universität wien

DISSERTATION

Titel der Dissertation

„Assessing the Evolutionary Patterns of Plastid Genome Reduction in a Group of Non-Photosynthetic Parasitic Angiosperms (Orobanchaceae)“

Verfasserin

Dipl.-Biol. Susann Wicke

angestrebter akademischer Grad

Doktorin der Naturwissenschaften (Dr.rer.nat.)

Wien, 2012

Studienkennzahl lt. Studienblatt:	A 091 438
Dissertationsgebiet lt. Studienblatt:	Dr.-Studium der Naturwissenschaften Botanik (Stzw)
Betreuerin / Betreuer:	Prof. Dr. Gerald M. Schneeweiss

CONTENTS

List of Figures

LIST OF TABLES

PREFACE AND ACKNOWLEDGEMENTS

Many people have contributed substantially to the successful accomplishment of my dissertation research. The impact of some of those has been more fundamental throughout the entire time. For that reason, I am going to mention them here first without diminishing the impacts and support of those named later, who have been equally important during this work.

First, I would like to express my deep thanks to my principal advisor Prof. Dr. Gerald Schneeweiss for supervision and continuous support during the past four years. He patiently (and in the end successfully) directed my attention to plastid genome evolution and parasitic plants, and gave me invaluable advices during various phases of this work. Besides, I am grateful for fruitful and critical discussions as well as essential and helpful comments on the various chapters of this thesis. I am grateful for his incredible patience and trust throughout my long sojourns in other research labs, and for providing the space and freedom to develop and tackle own ideas and research interests.

Owing to the project and frequent travelling, the number of people that I appreciate not only as important scientific advisors grew on a yearly basis. In equal measure, I am indebted to my co-advisors Prof. Dr. Dietmar Quandt (University of Bonn, Germany), Prof. Dr. Kai Müller (University of Muenster, Germany) and Prof. Dr. Claude dePamphilis (Pennsylvania State University, USA). All of them had fundamental impact during my PhD research with invaluable advices and suggestions, critical discussions and comments on various aspects of manuscripts, last-minute corrections, and – most importantly – patient and continuous support. Thanks also for relaxed and great evenings with fine wine and delicious foods!

At the time I started to work on this thesis, sequencing of plastid chromosomes from parasitic plants had been an adventure – and remained so up to now. The number of unknown parameters differs in several orders of magnitude from one species to the next making "gut-feeling" and "trial and error" the most suitable experimental and scientific approaches. Many people have contributed substantially to minimize the overall error rate and successfully manage experimental challenges and problems:

I am deeply indebted to Monika Ballmann and Karola Maul and people of the Nees-Institute for Plant Biodiversity (University of Bonn, Germany) for excellent technical assistance as well as continuous support and great working atmosphere at every time of the day (and night). Thanks also for the many exhilarant coffee breaks and "Spanish evenings"!!

I would like to thank Dr. Felix Grewe, Prof. Dr. Volker Knoop and Monika Polsakiewicz (all University of Bonn, Germany) for sharing their great experience regarding experimental procedures with me and helpful comments and suggestions, and nice hours inside and outside the lab.

I am very grateful to the PennState lab crew, especially Lena Landherr-Sheaffer, Dr. Norman J. Wickett, Eric Wafula, and Paul Ralph for innumerable helpful suggestions on wet lab and dry lab procedures and for making my trip to the U.S. an unforgettable experience.

I would like to gratefully acknowledge Dr. Thomas Münster, Yvonne Steinbüchel and Diana Kühn (all Max Plank Institute for Plant Breeding Research, Cologne, Germany) for assistance with robotics systems. Thanks are also due to Prof. Dr. Christoph Neinhuis (TU Dresden, Germany) and his working group for help during the very first phase of my PhD research.

Performing molecular evolutionary research in a non-standard model group of organisms requires quite some effort for collecting and cultivation of the desired plants. I am indebted to all people who have contributed to obtain tissue of Orobanchaceae species that have been the focus of my PhD-research or relate to it. Great thanks are due to Dr. Wolfram Lobin, Klaus Bahr and Jörg Dombrovski (all Botanical Garden Bonn, Germany) for cultivation of "my" broomrapes. Dr. Barbara Ditsch (Botanical Garden Dresden, Germany), Dr. Alison E. Colwell (U.S. Geological Survey, USA), Dr. Kay Kirkman (Joseph W. Jones Ecological Research Center, USA), Dr. Mats Hjertson (Uppsala University, Sweden), Dr. Olaf Werner, Prof Dr. Rosa María Ros (both University of Murcia, Spain), and Dr. Jesús Muñoz (Botanical Garden Madrid, Spain) have provided DNA or fresh plant material which is most gratefully acknowledged.

I would also like to thank all people of the working groups in Vienna, Bonn, Münster and State College/PA for their generous help and support and a great PhD-time.

Financial support of this project by the Austrian Science Fund (FWF, grant 19404 to G. M. Schneeweiss) and travel support from the University of Vienna (KWA program) and the Genetics Section of the Botanical Society of America are gratefully acknowledged.

Last but not least, I would like express my deepest thanks to my parents! Thanks so much for supporting me time and again with the most incredible patience, benevolence and faith. This work is for you!

ABSTRACT

The prime function of the plastid organelle is to carry out photosynthesis thereby providing autotrophy to the plant kingdom. Plastids retain a semi-autonomous genetic system including a genome (plastome) encoding subunits for photosynthesis-related and unrelated processes as well as proteins for basic functions of the genetic apparatus. Due to the transition from an autotrophic to a semi- or holo-heterotrophic lifestyle, parasitic plants show major plastomic reconfigurations with extreme reductions of plastome size and coding capacity as well as extraordinarily elevated nucleotide substitution rates. Using the broomrape family (Orobanchaceae) as a model group, this dissertation thesis reconstructs molecular evolutionary patterns of reductive plastome evolution of the plastid chromosome under relaxed evolutionary constraints. Employing comparative-evolutionary analyses of completely sequenced plastid genomes from several hemi- and holoparasitic members of Orobanchaceae this work examines aspects concerning the (i) co-linearity and structural rearrangements of plastomes, (ii) potential functionality of genes involved in photosynthesis, (iii) pseudogenization and gene loss, and (iv) accelerated substitution rates in plastid genomes. In addition, one chapter evaluates methodological aspects of plastid genome sequencing employing whole-genome shotgun pyrosequencing. This work reveals that genetic and genomic changes concerning plastome structure, nucleotide substitution rates and selectional constraints occur in a complex and highly lineage specific manner, and it provides novel insights into factors influencing reductive evolution of plastome. Increasing host-dependency notably seeds excessive non-functionalization of plastid genes due to pseudogenization or deletion, and severely relaxes the structural maintenance of the plastid chromosome. Pseudogenization and segmental deletions of newly dispensable regions depend significantly on the operon-structure of the plastid chromosomes as well as on the distribution of essential genes in Orobanchaceae. There is evidence for maintained or alternative function of a photosynthesis-related complex as well as for putatively increased rates of intracellular gene transfer in parasitic plants. Analyses of nucleotide substitutions reveal significantly elevated rates in both housekeeping and photosynthesis genes already in photosynthetic heterotrophs indicating that relaxation of selective constraints relates to the transition to a parasitic lifestyle. Compared to hemiparasites and autotrophs, distinctive trends of rate and selectional changes exist among holoparasite lineages including both local accelerations and rate reductions. Above that, this thesis shows for the first time that the successful reconstruction of plastid chromosomes from whole-genome shotgun pyrosequencing strongly depends on the size of the assembled read pool. Using the results of simulated and empirical 454 datasets in combination with a resampling scheme for automated quality assessment, a method for a parameter-less *a priori* assessment of the optimal read pool size is established that should ease assembly efforts.

ZUSAMMENFASSUNG

Der Plastid, der als Schlüsselfunktion der autotrophen Lebensweise die Photosynthese ausführt, besitzt ein semi-autonomes genetisches System mit eigenem Genom (Plastom), welches für Proteine des Photosyntheseapparates sowie wenige Enzyme anderer metabolischer Prozesse und Untereinheiten grundlegender genetischer Prozesse kodiert. Aufgrund des Überganges zu einer unterschiedlich stark ausgeprägten heterotrophen Lebensweise weisen parasitische Pflanze enorm modifizierte Plastome auf. Diese sind durch eine extreme funktionelle und strukturelle Reduktion sowie stark erhöhte DNS-Substitutionsraten charakterisiert. Gegenstand dieser Dissertation ist es, evolutive Trends der Plastomreduktion bei verminderten Selektionsdrücken zu rekonstruieren. Zu diesem Zweck wurden die Plastome verschiedener zur Photosynthese fähiger und unfähiger Vertreter der Sommerwurzgewächse (Orobanchaceae) vollständig sequenziert und mittels moderner Methoden der vergleichenden Genomanalyse hinsichtlich folgender Aspekte analysiert: (i) strukturelle Änderungen, (ii) potentielle Funktionalität von Photosynthese-assoziierten Plastidengenen, (iii) Pseudogenisierung und Gendeletion, und (iv) Evolution und Auswirkung erhöhter plastidärer DNS-Substitutionsraten. Darüber hinaus behandelt ein Kapitel methodologische Aspekte der Sequenzierung von Plastidengenomen mittels Pyrosequenzierung gesamt-genomischer DNS-Extrakte. Die Analyse plastidärer DNS nah-verwandter Orobanchaceae erlaubt es erstmals, komplexe Muster der Genomreduktion in parasitischen Pflanzen aufzudecken. Unter anderem kann gezeigt werden, dass bereits der Übergang zu einer heterotrophen Lebensweise für strukturelle Änderungen des Plastoms ausschlaggebend ist und zu einem Verlust der Funktionalität bestimmter Gene und einer Erhöhung der Substitutionsraten führt. Innerhalb der Orobanchaceae schreitet die Plastomreduktion mit zunehmender Heterotrophie mit linienspezifischer Geschwindigkeit voran. Das Ausmaß von Pseudogenisierung und Deletion nicht-essentieller Genomabschnitte wird dabei maßgeblich durch die Distanz zu essentiellen genischen Elementen und von der plastidären Operonstruktur beeinflusst. Darüber hinaus weisen die zusammengetragenen Ergebnisse darauf hin, dass parasitische Pflanzen die Funktion einzelner Photosynthese-assoziierter Proteinkomplexe möglicherweise aufrechterhalten und eine erhöhte Rate an intrazellulärem DNS-Transfer aufweisen. Veränderungen der DNS-Substitutionsmuster bei zur Photosynthese fähigen heterotrophen Orobanchaceae implizieren eine Korrelation des Übertritts zur parasitischen Lebensweise mit der Verminderung der Selektion bestimmter plastidärer Genen. Im Vergleich zu photosynthetisch aktiven Pflanzen, weisen Vollparasiten differenzierte Muster bezüglich DNS-Substitutionen auf, einschließlich linienspezifischer Ratenerhöhung und –reduktion. In der vorliegenden Arbeit wird anhand der Analyse von simulierten und experimentell generierten 454-Datensätzen erstmals gezeigt, dass die erfolgreiche Plastomrekonstruktion signifikant von der verwendeten Sequenzdatenmenge bestimmt wird. Darüber hinaus wird eine Methode zur *a priori* Schätzung der optimalen Datenmenge unter Verwendung weniger Parameter erarbeitet.

– CHAPTER I –

GENERAL INTRODUCTION

CONTENTS.

This chapter contains approx. 1,900 and 3 figures.

1. A distinguishing organelle: The primary plastid of the plant kingdom and its remnant genetic system in angiosperms.

The autotrophic lifestyle is the distinguishing feature of the plant kingdom. The primary uptake of a cyanobacterium-like prokaryote by a eukaryotic cell has been an "event of global significance" [1], providing the possibility of an autotrophic lifestyle by converting light into biochemical energy. Over time, host and endosymbiont deeply intertwined their metabolic apparatuses and cell cycles and evolved complex systems of transport and genetic signaling [2–5]. As part of this process, genetic information was transferred from the endosymbiont to the nuclear genome of the eukaryotic cell [6]. Eventually, the endosymbiont lost its genetic autonomy and became subjected to nuclear regulation as the plastid organelle. Since its enclosure into the eukaryotic cell, more than 90% of the genetic information of the endosymbiont were functionally transferred into the nuclear genome or they were lost [7].

