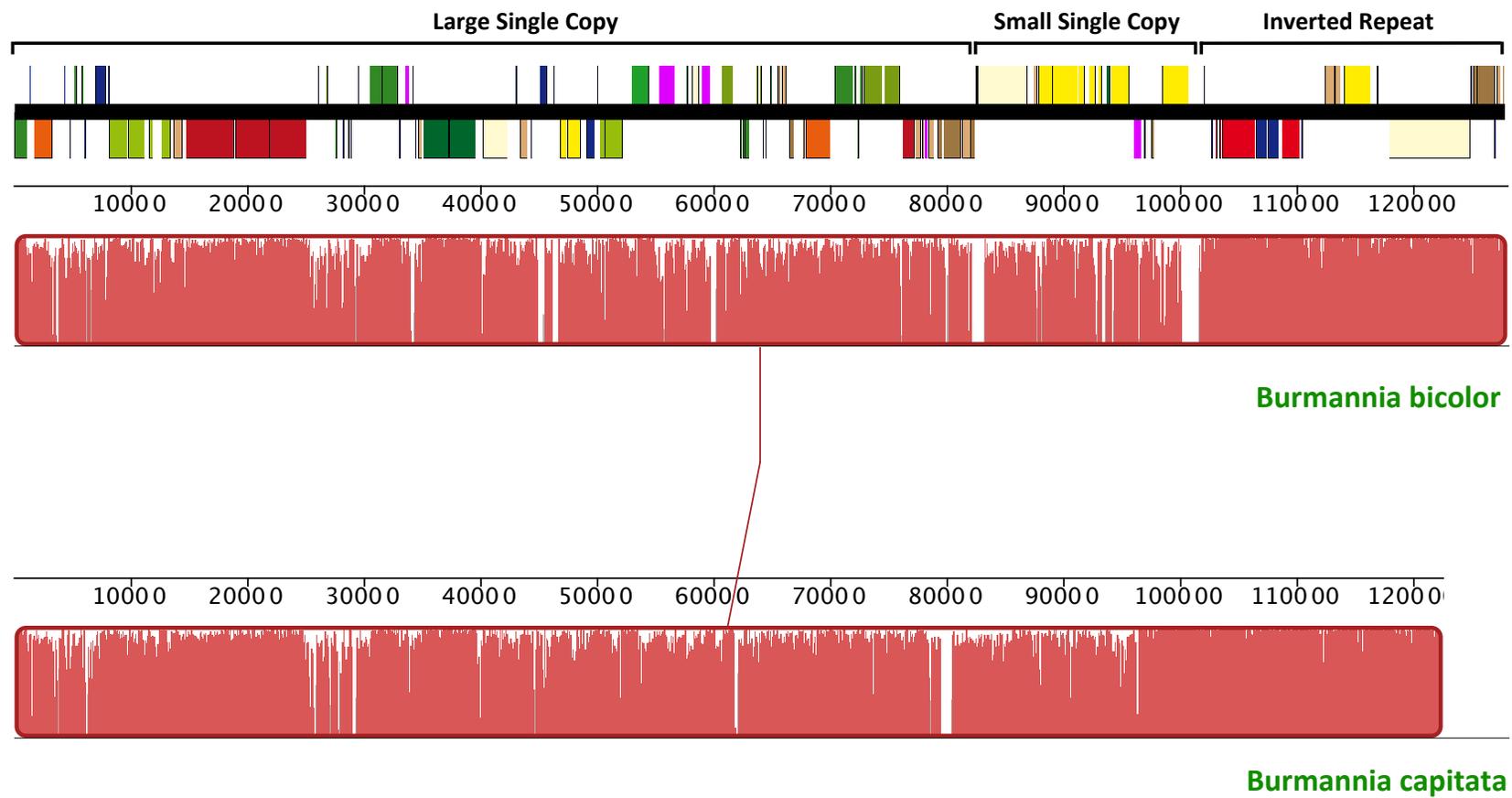


Figure C.14 Mauve-based alignment comparing autotrophic *Sciaphila densiflora* (Triuridaceae) and *Carludovica palmata* (Cyclanthaceae) (a linear-map of *C. palmata* appears first for reference, see also Fig. 3.3). A single copy of the inverted repeat region was included in this comparison. Coloured blocks have shared gene order between genomes, referred to as ‘locally colinear blocks’ (LCBs). LCBs appearing above the central line for the mycoheterotroph are colinear and in the same orientation as the reference sequence; those below align in reverse complement. Coloured lines link LCBs shared between taxa.