Various forms of (primary) plastids are distinguished based upon their developmental stage, localization or physiological significance and function [reviewed in 8]. While the undifferentiated form, known as the *proplastid* can only be found in meristematic and meiotic tissue, c*hloroplasts* represent the fully differentiated and pivotal form that carry out photosynthesis and carbon fixation. Beyond this, important steps during the synthesis of amino acids, lipids and pigments as well as crucial (intermediary) steps during the assimilation of mineral nutrients such as nitrogen, phosphorous and sulfur [9] take place in plastids in general.

Today, many aspects of plastid-cell biochemical and genetic pathways are still unknown. However, one thing that may be referred to as one of the best-studied facets of the plastid organelle is its genome (plastome). Due to the constantly high selective pressure on its genes that majorly encode subunits of photosynthesis pathways and elements for their transcription and translation, the plastid chromosome exhibits an extraordinary extent of functional and structural conservation among land plants. The second chapter of this thesis will review our current knowledge regarding aspects of the evolution of plastomes in land plants in detail. Special attention will be paid to the genomic rearrangement history of the plastome. Another focus will be the plastomic gene content, and the function of plastid-encoded genes.

Across lineages of *Plantae*, we observe a great number of most diverse plastid chromosome structures with complex and gene-rich structures in algal lineage [10–17] to highly compact plastomes in angiosperms [18–21]. The compact arrangement of land plant plastid chromosomes is the result of an ongoing reductive evolutionary process during which

genes are functionally transferred. Within angiosperm, several genes are independently translocated from the plastid chromosome to the nucleus [22–26]. The process of functional gene transfer appears to become significantly more prominent in lineages that display non-canonical structural evolution of the plastid chromosome [24,26–28]. Among angiosperm lineages with the most severe plastomic reconfigurations, parasitic plants are certainly the most exceptional group.

2. Parasitic flowering plants – An evolutionary experiment that provides an excellent natural model system for studying processes of molecular evolution.

Parasitism is a proven and successful life strategy that has evolved in all organismal kingdoms. The transition from autotrophy to heterotrophy within land plants occurs via two different mechanisms: On the one hand, *myco-heterotrophic* plants establish a unidirectional connection to the roots of an autotrophic plant via a mycorrhizal fungus. On the other hand, heterotrophic plants can connect directly to an autotrophic plant. In order to distinguish between these two fundamentally different forms, the term *parasitic plant* will be used throughout the entire thesis to describe the latter type of non-mutualistic plant interaction. As of this writing, parasitic plants are only known to originate within flowering plants, whereas myco-heterotrophy has evolved independently in nearly all land plant lineages (e.g. in liverworts, ferns, seed plants). Parasitic plants penetrate and connect to their host plants with the help of a highly specialized organ, the haustorium. In root parasitic plants, the haustorium forms either from the terminal root meristem, or it originates from lateral roots [29–31]. In Orobanchaceae, parasite-host connections may also develop from non-root tissue such as the hypocotyl [32,33].

Parasitism evolved at least 12 times independently in different angiosperm lineages [Fig. I-1; 31,34,35]. Heterotrophic plants display a wide range of different degrees of heterotrophy and developed a number of different and highly derived growth forms [Fig. I-1; 36,37]. Assignment of heterotrophic plants to their closest autotrophic relatives is often challenging due to their commonly highly derived morphology. In general, heterotrophic plants can be subdivided into photosynthetic heterotrophs that retain the ability of autotrophic nutrition to a greater or lesser extent (hemiparasites, hemi- mycoheterotrophs), and non-photosynthetic heterotrophs (holoparasites) that are completely dependent on a host plant. While

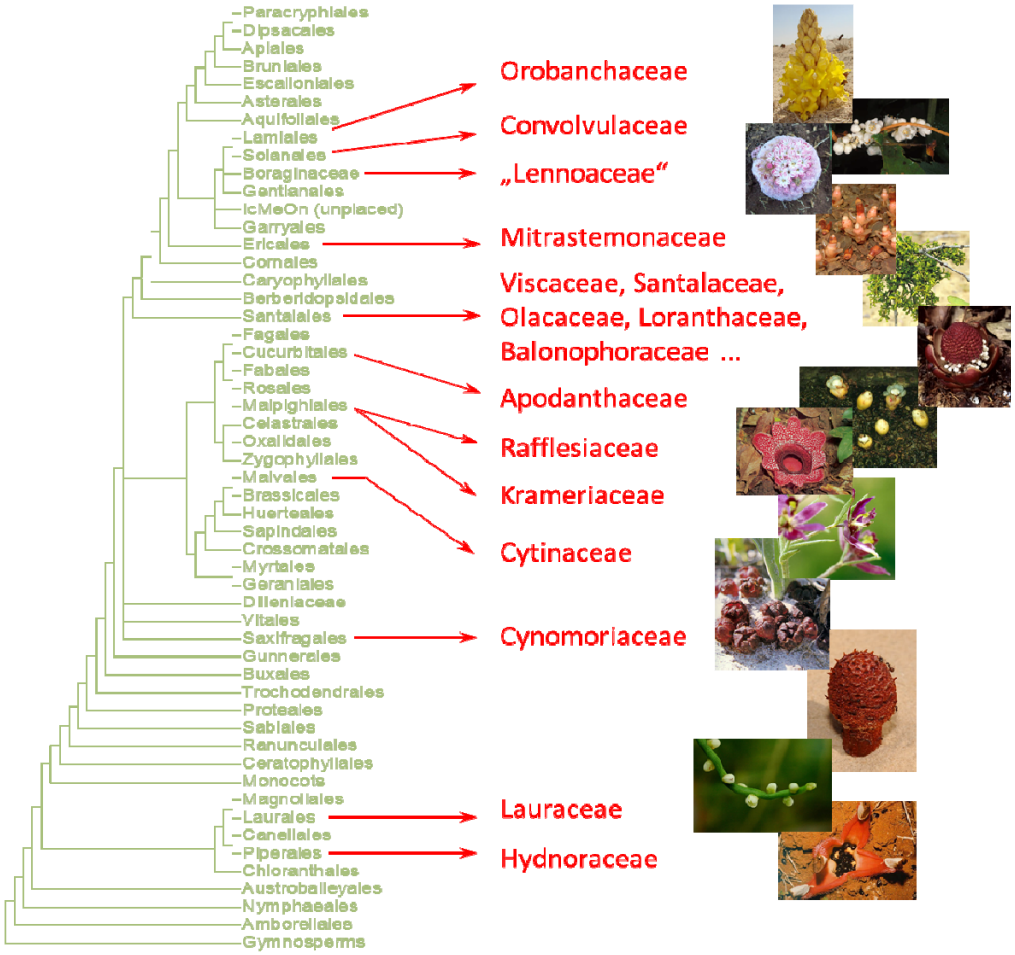


Fig. I-1 **Evolution of parasitic plants in flowering plants.** Based upon sequence data, families of parasitic plants have been assigned to their closest autotrophic relatives. The transition to parasitism occurred at least 12 time independently; Holo-heterotrophic stages evolved in at least 10 families.
[Tree topology after (32). All photographs are available online from the *Parasitic Plant Connection* at http://www.parasiticplants.siu.edu/.]

holoparasites and obligate hemiparasites completely rely on their host to reach their reproductive stage, facultative hemiparasites are able to fulfill their lifecycle without attaching to a host plant [31,36]. Within basal angiosperms, transition from an autotrophic lifestyle to hemi- and holo-heterotrophy evolved twice. All other parasitic angiosperm lineages belong to the large clade of eudicots. Among parasitic angiosperms, the complete ranges of parasitic forms occur only within a single family: The broomrape family (Orobanchaceae, Lamiales; Fig I-2).

Orobanchaceae are nearly cosmopolitan and comprise about 2,000 species in nearly 90 genera [38]. Known for their devastating effect in monocultures of crops, several members

Fig. I-2 **Representatives of Orobanchaceae.** a. *Lindenbergia philippinensis*; b. *Schwalbea americana*; c. *Striga asiatica*; d. *Conopholis americana*; e. *Cistanche phelypaea*; f. *Boulardia latisquama*; g. *Orobanche gracilis*; h. *O. cernua var. cumana*; i. *Myzorrhiza californica*; j. *Phelipanche ramosa*. k. close-up of the flower of *O. crenata*; l. Roots of *Vigna* infected with *O. crenata* showing the vegetative phase of the parasitic plant as small bulbs. m. Haustorium of *O. crenata* at the reproductive phase.
[Photo sources: a-e, j: obtained from the *Parasitic Plant Connection* at http://www.parasiticplants.siu.edu/; g: K.F. Müller; h: G.M. Schneeweiss; f,i,k-m: own.]

are of ecological and economic importance (e.g. *Orobanche crenata, Phelipanche ramosa, Striga*species). Spanning the complete set of evolutionary transition forms makes this family an excellent natural model system for studying molecular evolutionary processes in relation

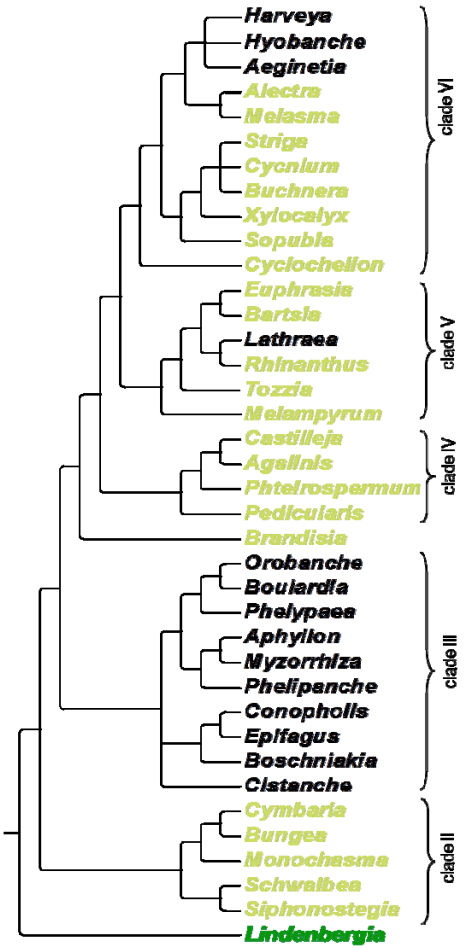


Fig. I-3 **Relationships among Orobanchaceae.** Based upon a nuclear gene marker, the inferred relationships of Orobanchaceae revealed that holoparasitism emerged several times (black). With the exception of autotrophic *Lindenbergia* (green), the remaining species of Orobanchaceae are photosynthetic heterotrophs (light green). [Topology after (36).]

to life form shifts and different levels of relaxed evolutionary constraints. Besides non-parasitic autotrophs, the family includes a great diversity of parasitic autotrophs (facultative and obligate hemiparasites with different degrees of photosynthetic capabilities) as well as non-photosynthetic heterotrophs (holoparasites) that evolved at least three times independently [31,38]. Above that, the phylogenetic relationships within Orobanchaceae are worked out well [Fig. I-3; 38–41], and provide a solid basis for testing hypotheses concerning the molecular evolution of genes, operons, and chromosome structure as well as non-coding genome regions in the light of different selective constraints. Above that, Orobanchaceae have been in the focus of numerous studies that provide invaluable insights in the dynamics of genome and transcriptome evolution of parasitic plants [42–46]. For instance, the close interaction between the parasites and their host has made them excellent case studies to investigate horizontal gene transfer (HGT). For instance, HGT appears to occur in lineages of *Orobanche* and *Phelipanche* [47]; and has been demonstrated for hemiparasitic Orobanchaceae

[48,49]. Similar studies in other parasitic plant lineages corroborate that rampant horizontal gene transfers seems to occur frequently between parasites and their hosts [49–52].