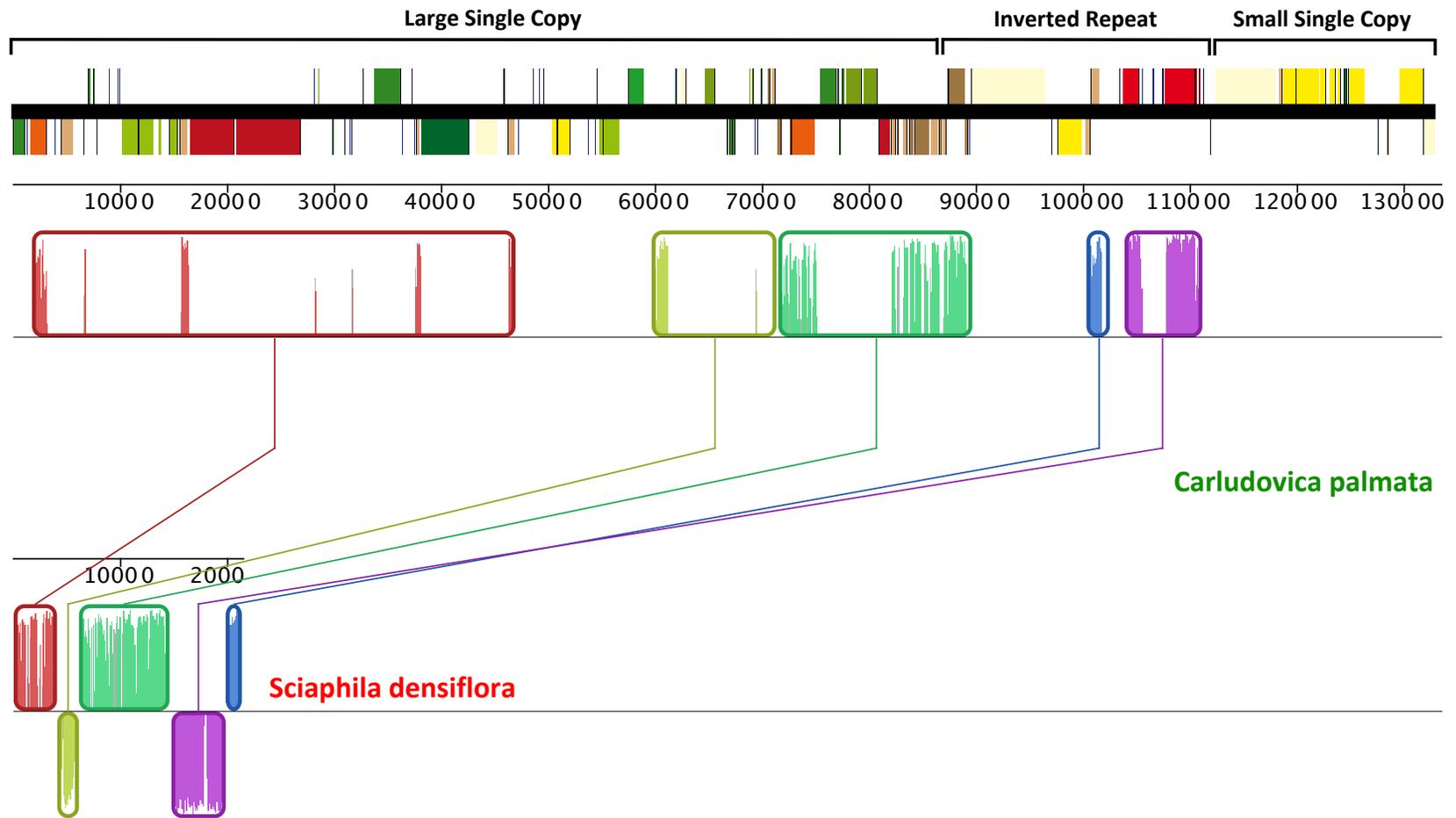


Figure C.15 Phylogeny of photosynthetic and mycoheterotrophic monocots based on an unpartitioned maximum likelihood analysis of 83 plastid genes. Thick lines indicate branches with 100% bootstrap support, those with less than 100% bootstrap support are indicated numerically, with values <50% indicated with a dash ('-'). Lineages outside orchids with mycoheterotrophs are indicated in red. The scale bar indicates inferred substitutions per site.

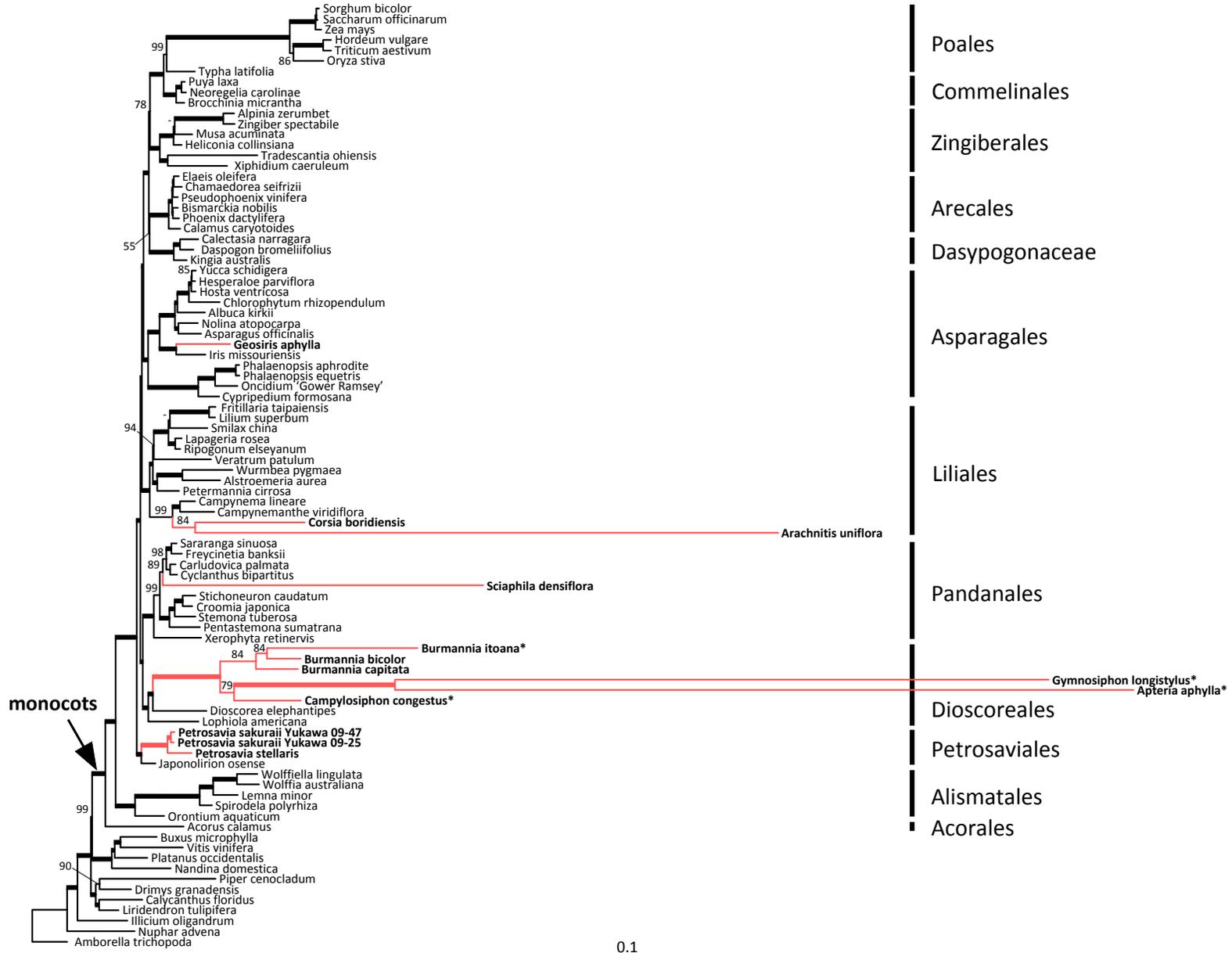
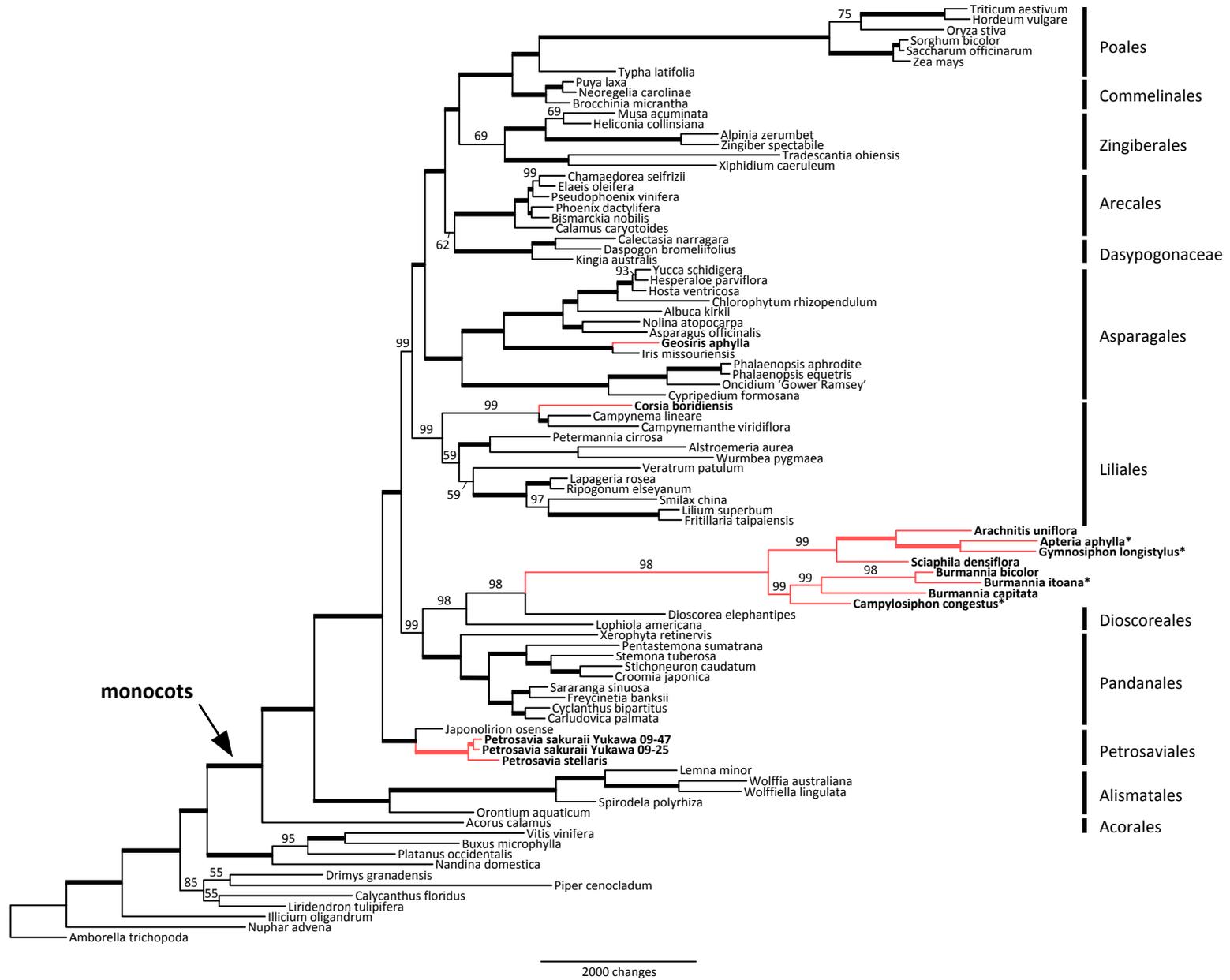