Accelerated reductive evolution due to relaxed selective pressure upon the photosynthetic apparatus led parasitic flowering plants to retain only a subset of the plastid genes present in autotrophic angiosperms. It will be the major focus of this thesis to investigate the evolution of the plastid chromosome in a set of closely related species of parasitic plants of the broomrape family.

3. Motivation and aims of this work

The general pattern of plastid genome reduction applies to all non-photosynthetic species investigated so far, although they exhibit considerable differences [53–60]. In 1990, dePamphilis and Palmer revealed that all genes for the photosynthetic apparatus have been lost in the holoparasite *Epifagus*. Later on, *Conopholis* was shown to share the majority of gene losses with its sister lineage *Epifagus* [55,61]. Studies using a wider taxon sampling but focusing only on a smaller sampling of plastomic gene regions suggested that reductive evolution within Orobanchaceae (and supposedly other parasite plant lineages as well) appears to be a highly lineage-specific process. For instance, some non-photosynthetic lineages of the broomrape family (Orobanchaceae) preserve and even express the *rbcL* gene coding for the large subunit of RuBisCO [62–64]. In contrast, several other species harbor only a diverged and pseudogenized copy [65,66]. Similarly, the set of plastid-encoded tRNAs differs substantially within and among various lineages of parasitic plants [54,56,59,60,67–70]. Apart from the distortion of gene order due to gene deletions, there are prominent differences in the structural evolution of the plastid chromosomes among holo-heterotrophs. While the *Epifagus* plastome structure does not show modifications in relative gene synteny compared to autotrophs [54], restriction mapping analyses of two Orobanchaceae species revealed that the obligate hemiparasites *Striga* and *Conopholis* have lost one important plastomic segment independently [71]. Similarly, a contraction of the large inverted repeat regions (IR) occurs in the Australian Earth orchid *Rhizanthella* [59], whereas a very conservative structure with widely co-linear gene synteny relative to closely related autotrophs was reported for the Bird's nest orchid, *Neottia* [60]. IR constriction also occurs in *Cuscuta* species, and so do several inversions in the single copy regions [69,70]. Besides structural changes, parasites evolve at high mutational rates in most genes of all three genomes [39,60,72–74]. Particularly, enhanced

nucleotide substitution rate are observed in the majority of plastid genes, although many of the retained genes still evolve under purifying selection [39,66,72]. In 1997, dePamphilis and colleagues (p. 7367) state that

> *"… we know nothing about the dynamics of the events that led to the extraordinary plastid DNAs seen in parasitic plants. For example, is there some predictable order to the genetic changes associated with the loss of photosynthesis? Have the profound alterations to the plastid genome seen in Epifagus and Conopholis all followed the loss of photosynthesis, or might some have begun before its loss? Are the similarities between the plastid genomes of these plants due to convergence or to shared ancestry? How rapidly do the structural and other changes to the plastid genome come about in parasitic plants? […]".*

In the following chapters, this dissertation will attempt to provide answers to some of those questions, and assess molecular evolutionary patterns of reductive evolution in the plastid chromosome of a group of parasitic plants. Employing comparative-evolutionary analyses of completely sequenced plastid genomes from several hemi- and holoparasitic members of the broomrape family, this work will analyze the following aspects: (i) co-linearity and structural rearrangements, (ii) potential functionality of genes involved in photosynthesis, (iii) pseudogenization and gene loss, and (iv) accelerated substitution rates in plastid genomes. In addition, one chapter of this thesis will evaluate methodological aspects of plastid genome sequencing using whole-genome shotgun sequencing with high-throughput sequencing techniques.

4. References

1. Keeling PJ (2010) The endosymbiotic origin, diversification and fate of plastids. Phil Trans R Soc B 365: 729–748.

2. Bölter B, Soll J, Schulz A, Hinnah S, Wagner R (1998) Origin of a chloroplast protein importer. Proc Natl Acad Sci USA 95: 15831–15836.

3. Leon P, Arroyo A, Mackenzie S (1998) Nuclear control of plastid and mitochondrial development in higher plants. Ann Rev Plant Physiol Plant Mol Biol 49: 453–480.

4. Vesteg M, Vacula R, Krajčovič J (2009) On the origin of chloroplasts, import mechanisms of chloroplast-targeted proteins, and loss of photosynthetic ability — review. Folia Microbiol 54: 303–321.

5. Samuel I. B (2011) Chloroplast signaling: Retrograde regulation revelations. Curr Biol 21: R391–R393.

6. Bock R, Timmis JN (2008) Reconstructing evolution: Gene transfer from plastids to the nucleus. BioEssays 30: 556–566.

7. Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, et al. (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA 99: 12246–12251.

8. Pyke KA (2007) Plastid biogenesis and differentiation. In: Cell and Molecular Biology of Plastids. Topics in Current Genetics. Berlin/Heidelberg: Springer, Vol. 19. pp. 1–28.

9. Neuhaus HE, Emes MJ (2010) Nonphotosynthetic metabolism in plastids. Annu Rev Plant Physiol Plant Mol Biol 51: 111–140.

10. Reith M, Munholland J (1995) Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome. Plant Mol Biol Rep 13: 333.

11. Stirewalt V, Michalowski C, Löffelhardt W, Bohnert H, Bryant D (1995) Nucleotide sequence of the cyanelle genome from *Cyanophora paradoxa*. Plant Mol Biol Rep 13: 327–332.

12. Turmel M, Otis C, Lemieux C (1999) The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: Insights into the architecture of ancestral chloroplast

genomes. Proc Natl Acad Sci USA 96: 10248–10253.

13. Lemieux C, Otis C, Turmel M (2000) Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. Nature 403: 649–652.

14. Maul JE, Lilly JW, Cui L, dePamphilis CW, Miller W, et al. (2002) The *Chlamydomonas reinhardtii* plastid chromosome: Islands of genes in a sea of repeats. Plant Cell 14: 2659–2679.

15. Turmel M, Otis C, Lemieux C (2002) The chloroplast and mitochondrial genome sequences of the charophyte C*haetosphaeridium globosum*: Insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. Proc Natl Acad Sci USA 99: 11275–11280.

16. Turmel M, Otis C, Lemieux C (2005) The complete chloroplast DNA sequences of the charophycean green algae *Staurastrum* and *Zygnema* reveal that the chloroplast genome underwent extensive changes during the evolution of the Zygnematales. BMC Biology 3: 22.

17. Pombert JF, Otis C, Lemieux C, Turmel M (2005) The chloroplast genome sequence of the green alga *Pseudendoclonium akinetum* (Ulvophyceae) reveals unusual structural features and new insights into the branching order of chlorophyte lineages. Mol Biol Evol 22: 1903–1918.

18. Palmer JD (1985) Comparative organization of chloroplast genomes. Ann Rev Genet 19: 325–354.

19. Sugiura M (1992) The chloroplast genome. Plant Mol Biol 19: 149–168.

20. Raubeson LA, Jansen RK (2005) Chloroplast genomes of plants. In: Diversity and evolution of plant-genotypic and phenotypic variation in higher plants: CABI Publishing London, pp. 45–68.

21. Bock R (2007) Structure, function, and inheritance of plastid genomes. In: Cell and Molecular Biology of Plastids. Topics in Current Genetics. Springer Berlin/Heidelberg, Vol. 19. pp. 29–63.

22. Baldauf SL, Manhart JR, Palmer JD (1990) Different fates of the chloroplast *tufA* gene following its transfer to the nucleus in green algae. Proc Natl Acad Sci USA 87: 5317–5321.

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

24. Chris Blazier J, Guisinger-Bellian MM, Jansen RK (2011) Recent loss of plastid-encoded *ndh* genes within *Erodium* (Geraniaceae). Plant Mol Biol 76: 1–10.

25. Jansen RK, Saski C, Lee S-B, Hansen AK, Daniell H (2011) Complete plastid genome sequences of three rosids (*Castanea, Prunus, Theobroma*): evidence for at least two independent transfers of *rpl22* to the nucleus. Mol Biol Evol 28: 835–847.

26. Magee AM, Aspinall S, Rice DW, Cusack BP, Sémon M, et al. (2010) Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Res 20: 1700–1710.

27. Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, et al. (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA 104: 19369–19374.

28. Guisinger MM, Kuehl JV, Boore JL, Jansen RK (2011) Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Mol Biol Evol 28: 583–600.

29. Kujit J (1969) The Biology of Parasitic Flowering Plants. 1st ed. Berkeley, CA.: University of California Press.

30. Kujit J, Toth R (1976) Ultrastructure of angiosperm haustoria - A review. Ann Bot 40: 1121–1130.

31. Westwood JH, Yoder JI, Timko MP, dePamphilis CW (2010) The evolution of parasitism in plants. Trends Plant Sci 15: 227–235.

32. Weber HC (1993) Parasitismus von Blütenpflanzen. Darmstadt: Wissenschaftl. Buchgesellschaft.

33. Joel DM (2007) Direct infection of potato tubers by the root parasite *Orobanche aegyptiaca*. Weed Res 47: 276–279.

34. The Angiosperm Phylogeny Group (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot J Linn Soc 161: 105–121.

35. Filipowicz N, Renner S (2010) The worldwide holoparasitic Apodanthaceae confidently placed in the Cucurbitales by nuclear and mitochondrial gene trees. BMC Evolutionary Biology 10: 219.

36. Press MC, Graves J (1995) Parasitic Plants. 1st ed. Springer Netherlands.

37. Heide-Jørgensen HS (2008) Parasitic Flowering Plants. Brill Academic Publishers.

38. Bennett JR, Mathews S (2006) Phylogeny of the parasitic plant family Orobanchaceae inferred from phytochrome A. Am J Bot 93: 1039–1051.

39. dePamphilis CW, Young ND, Wolfe AD (1997) Evolution of plastid gene *rps2* in a lineage of hemiparasitic and holoparasitic plants: Many losses of photosynthesis and complex patterns of rate variation. Proc Natl Acad Sci USA 94: 7367–7372.

40. Schneeweiss GM, Colwell A, Park J-M, Jang C-G, Stuessy TF (2004) Phylogeny of holoparasitic *Orobanche* (Orobanchaceae) inferred from nuclear ITS sequences. Mol Phylogenet Evol 30: 465–478.

41. Park J-M, Manen J-F, Colwell A, Schneeweiss GM (2008) A plastid gene phylogeny of the non-photosynthetic parasitic *Orobanche* (Orobanchaceae) and related genera. J Plant Res 121: 365–376.

42. Weiss-Schneeweiss H, Greilhuber J, Schneeweiss GM (2005) Genome size evolution in holoparasitic *Orobanche* (Orobanchaceae) and related genera. Am J Bot 93: 148–156. doi:10.3732/ajb.93.1.148.

43. Park J-M, Schneeweiss GM, Weiss-Schneeweiss H (2007) Diversity and evolution of *Ty1-copia* and *Ty3-gypsy* retroelements in the non-photosynthetic flowering plants *Orobanche* and *Phelipanche* (Orobanchaceae). Gene 387: 75–86.

44. Westwood JH, Roney JK, Khatibi PA, Stromberg VK (2009) RNA translocation between parasitic plants and their hosts. Pest Management Science 65: 533–539.

45. Wickett NJ, Honaas LA, Wafula EK, Das M, Huang K, et al. (2011) Transcriptomes of the parasitic plant family Orobanchaceae reveal surprising conservation of chlorophyll synthesis. Curr Biol 21: 2098–2104.

46. Yoshida S, Ishida J, Kamal N, Ali A, Namba S, et al. (2010) A full-length enriched cDNA library and expressed sequence tag analysis of the parasitic weed, *Striga hermonthica*. BMC Plant Biol 10: 55.