Figure C.16 Phylogeny of photosynthetic and mycoheterotrophic monocots based on a parsimony analysis of 83 plastid genes. Thick lines indicate branches with 100% bootstrap support, those with less than 100% bootstrap support are indicated numerically, and values <50% are indicated with a dash ('-'). Lineages outside orchids with mycoheterotrophs are indicated in red. The scale bar indicates inferred number of changes.



Appendix D: Supplementary tables and figures for Chapter 5

Table D.1 Voucher and Genbank information for taxa used in this study. Newly sequenced taxa specimen source information; herbarium abbreviations follow Thiers (continually updated). Genbank accessions for previously published plastome sequences included in this study.¹

Species	Family	Voucher/Study	Genbank Accession
<i>Manoao colensoi</i> (Hook.) Molloy	Podocarpaceae	T. Brodribb, living population, HO ²	XXXXXXXXXX
<i>Lepidothamnus laxifolius</i> (Hook.f.) Quinn	Podocarpaceae	T. Brodribb, living population, HO ²	XXXXXXXXXX
<i>Parasitaxus usta</i> (Vieill.) de Laub.	Podocarpaceae	M. Hollingsworth 2107, no voucher	XXXXXXXXXX
<i>Parasitaxus usta</i> (Vieill.) de Laub.	Podocarpaceae	M. Hollingsworth 2170, E ²	XXXXXXXXXX
<i>Parasitaxus usta</i> (Vieill.) de Laub.	Podocarpaceae	M. Hollingsworth 2171, E ²	XXXXXXXXXX

¹Additional sequences: *Abies firma* Siebold & Zucc. (FJ899565.1), *Agathis dammara* (Lamb.) Rich. (NC_023119.1), *Araucaria montana* Brongn. & Gris (KM678422.1), *Bowenia serrulata* (W.Bull) Chamb. (NC_026036.1), *Calocedrus formosana* (Florin) Florin (NC_023121.1), *Cathaya argyrophylla* (NC_014589.1), *Cedrus deodara* (Roxb. Ex D.Don) G.Don (NC_014575.1), *Cephalotaxus harringtonii* (Knight ex J.Forbes) K.Koch (NC_016063.1), *Ceratozamia hildae* G.P. Landry & M.C. Wilson (NC_026037.1), *Cryptomeria japonica* (Thunb. Ex L.f.) D.Don (NC_010548.1), *Cunninghamia lanceolata* (Lamb.) Hook. (NC_021437.1), *Cycas revoluta* Thunb. (NC_020319.1), *Dioon spinulosum* Dyer ex Eichl. (JX512656.1), *Ginkgo biloba* L. (NC_016986.1), *Juniperus virginiana* L. (NC_024024.1), *Keteleeria davidiana* (C.E.Bertrand) Beissn. (NC_011930.1), *Larix decidua* Mill. (NC_016058.1), *Nageia nagi* (Thunb.) Kuntze (NC_023120.1), *Picea abies* (L.) H.Karst. (NC_021456.1), *Pinus contorta* Douglas ex Loudon (NC_011153.4), *Pinus nelsonii* Shaw (NC_011159.4), *Pinus pinea* L. (JN854173.1), *Pinus strobus* L. (FJ899560.1), *Pinus thunbergii* Parl. (NC_001631.1), *Podocarpus lambertii* Klotzsch ex Endl. (NC_023805.1), *Podocarpus totara* G.Benn. ex D.Don (NC_020361.1), *Pseudotsuga menziesii* (Mirb.) Franco (JN854170.1), *Retrophyllum piresii* (Silba) C.N.Page (KJ617081.1), *Sciadopitys verticillata* (Thunb.) Siebold & Zucc. (AB645770.1- AB645822.1), *Stangeria eriopus* (Kunze) Baill. (NC_026041.1), *Taiwania cryptomerioides* Hayata (NC_016065.1), *Taxus mairei* (Lemée & H.Lév.) S.Y.Hu (KJ123824.1)