47. Park J-M, Manen J-F, Schneeweiss GM (2007) Horizontal gene transfer of a plastid gene in the non-photosynthetic flowering plants *Orobanche* and *Phelipanche* (Orobanchaceae). Mol Phyl Evol 43: 974–985.

48. Yoshida S, Maruyama S, Nozaki H, Shirasu K (2010) Horizontal gene transfer by the parasitic plant *Striga hermonthica*. Science 328: 1128.

49. Mower JP, Stefanovic S, Young GJ, Palmer JD (2004) Gene transfer from parasitic to host plants. Nature 432: 165–166.

50. Davis C, Wurdack KJ (2004) Host-to-parasite gene transfer in flowering plants: phylogenetic evidence from Malpighiales. Science 305: 676–678.

51. Barkman TJ, McNeal JR, Lim SH, Coat G, Croom HB, et al. (2007) Mitochondrial DNA suggests at least 11 origins of parasitism in angiosperms and reveals genomic chimerism in parasitic plants. BMC Evol Biol 7: 248.

52. Mower J, Stefanovic S, Hao W, Gummow J, Jain K, et al. (2010) Horizontal acquisition of multiple mitochondrial genes from a parasitic plant followed by gene conversion with host mitochondrial genes. BMC Biology 8: 150.

53. dePamphilis CW, Palmer JD (1990) Loss of photosynthetic and chlororespiratory genes from the plastid genome of a parasitic flowering plant. Nature 348: 337–339.

54. Wolfe KH, Morden CW, Palmer JD (1992) Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc Natl Acad Sci USA 89: 10648–10652.

55. Wimpee C, Wrobel R, Garvin D (1991) A divergent plastid genome in Conopholis americana, an achlorophyllous parasitic plant. Plant Mol Biol 17: 166, 161.

56. Wickett NJ, Fan Y, Lewis P, Goffinet B (2008) Distribution and evolution of pseudogenes, gene losses, and a gene rearrangement in the plastid genome of the nonphotosynthetic liverwort, *Aneura mirabilis* (Metzgeriales, Jungermanniopsida). J Mol Evol 67: 111–122.

57. Stefanović S, Olmstead R (2005) Down the slippery slope: Plastid genome evolution in Convolvulaceae. J Mol Evol 61: 292–305.

58. Nickrent D, García M (2009) On the brink of holoparasitism: Plastome evolution in Dwarf Mistletoes (*Arceuthobium*, Viscaceae). J Mol Evol 68: 603–615.

59. Delannoy E, Fujii S, des Francs CC, Brundrett M, Small I (2011) Rampant gene loss in the underground orchid *Rhizanthella gardneri* highlights evolutionary constraints on plastid genomes. Mol Biol Evol 28: 2077–2086.

60. Logacheva MD, Schelkunov MI, Penin AA (2011) Sequencing and analysis of plastid genome in mycoheterotrophic orchid *Neottia nidus-avis*. Genome Biol Evol 3: 1296–1303.

61. Wimpee CF, Morgan R, Wrobel RL (1992) Loss of transfer RNA genes from the plastid 16S–23S ribosomal RNA gene spacer in a parasitic plant. Curr Genet 21: 417–422.

62. Delavault PM, Sakanyan V, Thalouarn P (1995) Divergent evolution of two plastid genes, *rbcL* and *atpB*, in a non-photosynthetic parasitic plant. Plant Mol Biol 29: 1071–1079.

63. Lusson NA, Delavault PM, Thalouarn P (1998) The *rbcL* gene from the non-photosynthetic parasite *Lathraea clandestina* is not transcribed by a plastid-encoded RNA polymerase. Curr Genet 34: 212–215.

64. Randle CP, Wolfe AD (2005) The evolution and expression of RBCL in holoparasitic sister-genera *Harveya* and *Hyobanche* (Orobanchaceae). Am J Bot 92: 1575–1585.

65. Wolfe AD, dePamphilis CW (1997) Alternate paths of evolution for the photosynthetic gene *rbcL* in four nonphotosynthetic species of *Orobanche*. Plant Mol Biol 33: 965–977.

66. Young ND, dePamphilis CW (2005) Rate variation in parasitic plants: correlated and uncorrelated patterns among plastid genes of different function. BMC Evol Biol 5: 16.

67. Taylor GW, Wolfe KH, Morden CW, dePamphilis CW, Palmer JD (1991) Lack of a functional plastid $tRNA^{Cys}$ gene is associated with loss of photosynthesis in a lineage of parasitic plants. Curr Genet 20: 515–518.

68. Lohan AJ, Wolfe KH (1998) A subset of conserved tRNA genes in plastid DNA of nongreen plants. Genetics 150: 425–433.

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

70. McNeal JR, Kuehl J, Boore J, de Pamphilis C (2007) Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the

parasitic plant genus *Cuscuta*. BMC Plant Biol 7: 57.

71. Downie SR, Palmer JD (1992) Restriction site mapping of the chloroplast DNA inverted repeat - a molecular phylogeny of the Asteridae. Ann Mo Bot Gard 79: 266–283.

72. Wolfe KH, Morden CW, Ems SC, Palmer JD (1992) Rapid evolution of the plastid translational apparatus in a nonphotosynthetic plant: loss or accelerated sequence evolution of tRNA and ribosomal protein genes. J Mol Evol 35: 304–317.

73. Nickrent DL, Duff RJ, Colwell AE, Wolfe AD, Young ND, et al. (1998) Molecular phylogenetic and evolutionary studies of parasitic plants. Molecular Systematics of Plants II DNA Sequencing: 211–241.

74. Lemaire B, Huysmans S, Smets E, Merckx V (2011) Rate accelerations in nuclear 18S rDNA of mycoheterotrophic and parasitic angiosperms. J Plant Res 124: 561–576.

– CHAPTER II –

THE EVOLUTION OF THE PLASTID CHROMOSOME IN LAND PLANTS: GENE CONTENT, GENE ORDER, GENE FUNCTION

ABSTRACT. This review bridges functional and evolutionary aspects of plastid chromosome architecture in land plants and their putative ancestors. We provide an overview on the structure and composition of the plastid genome of land plants as well as the functions of its genes in an explicit phylogenetic and evolutionary context. We will discuss the architecture of land plant plastid chromosomes, including gene content and synteny across land plants. Moreover, we will explore the functions and roles of plastid encoded genes in metabolism and their evolutionary importance regarding gene retention and conservation. We suggest that the slow mode at which the plastome typically evolves is likely to be influenced by a combination of different molecular mechanisms. These include the organization of plastid genes in operons, the usually uniparental mode of plastid inheritance, the activity of highly effective repair mechanisms as well as the rarity of plastid fusion. Nevertheless, structurally rearranged plastomes can be found in several unrelated lineages (e.g. ferns, Pinaceae, multiple angiosperm families). Rearrangements and gene losses seem to correlate with an unusual mode of plastid transmission, abundance of repeats, or a heterotrophic lifestyle (parasites or myco-heterotrophs). While only a few functional gene gains and more frequent gene losses have been inferred for land plants, the plastid *Ndh* complex is one example of multiple independent gene losses and will be discussed in detail. Patterns of *ndh*-gene loss and functional analyses indicate that these losses are usually found in plant groups with a certain degree of heterotrophy, might rendering plastid encoded *Ndh* subunits dispensable.

KEYWORDS. Plastid genome, land plants, genome evolution, plastid gene function, gene retention

CONTENTS.

This chapter contains approx. 19,000 words, 2 figures and 1 table.

1. INTRODUCTION

Plastids are one of the main distinguishing characteristics of the plant cell. The central function of the plastid is to carry out photosynthesis, but other major cellular functions also take place in plastids, including synthesis of starch, fatty acids, pigments, and amino acids (reviewed by Neuhaus and Emes 2010). As early as 1905, Konstantin S. Mereschkowski hypothesized that plant "chromatophores" are the result of the uptake of a cyanobacterium by a eukaryotic organism (English translation available by Martin and Kowallik 1999). It is now generally accepted that the plastid originated via incorporation of a free-living cyanobacterial-like prokaryote into a eukaryotic cell (primary endosymbiosis), thereby enabling the transition from heterotrophy to autotrophy by gaining the ability of utilizing photoenergy. Recent phylogenetic analyses of plastid genes from major plant lineages have converged on the hypothesis that plastids of the plant kingdom, i.e. the clade including Glaucophytes, Rhodophytes, Chlorophytes, and Streptophytes (Fig. II-1; Keeling 2004), are derived from a single origin (Palmer 2000; McFadden and van Dooren 2004; Keeling 2010). This is also supported by several biochemical features, such as the composition of light harvesting complexes and their components, structural RNAs, membrane structure, and the protein import/targeting machinery (Weeden 1981; Bölter et al. 1998; Keeling 2004; Yang and Cheng 2004; Koziol et al. 2007; Vesteg et al. 2009).

Over evolutionary time, genetic information was functionally or more often non-functionally transferred from the endosymbiont's genetic system to the host nuclear genome, genetically intertwining the two genomes. Except for genes involved in photometabolic processes, most other genes have been incorporated into the nuclear genome. This has resulted in a highly reduced plastid genome in Streptophytes (land plants plus their closest algal relatives), comprising less than 5–10% of the genes hypothesized for the ancestral cyanobacterial genome (ca. 2000 to 3000 genes; Martin et al. 2002). A corollary of this process is that the plastid genome (plastome) became subjected to nuclear regulation (Timmis et al. 2004), locking in their symbiotic relationship. The transfer of sequences and both functional and non-functional genes from the plastid genome to both the nuclear and the mitochondrial genome remains an ongoing process (Stern and Lonsdale 1982; Stern and Astwood 1986; Nakazono and Hira 1993; Ayliffe et al. 1998; Shahmuradov et al. 2003; Matsuo et al. 2005; Guo et al. 2008; Sheppard and Timmis 2009). This intracellular gene transfer is considered "frequent and [to occur] in big chunks" (Martin 2003:1; Stegemann et al. 2003; Noutsos et al. 2005). The question of how many genes can eventually be transferred to the nuclear genome (and whether the plastome could eventually be lost) has been discussed for some time (Barbrook et al. 2006). Massive gene loss has been observed in several parasitic plants (e.g. Orobanchaceae: Wolfe et al.

1992; *Cuscuta*: Funk et al. 2007, McNeal et al. 2007). In these plants, gene loss is not restricted to genes that are primarily involved in photosynthesis and related pathways (Wolfe et al. 1992; Krause 2008); additional losses or pseudogenization is seen in genes encoding subunits of the genetic apparatus (e.g., plastid-encoded RNA polymerase, some tRNAs, some ribosomal proteins; dePamphilis and Palmer 1990; Wolfe et al. 1992; Lohan and Wolfe 1998).

Four decades of genetic, genomic and physiological research have contributed substantially to assign genes and gene functions to land plant plastid encoded proteins. Plastid genes have been grouped into functionally defined classes, including (i) those involved in primary and secondary photosynthesis pathways (photosynthetic light and dark reactions), (ii) genes not involved in photosynthetic pathways, such as sulfate transport and lipid acid synthesis, (iii) genes involved in transcription and translation, and (iv) a number of structural RNA genes (Palmer 1991; Sugiura 1992; Bock 2007). Subsequent studies have identified the roles of additional genes not falling into any of these genes classes, including genes involved in post-transcriptional modification (*mat*K, Liere and Link 1995), protein turnover or protein complex assemblies (Peltier et al. 2004). Currently, only two genes remain, *ycf*1 and *ycf*2, whose metabolic or genetic roles have not yet been unambiguously defined (Bock 2007).

In this review, we will discuss functional and evolutionary insights from research on land plant plastid chromosomes, providing a synthesis of our knowledge of their evolution and conservation. Accordingly, particular emphasis will be placed on genetics of plastomes in the context of land plant diversification, with special attention to the roles of plastid-encoded proteins in photosynthesis and other principal genetic pathways.