²No vouchers were made for these samples from living populations. Living populations can be found at University of Tasmania, Hobart (HO). For *Parasitaxus*, photos with locality, date and collecting number are available at the herbarium at the Royal Botanic Gardens, Edinburgh (E).

Table D.2 Data partitioning schemes inferred using PartitionFinder or PartitionFinder Protein with the BIC criterion. (a) “GxC” (gene x codon) partitioning scheme of the 80-gene plastome dataset (see text for details). Plastid genes are indicated before the underscore; the ‘pos’ term after an underscore indicates the codon position, the ‘exon’ position indicates genes with exons.

Partition no.	Best Model	Partition subsets
1	GTR+I+G	accD_pos1, clpP_pos1, psaM_pos1, psbM_pos1
2	GTR+I+G	accD_pos2, clpP_pos2, petL_pos1, rpl14_pos3, rpl16_pos3, rpl23_pos1, rpl33_pos1, rps12_pos3, rps18_pos1, rps19_pos2, rps2_pos3
3	GTR+I+G	accD_pos3, cemA_pos3, chlN_pos3, clpP_pos3, infA_pos3, matK_pos3, psaM_pos3, rpl22_pos3, rpl23_pos3, rpl32_pos3, rpoA_pos3, rpoC2_pos3, rps18_pos3, rps19_pos3, rps3_pos3, rps4_pos3, rps7_pos3
4	GTR+I+G	atpA_pos1, atpB_pos1, chlL_pos1, ndhH_pos1, psaA_pos1, psaB_pos1, psbB_pos1, psbE_pos1, rbcL_pos1, rps12_pos1, ycf3_pos1
5	GTR+I+G	atpA_pos2, atpB_pos2, chlB_pos2, chlL_pos2, chlN_pos2, ndhH_pos2, ndhI_pos2, ndhJ_pos2, petA_pos2, petD_pos2, psbF_pos1, psbI_pos1, psbL_pos1, psbL_pos2, psbT_pos1, rpl14_pos2, ycf3_pos2
6	GTR+G	atpA_pos3, atpB_pos3, atpE_pos3, atpF_pos3, atpI_pos3, ccsA_pos3, chlB_pos3, chlL_pos3, lhbA_pos3, ndhA_pos3, ndhB_pos3, ndhC_pos3, ndhD_pos3, ndhF_pos3, ndhG_pos3, ndhH_pos3, ndhI_pos3, ndhJ_pos3, ndhK_pos3, petA_pos3, petD_pos3, petL_pos3, psaI_pos1, psaI_pos3, psaJ_pos3, psbH_pos3, psbJ_pos3, psbK_pos3, psbM_pos3, rpl20_pos3, rpl2_pos3, rpl33_pos3, rpoB_pos3, rpoC1_pos3, rps11_pos3, rps14_pos3, rps15_pos3, rps8_pos3, ycf4_pos3
7	GTR+G	atpE_pos1, atpF_pos1, infA_pos1, rpl22_pos1, rpoA_pos1, rpoB_pos1, rpoC1_pos1, rpoC2_pos1, rps11_pos1, rps19_pos1, rps2_pos1, rps3_pos1, rps7_pos1, rps8_pos1
8	GTR+G	atpE_pos2, infA_pos2, psbH_pos1, rpl20_pos1, rpl22_pos2, rpl23_pos2, rpl32_pos2, rpl36_pos1, rps14_pos1, rps15_pos1, rps2_pos2, rps3_pos2, rps4_pos1, rps7_pos2
9	GTR+G	atpF_pos2, cemA_pos2, matK_pos2, ndhK_pos2, psbF_pos3, rpl20_pos2, rpoA_pos2, rpoB_pos2, rpoC1_pos2, rpoC2_pos2, rps15_pos2, rps4_pos2
10	GTR+G	atpH_pos1, petB_pos1, petG_pos1, psaC_pos1, psaC_pos2, psbA_pos1, psbC_pos1, psbD_pos1, psbN_pos1
11	GTR+G	atpH_pos2, petB_pos2, psaA_pos2, psaB_pos2, psbA_pos2, psbB_pos2, psbC_pos2, psbD_pos2, psbE_pos2, psbN_pos2, rbcL_pos2
12	GTR+G	atpH_pos3, ndhE_pos3, petB_pos3, petG_pos3, petL_pos2, psaA_pos3, psaB_pos3, psaC_pos3, psaJ_pos1, psbA_pos3, psbB_pos3, psbC_pos3, psbD_pos3, psbE_pos3, psbI_pos3, psbL_pos3, psbN_pos3, rbcL_pos3, rpl36_pos3, ycf3_pos3
13	GTR+G	atpI_pos1, chlB_pos1, chlN_pos1, lhbA_pos1, ndhA_pos1, ndhB_pos1, ndhC_pos1, ndhD_pos1, ndhE_pos1, ndhI_pos1,

Partition no.	Best Model	Partition subsets
14	GTR+I+G	ndhJ_pos1, petA_pos1, petD_pos1, petN_pos1, rpl14_pos1, rpl16_pos1, rpl16_pos2, rpl2_pos1, rpl2_pos2, rpl33_pos2, rpl36_pos2, rps11_pos2, rps12_pos2, rps14_pos2, rps8_pos2, ycf4_pos1, ycf4_pos2
15	GTR+G	atpI_pos2, lhbA_pos2, ndhA_pos2, ndhB_pos2, ndhC_pos2, ndhD_pos2, ndhE_pos2, petG_pos2, petN_pos2, psbF_pos2, psbI_pos2, psbJ_pos2, psbT_pos2
16	GTR+G	ccsA_pos1, cemA_pos1, matK_pos1, ndhF_pos1, ndhG_pos1, ndhK_pos1, psbJ_pos1, psbK_pos1, psbM_pos2, psbT_pos3, rpl32_pos1, rps18_pos2
		ccsA_pos2, ndhF_pos2, ndhG_pos2, petN_pos3, psaI_pos2, psaJ_pos2, psaM_pos2, psbH_pos2, psbK_pos2

Table D.3 Overview of plastome sizes and characteristics of gymnosperms included in this study.

Species	Size (bp)	Inverted Repeat? (total) ^a	No. of unique genes protein/rDNA/tRNA	Coding percentage ^b	GC content (percentage)
<i>Ginkgo biloba</i>	156,988	Yes	80 / 4 / 32 (116)	56.6	39.0
<i>Manoao colensoi</i>	135,808	No	81 / 4 / 32 (117)	59.6	36.5
<i>Lepidothamnus laxifolius</i>	130,500	No	81 / 4 / 32 (117)	62.4	37.3
<i>Nageia nagi</i>	133,722	No	81 / 4 / 32 (117)	61.2 ^b	37.3 ^c
<i>Parasitaxus usta</i>	83,156	Yes	33 / 4 / 31 (68)	60.8	37.6
<i>Podocarpus lambertii</i>	133,734	No	81 / 4 / 31 (116)	58.7	37.1

^a Excludes pseudogenes. 5'-*rps12* and 3'-*rps12* counted as one gene, refer to Table 5.1 for more details for *Manoao* and *Parasitaxus*

^b Includes rDNAs and tRNAs, includes only one inverted repeat for *Ginkgo* and *Parasitaxus*

^c Values obtained from Wu and Chaw (2014)

Table D.4 Indel variation between three accessions of *Parasitaxus usta*, as identified by C-Sibelia. Two accessions of *Parasitaxus* (2170, 2171) belong to the same population and 2107 from a different population (see text for details). Number refers to the position of the indel in *Parasitaxus usta* 2170, which was used as the reference plastome in comparison to (a) *Parasitaxus usta* 2107 and (b) *Parasitaxus usta* 2171.

a)

Position in <i>P. usta</i> 2170	Indel in <i>P. usta</i> 2170	Indel in <i>P. usta</i> 2107
7247	T	TT
9497	TT	T
10247	TT	T
21719	TTT	T
22867	T	TGATAT
23103	ACATAA	AAGAT
24470	T	TTT
25200	TT	T
26387	AA	A
27855	AA	A
28468	GTAGG	G
23730	T	TTT
29068	T	TT
34152	T	TT
34979	T	TT
35557	T	TT
36444	T	TTTCAT
36828	T	TTT
37766	GGATTCGTCATTCTCGGA TTTGTCATTCTCA	GTCATTCTCGGATTCGTC ATTCTCG
39011	TTTTT	T
39272	G	GGG
40393	T	TT
40840	AA	A
41090	A	AA
42038	T	TT
54320	G	GTATACTATA
54387	AAT	ATAA
54443	AACTAA	A
61499	ATNNNNNNA	ANNNNNN
63323	TG	T
64393	A	AA
68646	TT	T
70234	T	TGATTT
70484	CGNNNNNNNN	CCNNNNNNNTGG
72442	C	CTCCTC
72594	T	TT
72744	TCAGAAAAAAAAAAT	TTCAGAAAAAAAAAA
74172	TAGGATTAATCATGAACG AGTCATTAACACTACTGG GAAAAAAAAAGAACTGCAAA	TCCGATTTATAAT

a)