2. PLASTID GENETICS AND SYNTENY OF LAND PLANT PLASTID GENOMES

2.1. Plastid inheritance

The transmission (inheritance) of plastids has been disputed for many years. For seed plants, mechanisms and occurrences of plastid inheritance have been studied in a great number of species (reviewed in Hagemann 2004; Bock 2007; Zhang and Sodmergen 2010). However, little is known about plastid transmission in earlier land plant lineages, probably due to methodological difficulties. Ultrastructural studies of functional sperm cells of bryophytes, lycophytes, horsetails and water ferns (heterosporous ferns) reported the presence of proplastids (reviewed in Sears 1980). In liverworts and mosses, the sperm cell's proplastids are "discarded" before fertilization (Sears 1980, and references therein). Maternal plastid transmission was subsequently demonstrated for the liverwort *Pellia*

(Pacak and Szweykowska-Kulińska 2002) and several moss representatives (*Rhizomnium*: Jankowiak et al. 2005; *Sphagnum*: Natcheva and Cronberg 2007; *Plagiomnium*: Jankowiak-Siuda et al. 2008). Maternal inheritance of plastids was shown for the horsetail *Equisetum variegatum* (Guillon and Raquin 2000), but nothing is known about the fate of the sperm cell's proplastid. Most, though probably not all, plastid-like structures are lost from the spermatozoids of lycophytes, and it seems as if there was a strong bias towards predominantly maternal plastid transmission caused by degradation prior or immediately after fertilization (Sears 1980). The absence of a plastid-like structure in sperm cells was shown in representatives of leptosporangiate ferns (*Pteridium*: Bell et al. 1966; *Thelypteris*: Sears 1980). This suggested maternal plastid transmission, which was later confirmed using molecular biological methods for *Cheilanthes* (Gastony and Yatskievych, 1992) and *Asplenium* (Vogel et al. 1998). In gymnosperms and angiosperms, uniparental inheritance is more frequent than biparental transmission (Hagemann 2004). Maternal inheritance is typical for angiosperms and the gymnosperm groups cycads and gnetophytes. In the majority of gymnosperms (conifers) paternal transmission is the dominant mode (Hagemann 2004; Zhang and Sodmergen 2010). However, biparental inheritance has evolved multiple times in seed plants, in particular in eudicot angiosperms such as Geraniaceae (e.g. Tilney-Bassett and Almouslem 1989), Campanulaceae (Corriveau and Coleman 1988) and Fabaceae (Corriveau and Coleman, 1988). In gymnosperms, biparental inheritance is much less frequent (Hagemann 2004).

2.2. Architecture of plastid chromosomes

In vivo structure and molecular conformation of the plastid chromosome has long been thought to be exclusively circular. However, several studies employing *in-situ* hybridization techniques demonstrated that often only a minor proportion of the molecules occur in a circular and covalently closed form. Instead, the majority of plastid chromosomes are arranged in concatemers of two or more molecules in either circularized or linear form (Deng et al. 1989; Bendich and Smith 1990; Bendich 1991, 2004; Harada et al. 1997; Lilly et al. 2001). It is still unknown how these concatemeric molecules are formed, and how linkage and breakage is carried out *in vivo*. It is speculated that the formation of these supermolecules might facilitate maintenance of gene organization and genome integrity (Day and Madesis 2007; Maréchal and Brisson 2010). However, the formation of supermolecules as a primary stabilizing factor needs to be evaluated carefully. Mitochondrial DNA forms concatemeric molecules as well, but exhibits a great variety of genome size and structure among land plants (Palmer and Herbon 1988; Bendich 2007).

The size of photosynthetic land plant plastid chromosomes ranges from 120 kb to 160 kb. The plastome in photosynthetic plants comprises 70 (gymnosperms) to 88

(liverworts) protein coding genes and 33 (most eudicots) to 35 (liverworts) structural RNA genes (Wakasugi et al. 1994; Ohyama 1996; Bock 2007), totaling 100–120 unique genes (Fig. II-1). The vast majority of these genes are arranged in operons (or operon-like structures) and transcribed as polycistronic precursor molecules that are subjected to splicing and nucleolytic cleavage in order to produce mature and translatable mRNAs (Stern et al. 2010). Functional gene classes (translation/transcription, electron transfer, and photosystems) are often arranged in close vicinity to one another (Fig. II-2; Cui et al. 2006). Using a parametric bootstrap-approach, Cui et al. (2006) showed that the genomic rearrangements of some chlorophytic algae (e.g. *Chlamydomonas*) relative to others are not random. Results indicated that the physical clustering of genes belonging to a similar functional class is positively selected. Furthermore, expression analysis indicated that some of these newly formed cluster are co-transcribed which led the authors to speculate that these could represent new regulons (Cui et al. 2006).

The plastid chromosome displays a quadripartite structure, i.e. it is divided into four major segments (Fig. II-2). Two of those contain only single copy (SC) genes and are referred to as Single Copy regions. The Large Single Copy region (LSC) harbors the majority of plastid genes; its smaller counterpart is known as the Small Single Copy region (SSC). The third segment is duplicated and exists in two nearly identical copies separating the SC regions (Kolodner and Tewari 1979). These copies are inverted and, therefore, termed large Inverted Repeats A and B (IR_A, IR_B). An IR is between 20 and 30 kb in size in angiosperms compared to only 10 to 15 kb in most non-seed plant lineages (Kolodner and Tewari 1979; Palmer 1991; Raubeson and Jansen 2005; Wu et al. 2009; Wolf et al. 2010). However, several lineages deviate strongly from the average, such as *Cycas* (25 kb, Wu et al. 2007), the cypress *Cryptomeria* (114 bp, Hirao et al. 2010) or the eudicot Geraniaceae (*Monsonia*: 7 kb, Guisinger et al. 2010; *Pelargonium*: 76 kb, Chumley et al. 2006). As the IRs are essentially identical, one might describe the plastid genome structure also as tripartite (as in Bock 2007), since the IRs share molecular evolutionary patterns that clearly differ from those observed in the SC regions. This quadripartite (or tripartite) architecture is already present in algal lineages including the closest relatives of land plants (e.g. *Chaetosphaeridium*, *Chara*; Turmel et al. 2002, 2006), implying a pre-land plant origin for this important conserved structural feature.

The plastid chromosomes of charophyte algae, the closest relatives of land plants (Qiu et al. 2006), are larger than those of land plants. They contain several genes that have either been lost or functionally transferred to the nuclear genome in Embryophytes (Turmel et al. 1999; 2006). Parsimony analyses reconstructing unambiguous changes in gene content among plants revealed that the gene *ycf*1 was gained in a common ancestor of several green algae and land plants (Maul et al. 2002). The gain of an intron in the *trn*K$_{UUU}$ coding regions, including an intact open reading frame (ORF; *mat*K), is shared by Charophytes and Embryophytes (Maul et al. 2002; Lewis and McCourt 2004; McNeal et al.

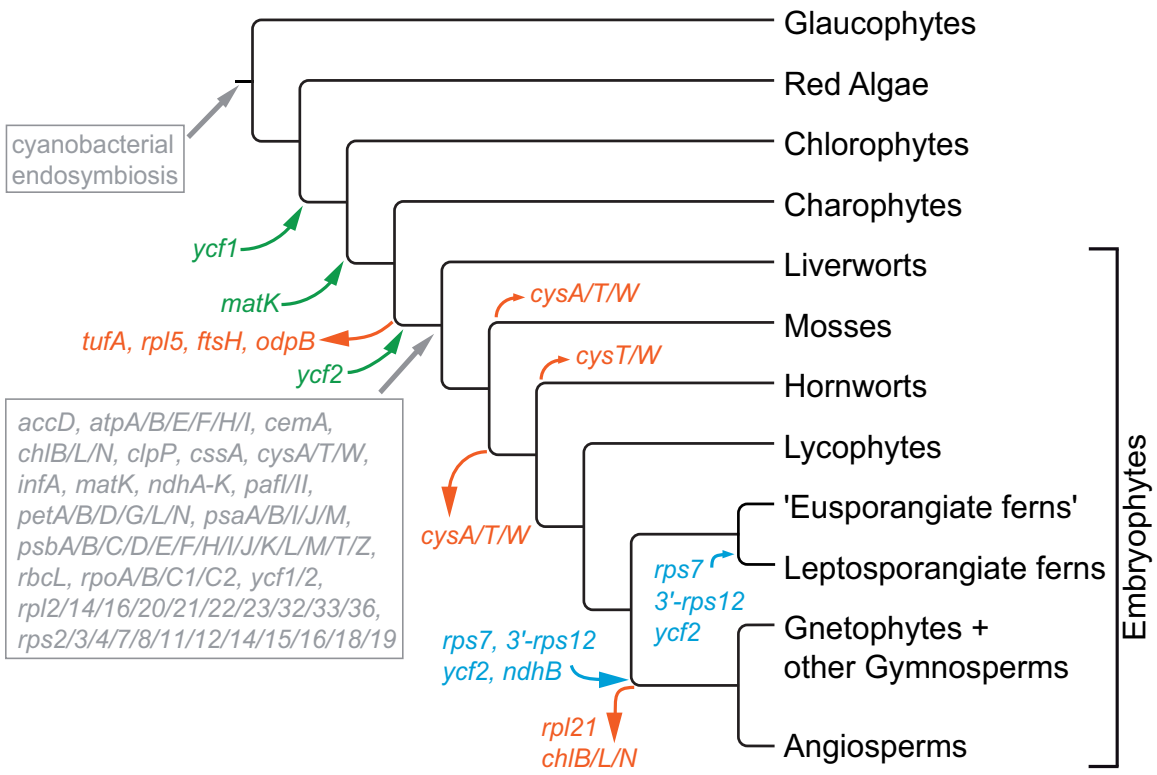


Fig.II-1 **Evolution of plastid gene content in land plants.** Events of gene losses in Embryophytes, as well as gains and duplication of protein coding genes in green plant lineages are depicted along the branches/nodes of the *Plant Tree of Life* (Palmer et al. 2004; Qiu et al. 2006; Zhong et al. 2010). The putatively ancestral gene content, as reflected in *Marchantia* and derived from parsimony analysis after Maul et al. (2002), is given at the first land plant node. Gene losses during the evolution of land plants are indicated by red arrows (those occurring before the emergence of Embryophytes are not considered here); a green arrow indicates the evolution of a novel gene prior to the transition to land; blue arrows refer to gene duplications. Changes in the content of transfer RNAs are not considered here (refer to Gao et al. 2010 for review). A detailed summary of gene losses during the evolution of angiosperms is provided by Jansen et al. (2007) and Magee et al. (2010). Although *chl*-subunits are still present in some gymnosperm plastomes, multiple losses and pseudogenizations indicate a functional transfer to the nuclear genome. As *chl* genes have been lost entirely from angiosperm plastomes, functional *chl*-gene transfer might have already occurred in a common ancestor.

2009). Comparative analysis revealed that the plastome structure and gene content in *Chaetosphaeridium*, a unicellular freshwater charophyte alga, is most similar to that of early land plants (Turmel et al. 2002): Large blocks of co-linear groups of genes are already present in this genus. Yet, in order to obtain the structural organization of early land plant plastomes, several functional gene transfers to the nuclear genome (e.g. *tuf*A, *fts*H, *odp*B, *rpl*5), one gene gain (*ycf*2), and a minimum of eight inversions are necessary (Turmel et al. 2006; Gao et al. 2010). One of those inversions involves a region of the LSC approximately 30 kb in length (Raubeson and Jansen 1992). A huge inversion of the complete *mat*K – *atp*A-I – *rpo*B-C1/2-region is shared between ferns and seed plants (Fig. II-2), whereas liverworts (Ohyama et al. 1988; Wickett et al. 2008), mosses (Sugiura et al. 2003; Oliver et al. 2010), hornworts (Kugita et al. 2003), and lycophytes (Wolf et al. 2005; Tsuji et al. 2007; Karol et al. 2010) show a more ancestral organization similar to that of *Chaetosphaeridium*

(Quandt and Stech 2002; Turmel et al. 2002). Generally, the presence of such rearrangements implies that additional transitional forms probably existed and might still be observable in lineages that have remained unstudied so far.