Position in <i>P. usta</i> 2170	Indel in <i>P. usta</i> 2170	Indel in <i>P. usta</i> 2107
	ACCGGATCCGATTTATAAT	
74976	T	TT
75217	ATA	A
79779	AAA	ATT
82298	GCGGTTCCAAAGTACGAG	GTGGTTCCAAAGTA
	GCATTTTATG	CGAGGCATTTTATT
82369	TTAGTT	T
82422	TAT	TTTA
82482	GNNNNNNNNN	GTATAGTATA

b)

Position in <i>P. usta</i> 2170	Indel in <i>P. usta</i> 2170	Indel in <i>P.usta</i> 2171
61499	ATNNNNNNA	ANNNNNN
70484	CGNNNNNNNN	CCCCNNNNNN

Figure D.1 Pairwise Mauve-based alignments of *Ginkgo* with a) *Podocarpus*, b) *Manoao* and c) *Parasitaxus* (a linear map of *Ginkgo biloba* appears first for reference), respectively. A single copy of the inverted repeat region was included for *Parasitaxus* for comparison. Coloured blocks are homologous regions with shared gene orders between two or more genomes, referred to as ‘locally colinear blocks’ (LCB). LCBs above the line are colinear and in the same orientation as the reference sequence, those below are reverse complements. Coloured lines link blocks of homology shared between taxa.

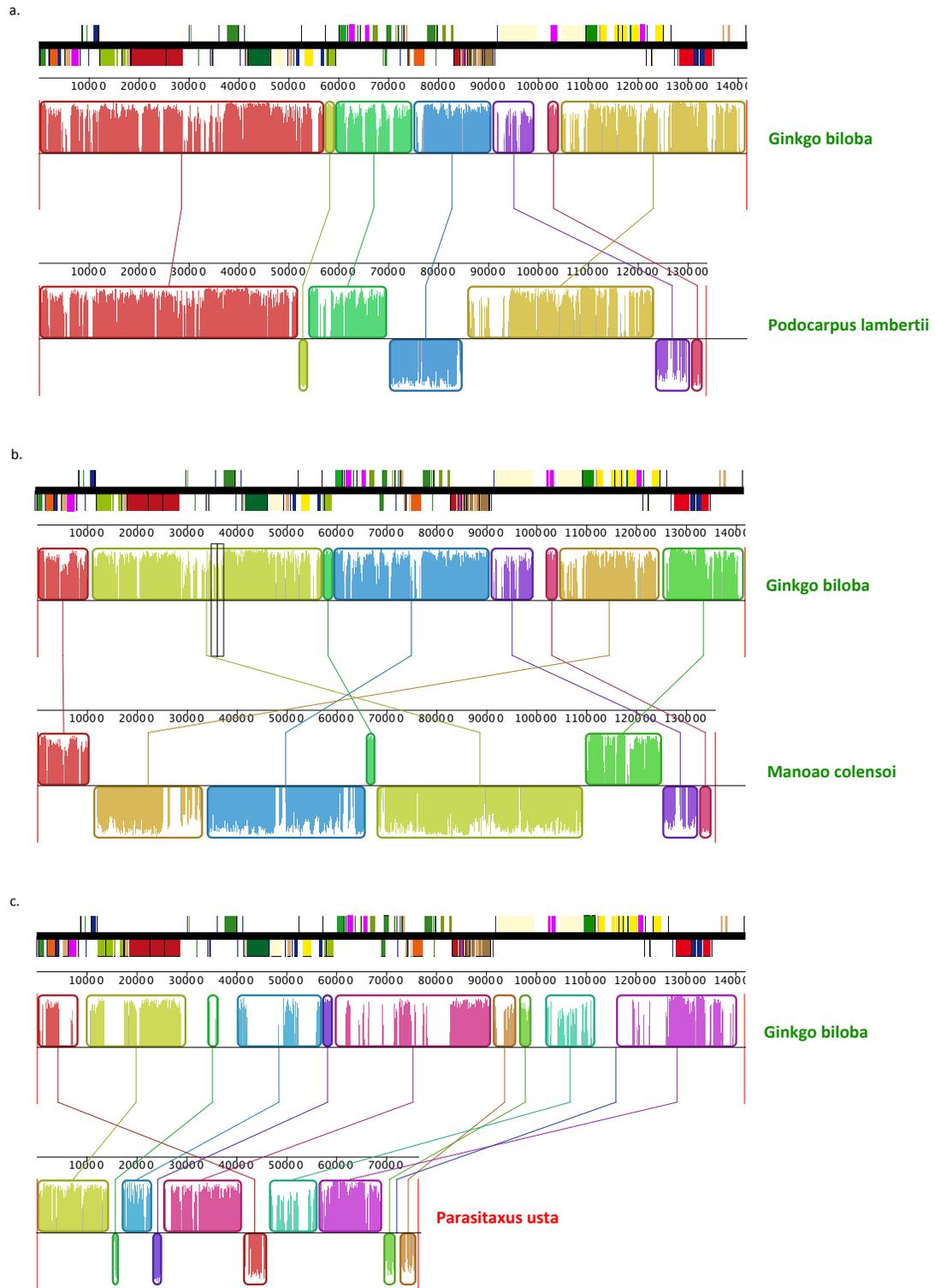


Figure D.2 BRIG-based diagram indicating the positions of nucleotide substitutions and indels between two accessions of *Parasitaxus usta* (Podocarpaceae; accession 2170, outer ring; 2107, inner ring). Mutations are shown as thickened tick marks between the outermost blue ring and innermost ring (tick marks not to scale). Two thick gray bars indicate the position of inverted repeats (IR). The innermost ring indicates the positions of plastid genes (shown as black segments); the gene labels are as indicated in the outermost blue ring.