2.3. Gene Synteny and structural rearrangements

2.3.1. Plastome rearrangements.

Hotspots for structural rearrangements within plastid genomes include the IRs, which are frequently subject to expansion, contraction, or even complete loss. Such changes occurred several times independently during the evolution of land plants and often are specific for single orders and families, sometimes even for just one or a few species within a genus (Downie and Bewley 1992; Goulding et al. 1996; Plunkett and Downie 2000; Daniell et al. 2006; Guisinger et al. 2010; Wolf et al. 2010). Furthermore, extensive changes within the IRs appear to have an effect on the structural integrity of the entire plastid chromosome beyond the IRs and their immediate neighborhood. This is likely due to their role as putatively important players in the stabilization of the plastid chromosome via homologous recombination-induced repair mechanisms (Maréchal et al. 2009; Rowan et al. 2010; reviewed in detail by Maréchal and Brisson 2010).

Early branching gymnosperms (McCoy et al. 2008; Wu et al. 2009), angiosperms (Goremykin et al. 2003; Cai et al. 2006) and derived leptosporangiate ferns possess much larger IRs than the remaining land plant lineages (Wakasugi et al. 1998; Roper et al. 2007; Karol et al. 2010). Thus, large scale expansions of the IRs most likely occurred at least twice independently over the evolution of major land plant groups, including once in the common ancestor of seed plants. Additional large- (Guisinger et al. 2010) and small-scale (Goulding et al. 1996) expansions have occurred within angiosperms. Because of the re-location into the IR, several previously SC genes became duplicated, including the largest plastid gene, *ycf*2 (Wolf et al. 2010). A duplication of the *ycf*2 gene occurs independently in derived leptosporangiate ferns (tree and polypod ferns), and might be functionally relevant for plant development. In angiosperms, *ycf*2 expression is highest in fruits (Drescher et al. 2000), but comparable data for leptosporangiate ferns (or other land plant lineages) are lacking so far. Interestingly, plastome re-structuring in ferns is correlated with an expansion of the IR (Thompson et al. 1986; Stein et al. 1992; Raubeson and Stein 1995; Wolf et al. 2010). Contraction of the large inverted repeats involves only few (tens to hundreds of) base pairs up to and including complete IR loss. The positions of the LSC-IR junctions vary slightly within groups, but usually this has only negligible effects on plastome size (Goulding et al. 1996; Daniell et al. 2006; Wang et al. 2008). It has been suggested that such positional changes of IR-junctions among species are the result of gene

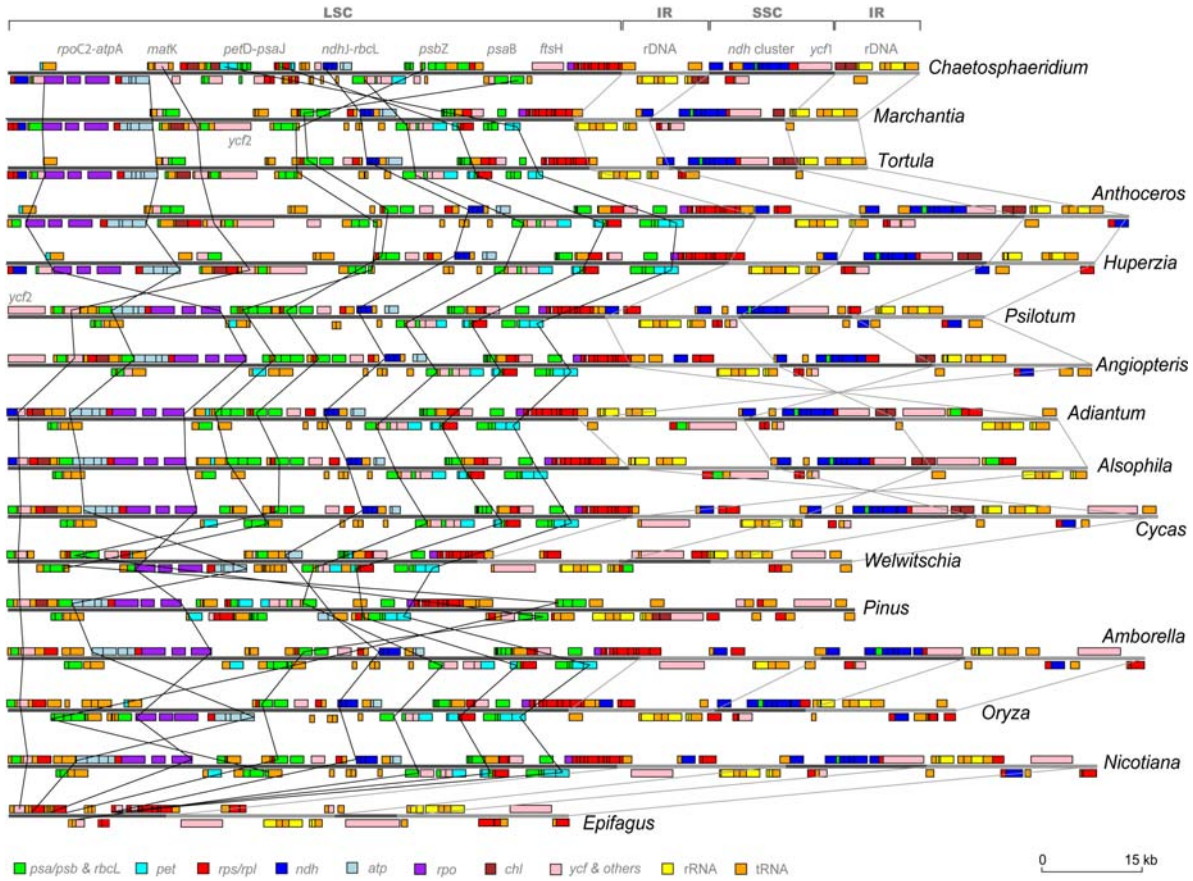


Fig. II-2 Synteny of land plant plastid chromosomes. The plastid chromosomes are shown in linearized form illustrating relative gene synteny. Genes are depicted by boxes colored according to their relevant functional class (see legend). Genes encoded by the leading strand (+ strand) or by the lagging strand (- strand) are shown above or below the grey chromosome bar, respectively. Lengths of boxes do not reflect lengths of genes, but are artificially increased to aid legibility (consequently, overlapping genes on +/- strand do not indicate overlapping reading frames). Lines from selected genes/gene-regions mentioned above the first chromosome bar roughly indicate genes clusters that have been reorganizated during land plant evolution. Not all regions that underwent genomic relocations prior or during land plant evolution are depicted here. The chromosome bars are colored gray to highlight the positions of the two large Inverted Repeat regions (IR_A/IR_B) and are connected by gray lines between the different lineages. Gray lines are discontinued once to indicate loss of the large inverted repeat in *Pinus*. Drawn with *GenomePixelizer* (Kozik et al. 2002) using genome annotations deposited in public sequence databases. Refer to the text for genome references and original publications.]

conversion (Goulding et al. 1996). In several groups, one of the IR-region has been completely lost, for instance in several legumes (Palmer et al. 1987b; Cai et al. 2008; Jansen et al. 2008; Tangphatsornruang et al. 2010), members of Geraniaceae (Guisinger et al. 2010), and some representatives of Orobanchaceae (Downie and Palmer 1992; S. Wicke, C. W. dePamphilis, D. Quandt and G. M. Schneeweiss, unpublished data). So far, no properties have been identified that are shared between these rather distantly related angiosperms and might provide an explanation for these IR losses. In legumes, the loss apparently affects overall structural stability, leading to mutational hotspots (Palmer et al. 1987b; Milligan et al. 1989; Cai et al. 2008; Magee et al. 2010) and an overall increase of nucleotide substitution rates (Perry and Wolfe 2002). The changes in gene order of a *Vigna angularis* cultivar relative to other members of Fabaceae have been proposed to either be caused by

a large inversion or mediated by a two-step model including IR expansion and contraction (Perry et al. 2002).

2.3.2. Small dispersed repeats.

Reorganizations are in many cases associated with small dispersed repeats (SDR), which are hypothesized to contribute to the double-strand break induced repair mechanism (Milligan et al. 1989; Maul et al. 2002; Odom et al. 2008). SDRs often contribute significantly to repeat space in genomes with highly rearranged gene order and add to structural polymorphism in even closely related lineages (Maul et al. 2002). SDRs mainly occur in non-coding DNA fractions (spacer, introns; Raubeson et al. 2007), where they are often associated with small hairpin structures (Quandt et al. 2003; Kim and Lee 2005). The greatest concentrations of SDRs have so far been reported in green algal plastid genomes (ca. 20 % of the *Chlamydomonas* plastome), although this seems to be highly lineage specific (Maul et al. 2002). Large repeats are assumed to be suppressed (or selectively eliminated) in plastid DNA because of their ability to cause recombination that may destabilize genome structure (Gray et al. 2009; Maréchal and Brisson 2010). Among angiosperms, the most abundant sizes of SDRs are on average smaller than 50 bp with direct repeats being more frequent than inverted repeats (Raubeson et al. 2007). A significant increase of repeats larger than the average has been reported in highly rearranged genomes such as Geraniaceae (Guisinger et al. 2010), Campanulaceae (Haberle et al. 2008), and Fabaceae (Cai et al. 2008), supporting the notion that repeats and genomic rearrangement are causally related. Possibly, tRNA genes might be recognized as repeated elements causing rearrangements by intramolecular or non-homologous recombination (Ogihara et al. 1988; Hiratsuka et al. 1989). In many cases, breakpoints of inversions are flanked by tRNA genes and short repetitive sequences (Hiratsuka et al. 1989; Haberle et al. 2008; Guisinger et al. 2010).

A unique switch in IR orientation (inversion) has occurred along the branch separating early diverging fern lineages *(Psilotum, Angiopteris:* Wakasugi et al. 1998; Roper et al. 2007; Karol et al. 2010) from derived leptosporangiate ferns (*Adiantum, Alsophila*: Wolf et al. 2003; Gao et al. 2009). This might be an outcome of the flip-flop recombination process proposed by Palmer (1983). Two smaller rearrangements occur at the breakpoint of the large inversion that is synapomorphic to all vascular plants except lycophytes (Raubeson and Jansen 1992; Wolf et al. 2003). The inversions reported in derived leptosporangiates are likely to be caused by two overlapping inversions during the evolution of leptosporangiate ferns (Wolf et al. 2003, 2010). Several small and large inversions that are not accompanied by expansion and contraction of an IR have been reported for diverse angiosperm lineages (Asteraceae: Jansen and Palmer 1987; Kim et al.

2005; *Spinacia*: Schmitz-Linneweber et al. 2001; some Oleaceae: Lee et al. 2007; Mariotti et al. 2010; grasses: Hiratsuka et al. 1989; Bortiri et al. 2008), but seem to be less frequent in early land plants lineages. However, one large inversion (71 kb), affecting nearly the entire LSC, is found in the model moss *Physcomitrella patens* (Sugiura et al. 2003). This inversion was shown to be autapomorphic to *Physcomitrella* and Funariales, but absent in other mosses (Goffinet et al. 2007). Due to the small number of plastid genomes sequenced from early land plant lineages, little is known about other structural rearrangements in bryophytes. As of this writing, no structural changes (inversions) have been identified in liverworts (L. L. Forrest and B. Goffinet, Ecology and Evolutionary Biology, University of Connecticut/USA, personal communication). Some of the largest inversions observed may be attributable to flip-flop recombination due to the existence of the large inverted repeats (Palmer 1983). In the few flowering plants studied so far, it has been shown that flip-flop recombination and inversions predominantly occur around the origin of replication (*ori*). In some angiosperms, the *ori*B maps to the rDNA-*ycf*1 region within the IR, which is located more closely to the IR-SSC-boundary than to the IR-LSC junction (Thompson et al. 1986; Lu et al. 1996; Kunnimalaiyaan and Nielsen 1997; Eisen et al. 2000; Mackiewicz et al. 2001).

2.3.3. Genome size reduction, gene transfer, and gene gains.

Genome size reduction is another major aspect of non-canonical structural evolution. The most dramatic changes in genome size and gene content have been reported for non-photosynthetic parasitic plants. The plastome of *Epifagus* (Wolfe et al. 1992) measures only about half the size of an average eudicot plastome (Bock 2007). This is mainly due to non-functionalization of most photosynthesis-related genes (dePamphilis and Palmer 1990) and some genes for transcription and translation (Morden et al. 1991). Although there is a general trend of (functional) plastid genome reduction in parasitic plants, the size and gene content seem to vary widely among different lineages because some highly heterotrophic species retain photosynthetic ability (Revill et al. 2005; Funk et al. 2007; McNeal et al. 2007; Nickrent and García 2009). Independent of parasitism, genome reduction was observed in Pinaceae and Gnetophytes (McCoy et al. 2008; Wu et al. 2009), due in large part to the loss of *ndh* genes. The plastomes of *Gnetum* and *Welwitschia* are also more compact than in other seed plant lineages due to the reduction of intron and spacer regions (McCoy et al. 2008; Wu et al. 2009). This genome reduction is speculated to be the result of a low-cost strategy that could facilitate rapid genome replication under disadvantageous environmental conditions (McCoy et al. 2008; Wu et al. 2009). Translocation of single genes is rare in plastid genomes, and this is likely a reflection of the overall rarity of inserted (vs. lost or rearranged) sequences in plastid genomes. Reports of

foreign DNA being naturally inserted into the plastid DNA are rare (Maul et al. 2002; Haberle et al. 2008; Guisinger et al 2010); perhaps in part because of the difficulty of detecting insertions in poorly conserved intergenic regions. Many of the repetitive elements found in highly rearranged genomes seem to be derived from plastid sequences (Cai et al. 2008; Haberle et al. 2008; Guisinger et al. 2010). However, some are unique which might suggest either rapid divergence or a non-plastid origin (Guisinger et al. 2010). As already mentioned by Park et al. (2007), the putatively horizontally acquired *rbc*L gene copies found in several *Phelipanche* species (Orobanchaceae) are most likely located in the nuclear or mitochondrial genome, and are not plastid encoded. *Rbc*L appears to be generally absent from *Phelipanche* plastid genomes (S. Wicke, D. Quandt, C. W. dePamphilis, G. M. Schneeweiss, unpublished data).

Gene gains, too, are exceptional during plant evolution (e.g. *mat*K, *ycf*1/2; Fig 1). The organization and regulation of genes in operons might be one stabilizing factor. Most often, localized changes of gene order are caused by the loss of single genes to the nuclear genome, or due to non-functionalization in parasitic or mycotrophic plants. Functional transfer of genes and subsequent loss of the plastid gene copy has been reported for some rosids (Jansen et al. 2010), some monocots (e.g. Hiratsuka et al. 1989; Masood et al. 2004; Saski et al. 2007) and the spikemoss *Selaginella uncinata* (Tsuji et al. 2007).

Contrasting with the overall high degree of conservation of plastome structure and gene content in land plants, massive structural changes are occasionally found in several unrelated lineages. These include derived angiosperm families such as Geraniaceae (Palmer et al. 1987a; Chumley et al. 2006; Guisinger et al. 2010), Fabaceae (Palmer et al. 1987b; Milligan et al. 1989; Cai et al. 2008; Tangphatsornruang et al. 2010), members of Onagraceae (*Oenothera*: Hupfer et al. 2000; Greiner et al. 2008), Campanulaceae (Knox and Palmer 1999; Cosner et al. 1997, 2004; Haberle et al. 2008), but also leptosporangiate ferns (Wolf et al. 2003, 2010; Gao et al. 2009). Because some of the extensively re-shuffled angiosperm plastomes occur in lineages with biparental plastid inheritance (Corriveau and Coleman 1988), it is tempting to speculate that the nature of plastid inheritance may affect plastid genome stability. Biparental inheritance combined with fusion of paternal and maternal plastids (although rare; Wellburn and Wellburn 1979) would likely result in homologous recombination between putatively divergent plastome copies (experimentally shown by Fejes et al. 1990), eventually leading to alteration of the genome structure. In other plants, major rearrangements, in particular gene losses, are obviously connected to a change in lifestyle from autotrophy to parasitism or myco-heterotrophy (*Aneura*: Wickett et al. 2008; Orobanchaceae: dePamphilis and Palmer 1990; Wolfe et al. 1992; Convolvulaceae: Funk et al. 2007; McNeal et al. 2007, 2009; Viscaceae: Nickrent and García 2009; and Lennoaceae: Y. Zhang and C.W. dePamphilis, unpublished data). The precise mechanisms underlying structural changes are as yet unknown, but they are often

associated with the presence of nearby repeat sequences, including small repeated sequences that are dispersed through the genome (Maul et al. 2002; Cui et al. 2006; Omar et al. 2008; Cai et al. 2008; Gray et al. 2009; Maréchal and Brisson 2010). Similarly to the plastid genome, in both the nuclear and mitochondrial genomes, structural reorganizations often are observed in proximity to structural RNA genes and short repetitive flanking sequence motifs (Grewe et al. 2009). In the nuclear genome, the latter is often associated with transposon activity (Woodhouse et al. 2010). In mitochondrial genomes, transposons are restricted to angiosperms (Knoop et al. 1996; Kubo et al. 2000; Notsu et al. 2002), but are absent in early land plant lineages (Ohyama 1996; Knoop 2004; Grewe et al. 2009). No (retro-) transposons, or traces thereof, have ever been reported from land plant plastomes. Yet, the plastid chromosome of the model green algae *Chlamydomonas* harbors two copies of the non-functional transposable element *Wendy* (Fan et al. 1995, Maul et al. 2002). Consequently, mechanisms suggested for nuclear and mitochondrial genomes are less likely for plastid genomes given the current knowledge on their evolution (reviewed in Palmer 1991; Raubeson and Jansen 2005; Bock 2007).

Other possible candidates for causing restructuring of plastid genomes are relaxed repair mechanisms and/or recombination processes. Recently, several nuclear encoded genes and gene families have been identified that mediate stabilization, repair and maintenance of the plastid chromosome (Day and Madesis 2007; Maréchal and Brisson 2010). It might be possible that mutations in these proteins could lead to impaired maintenance of the plastid genome structure (Guisinger et al. 2010).

3. GENE CONTENT AND FUNCTION OF THE PLASTID GENOME

The central function of the chloroplast is to carry out photosynthesis and carbon fixation. Besides genes encoding elements for the genetic apparatus, such as structural and transfer RNAs, the plastome encodes numerous proteins for photometabolic pathways (Palmer 1991; Sugiura 1992; Raubeson and Jansen 2005; Bock 2007). The following functional protein categories can be distinguished (Table II-A): proteins for the genetic apparatus, for non-photosynthesis related metabolic pathways, for primary (light-dependent) photosynthetic reactions, and for secondary (light-independent) photosynthesis pathways. In most cases, fully functional protein complexes are assembled from plastid encoded gene products and nuclear encoded subunits that are imported into the plastid organelle.

3.1. Plastid encoded elements for the plastid genetic apparatus

Many genes that encode pathways for the plastid genetic apparatus have been transferred to the nucleus and are now imported into the plastid. However, genes for transcription and protein biosynthesis are retained in the plastome. These comprise structural RNAs (rRNA, tRNA), some ribosomal proteins, and genes for a DNA-dependent RNA polymerase as well as few genes coding for DNA and protein processing enzymes.

3.1.1. Genes for DNA/RNA processing enzymes.

Plastid genetics is sometimes described as "chimeric" in that eukaryotic cytosolic (e.g. poly-A-binding proteins) and eubacterial components (e.g. Shine-Dalgarno interactions) are combined with novelties such as regulating stem loops in the 5'- and 3'-untranslated regions of plastid mRNAs (Zerges 2000). Transcription of plastid genes is carried out by a set of DNA-dependent RNA polymerases: nuclear encoded (phage-type) polymerase (NEP) and plastid-encoded (eubacterial-type) polymerase (PEP). Both transcribe distinct groups of genes (Hajdukiewicz et al. 1997; Cahoon and Stern 2001; Shiina et al. 2005) and require different transcription promoting signals (Weihe and Börner 1999). Promoter signals of PEP-transcribed genes are highly similar to those of eubacterial σ70-promoters with AT-rich sequences in the -35 promoter element (consensus 5'-TTGACA-3') and the -10 TATA-box (consensus 5'-TATAAT-3') upstream of the transcription initiation site (Briat et al. 1986). Promoter elements of NEP-transcribed genes are less conservative and share only short elements (Weihe and Börner 1999). Three different types are known. Two are characterized by a common core promoter YRT-element (i.e. purine-pyrimidine-thymidine stretch) that is highly conserved among flowering plants. This motif is localized in close proximity to the start codon (less than 10 bp away), where it can be preceded by a GAA-box. The different classes of promoters are recognized by two phage type polymerases. In *Arabidopsis*, the existence of at least two plastid targeted NEPs has been experimentally corroborated (Swiatecka-Hagenbruch et al. 2008), but evidence for differential usage or affinity to particular promoters is currently lacking. In eudicots, one of these NEPs is targeted to mitochondria *and* plastids (Kobayashi et al. 2001), which is reflected in partially shared promoter architectures between both organelles (Kühn et al. 2005). However, this dual-targeted phage type polymerase appears to be absent from other land plants including monocots and early diverging angiosperms (Yin et al. 2010).

Table II-A **Summary of plastid encoded genes in land plants.** Genes are divided primarily according their principal function (light-independent pathways, light-dependent pathways, genetic apparatus), and, secondarily according to the function of their respective subunits in a given protein. Where more than one subunit exists for a given gene class, subunit specifications are in alphanumeric order. The type of encoded protein is indicated as cytosolic, integral or extrinsic. The term 'cytosolic' is used here to describe localization in either the plastid stroma or lumen without associations to membranes. Genomic or genetic features as well as incidents of gene losses are provided for each gene. Refer to the text for further information about the function of certain genes. [Abbreviations: ne – nuclear encoded. References for structural characteristics/gene losses/pseudogenization are (if not stated otherwise): Jansen et al. 2007 and/or Magee et al. 2010 - photosynthetic angiosperms; Wolfe et al. (1992) - holoparasitic (non-photosynthetic) *Epifagus*; Wakasugi et al. 1994 and/or Wu et al. 2009 - Pinaceae and Gnetales; Wolf et al. 2010 - early vascular plants and ferns.]