Figure D.3 Gymnosperm phylogeny inferred in an unpartitioned maximum-likelihood analysis of 80 plastid genes (the best tree represented as a phylogram). Bootstrap support values are indicated beside branches: thick lines indicate branches with 100% bootstrap support. The scale bar indicates estimated substitutions per site.

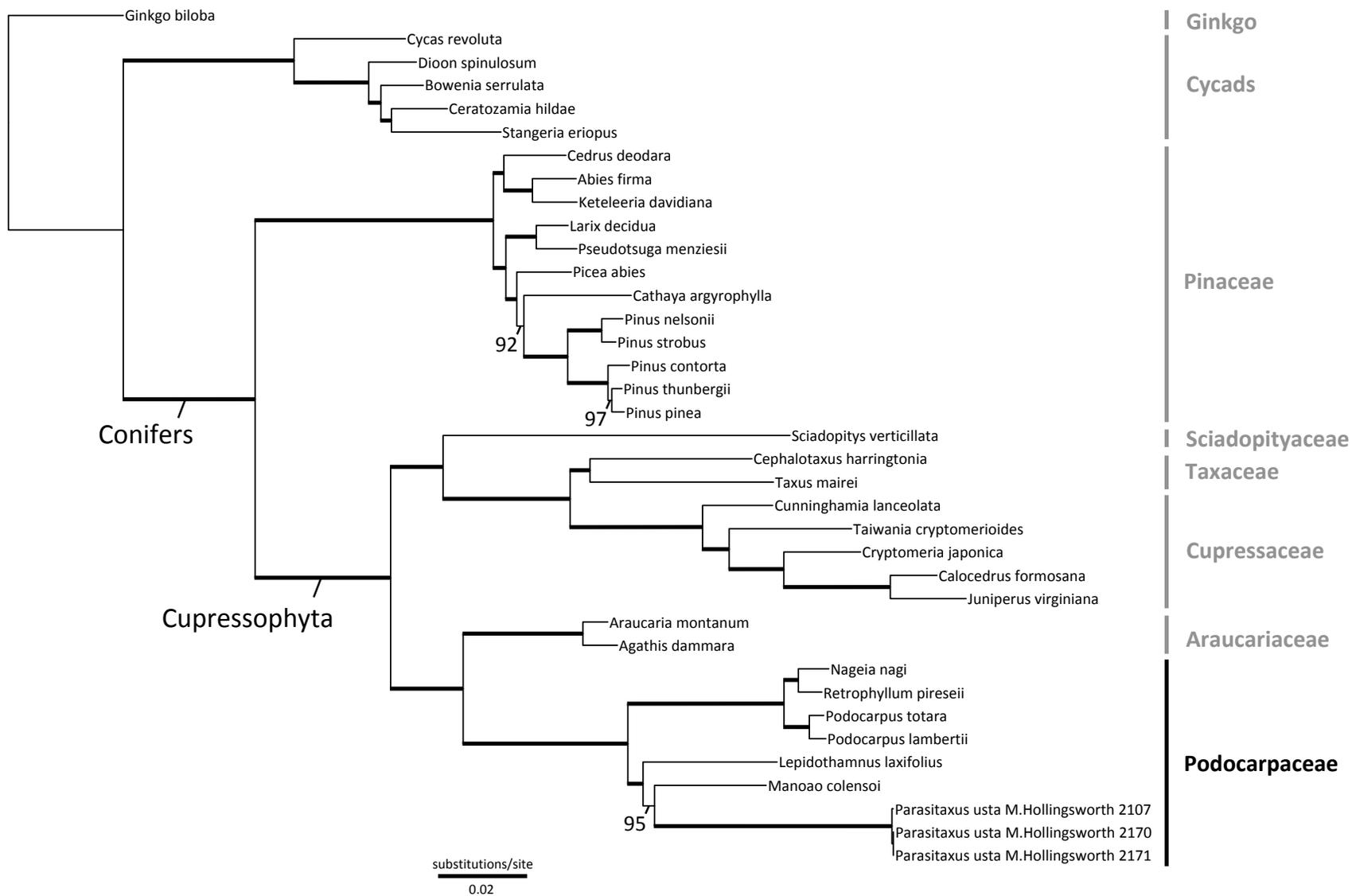


Figure D.4 Gymnosperm phylogeny based on the best tree inferred from a parsimony analysis of 80 plastid genes. Bootstrap support values are indicated beside branches: thick lines indicate branches with 100% bootstrap support. The scale bar indicates the number of changes.