Function	*Gene class*	*Subunits*	*Protein type*	*Functional/structural remarks*	**Losses and pseudogenizations**
light in-dependent ***- proteins related to photosynthetic dark reactions***	*cem*A inner membrane protein	--	integral	- mediates CO_2-uptake	- lost from *Epifagus*
	chl protochloro-phyllide reductase	B, L, N	--	--	- absent from angiosperm plastomes - lost/pseudogenized from some gymnosperms, e.g. *Gnetum, Welwitschia*
	ccsA cytochrome c biogenesis protein	--	--	mediates heme attachement to c-type cytochromes	- lost from *Epifagus*
	*rbc*L large subunit of RuBisCO	--	cytosolic	- presumably the most abundant protein on earth - primarily involved in photosynthetic carbon fixation - putative additional function in lipid acid metabolism, that is decoupled from photosynthesis	- pseudogenized in *Epifagus, Orobanche cernua*[3], *Hyobanche sp.*[4] - retained and expressed in *Cuscuta*[1] and non-photosynthetic broomrape *Lathraea*[2], some species of *Harveya*[4], *O. corymbosa*[4], *O. fasciculata*[4]; putatively functional in myco-heterotrophic *Aneura mirabilis*[5] but expression analysis are currently unavailable
- proteins not related to photosynthesis	*acc*D	--	cytosolic	- involved in lipid acid synthesis and not related to photosynthesis	- lost from Poales, *Acorus*, and *Trachelium*, some Geraniaceae[6], *Passiflora*
	cys putative ABC-containing sulfate transporter genes	A, T	-- extrinsic	- *cys*A: contains ATP-binding cassette and nucleotide binding motifs (Walker sites A, B) - involved in sulfate metabolism and not related to photosynthesis	- *cys*A present in the *Marchantia*[7] and *Anthoceros*[8] plastome, - *cys*T pseudogenized in *Aneura mirabilis*[5] -all subunits lost in mosses[9,10] and all vascular plants
light -dependent ***- proteins of the photosynthetic light reactions***	*atp* F-type ATP Synthase	A, B, E F, I, H	extrinsic integral	- extrinsic domains form catalytic complex F_1 - integral domain forms proton pumping complex F_0	- some subunits are lost from *Epifagus*, some are retained as pseudogenes
	ndh NAD(P)H dehydrogenase complex *(Ndh1)*	A, B, C, D, E, F, G, H, I, J, K	integral extrinsic	- subunits A-D and H-K homologous to mitochondrial complex I -subunits E, F, G are unique to plastid/ cyanobacterial *Ndh*1 complex – no homologs to any subunit of complex I	- *ndh* genes are either pseudogenized or lost from *Aneura mirabilis*[5], Pinaceae, Gnetales, some Erodium-species, orchids, *Epifagus*, some *Cuscuta*[1]; some members of carnivorous Lentibulariaceae[11]
	paf photosystem I assembly factor	I (*ycf*3), II (*ycf*4)	--	-*paf*I is indispensable for PSI assembly, whereas *paf*II might not be essential.	- *paf*II absent from some legumes; *paf* I/II absent from *Epifagus*
	pet cytochrome b_6/f complex	A, B, D, G, L, N	extrinsic integral	- A, B, D + ne PetC (Rieske protein) form core complex - G,L,N + ne PetM arranged peripherally around the core	- lost from *Epifagus* - some subunits are pseudogenized in *Cuscuta*-species[1], *and Aneura mirabilis*[5]

Function	*Gene class*	*Subunits*	*Protein type*	*Functional/structural remarks*	**Losses and pseudogenizations**
	psa photosystem I	A, B, C, I, J	integral	- extrinsic subunits are nuclear encoded	- some subunits are either lost from or pseudogenized in *Epifagus,* some are pseudogenized in *Cuscuta*-species[1], and *Aneura mirabilis*[5]
	psb photosystem II	A, B, C, D, E, F, H, I, J, K, L, M, N, T, Z	integral	- *psb*B/C form large extrinsic loops but are intrinsically bound to the membrane - extrinsic subunits *psb*O, P, Q are nuclear encoded	- some subunits are either lost from or pseudogenized in *Epifagus* - some subunits are pseudogenized in *Cuscuta*-species[1] *and Aneura mirabilis*[5]
genetic apparatus ***- proteins for transcription and post-transcriptional modification***	*mat*K maturase for most group-IIA-introns	--	cytosolic	- fast evolving gene with nearly equal substitution rates at all codon positions - only plastid gII-ORF localized in the trnK$_{UUU}$-intron	- lost from some *Cuscuta*-species[1] along with seven gIIA-introns - loss of the surrounding trnK$_{UUU}$ intron from most leptosporangiate ferns, some lycophytes[12]
	rpo DNA-dependent RNA polymerase	A, B, C1, C2	cytosolic	--	- *rpo*A highly diverged in Geraniaceae[6, 13], although remaining subunits are intact; - subunits are either lost from or pseudogenized in parasitic plants including *Epifagus,* and *Cuscuta*[1]
- proteins for translation and protein-modifying enzymes	*clp*P proteolytic subunit of Clp-protease	--	cytosolic	- contains two gII-introns, however, loss of introns occurs several times independently	- lost from *Passiflora, Scaevola*
	*inf*A translation initiation factor A	--	cytosolic		- pseudogenized in the lycophyte *Isoëtes*; - pseudogenized/lost in rosids, Solanales, *Lemna, Trachelium, Ranunculus, Cuscuta*[1]
	rpl large ribosomal proteins	2, 14, 16, 20, 21, 22, 23, 32, 33, 36	cytosolic	- primary 23S binding subunits[14]: L2, L16, L20, L22, 23, plus NE. L1, L3, L4, L9, L10, L11, L12, L15, L17, L18, L24, L29) - secondary 23S binding subunits[14]: , L14, L19, L21 plus NE L5, L6, L13, L27 - binding features unknown: L32, L33, L36, plus ne L31, L34 and 50S-PSRPs	- *rpl*21 lost from all seed plants - *rpl*22 lost from Fabaceae - *rpl23* pseudogenized in several angiosperms taxa - *rpl*32 lost from *Yucca, Populus*, some parasitic plants - *rpl*33 lost from *Phaseolus* - some *rpl* genes are pseudogenized or lost from some parasitic plant plastomes
	rps small ribosomal proteins	2, 3, 4, 7, 8, 11, 12, 14, 15, 16, 18, 19	cytosolic	- primary 16S binding subunits[14]: S4, S7, S8, S15, (ne S13, S20) - secondary 16S binding subunits[14]: S11, S12, S16, S18, S19, (ne S6, S9, S13, S5) - tertiary 16S subunits[14]: S14, S10, S3, S2, (ne S21) - binding unknown: 30S-PSRPs - 3′*rps*12 (exons 2/3) are trans-spliced to 5′-*rps*12 (exon 1)	- *rps*7 lost from *Passiflora* - *rps*11 lost from *Scaevola* - *rps*16 gene lost multiple times during land plant evolution, multiple intron losses - some *rps* genes are either pseudogenized or lost from some parasitic plant plastomes
- structural RNAs	*rrn* ribosomal RNAs	4.5S, 5S, 16S, 23S	--	--	--
	trn[15] tranfer RNAs	Ala$^{(UGC)}$ His$^{(GUG)}$, Lys$^{(UUU)}$, Gln$^{(UUG)}$, Cys$^{(GCA)}$, Asp$^{(GUC)}$, Glu$^{(UUC)}$, Tyr$^{(GUA)}$, fMet$^{(CAU)}$, Phe$^{(GAA)}$, Met$^{(CAU)}$, Trp$^{(CCA)}$, Asn$^{(GUU)}$, Gly$^{(UCC, GCC)}$, Thr$^{(GGU, UGU)}$, Ile$^{(GAU, CAU)}$, Val$^{(GAC, UAC)}$, Pro$^{(UGG, GGG)}$,	--	- anticodon of tRNA-Leu^{UAA} mutated to CAA, RNA-edited to UAA → loss of tRNA-Leu^{CAA} in some ferns	- trnP$_{GGG}$: loss is synapomorphic in angiosperms - tRNA-Arg^{CCG}: pseudogenized in gymnosperms plus Gnetales, absent from angiosperm plastomes - tRNA-SeC^{UCA} gene in *Adiantum* has not yet been found in any other Embryophyte plastome[15] - some tRNAs have been lost from the plastomes of parasitic plants

Function	*Gene class*	*Subunits*	*Protein type*	*Functional/structural remarks*	**Losses and pseudogenizations**
		$Arg^{(UCU,ACG,CCG)}$, $Leu^{(UAA,CAA,UAG)}$, $Ser^{(GCU,GGA,UAG)}$			
proteins of unknown function	*ycf1*	--	--	--	- lost from Poales, Acorales, *Passiflora* - highly diverged in some ferns and some members of carnivorous Lentibulariaceae[11]
	ycf2	--	--	--	- lost from Poales - highly diverged in some ferns and some members of carnivorous Lentibulariaceae[11]

1 - McNeal et al. 2007, Funk et al. 2007, McNeal et al. 2009; 2 - Delavault et al 1998; 3 - Wolfe and dePamphilis 1997, Leebens-Mack and dePamphilis 2002; 4 - Randle and Wolfe 2005; 5 - Wickett et al. 2008a, b; 6 - Guisinger et al. 2010; 7 - Ohyama 1996 ; 8 - Kugita et al. 2003; 9 - Sugiura et al. 2003; 10 - Oliver et al. 2010; –11 - B. Schäferhoff, S. Wicke, C. W. dePamphilis & K. F. Müller, unpublished.data.; 12 - Tsuji et al. 2007; 13 – Chumley et al. 2006; 14 – Refers to eubacterial-type ribosomes, see Grondek and Culver 2004 for assembly maps. No assembly maps are currently available for binding of plastid ribosomal proteins to 16S rRNA or 23S rRNA, respectively. PSRP do not harbor homologues in *E. coli* ribosomes. *E.coli* ribosomal proteins L7, L8, L30, L25, and L26 have not been detected in chloroplast ribosomes by Yamaguchi & Subramanian (2000); 15 - List may be incomplete due to additional changes/mutations or RNA-editing of anticodon sequences as reported from *Adiantum* by Wolf et al. 2004 A more thorough overview of tRNA changes among land plants is provided by Gao et al. 2010. The set of tRNAs shown here refers to the reference plastome *Nicotiana tabacum*, Shinozaki et al. 1986*).*

PEP is lost or pseudogenized in some parasitic plants with minimal or no photosynthetic activity such as *Cuscuta* (Funk et al. 2007; McNeal et al. 2007) and Orobanchaceae (Wolfe et al. 1992; Delavault et al. 1996). The loss of PEP subunits renders its promoters dispensable, potentially allowing them to be lost from the plastome (Krause et al. 2003). However, NEP seems to be able to take over at least some of PEP's transcriptional functions as suggested by the frequent presence of both NEP and PEP promoters upstream of several plastid transcription units, for instance in the *rrn*16-*trn*V region (Krause et al. 2003). In both *Cuscuta* (Berg et al. 2004) and *Lathraea* (Lusson et al. 1998) expression of the *rbc*L gene is accomplished by NEP after the loss of PEP.

3.1.2. matK – a general group IIA intron maturase?

Protein coding genes that are related to (post-) transcriptional activity include the *mat*K gene. The *mat*K-gene product is thought to act as a splicing factor for plastid group IIA (gIIA) introns (Liere and Link, 1995). It is commonly referred to as a 'general' maturase associated with several different intron-containing plastid mRNAs (Zoschke et al. 2010). MatK is transcribed from the sole intact plastid gII intron ORF localized between the exons coding for the lysine-tRNA (*trn*K_{UUU}). In contrast to other gII ORFs, MatK has lost domains assigned to a reverse transcriptase and endonuclease function. Similarity to typical gII ORF maturases is only retained in the DNA-binding domain (Mohr et al. 1993; San Filippo and Lambowitz 2002; Mohr and Lambowitz 2003; Lambowitz and Zimmerly 2004; Pyle and Lambowitz 2006; Hausner et al. 2006). The molecular evolution of the *mat*K codingregion is unusual compared to other plastid genes in that all three codon positions evolve at nearly equal rates (Hilu and Liang 1997). This feature makes it particularly useful for phylogenetic reconstruction (Müller et al. 2006; Wicke and Quandt 2009). Equal substitution rates at all codon positions, however, are indicative of relaxed purifying

selection (Müller et al. 2006; Duffy et al. 2009), which led several authors to question its function or functionality in land plants (Hausner et al. 2006). Substitution rate analysis, however, demonstrated purifying selection for *mat*K in parasitic lineages including Orobanchaceae (Young and dePamphilis 2000) and some *Cuscuta* species (McNeal et al. 2009), providing evidence for sustained functionality. In *Cuscuta*, however, *mat*K is absent from species (Funk et al. 2007; McNeal et al. 2007) that have lost all of the seven gIIA introns that likely depend upon the *mat*K maturase for splicing (McNeal et al. 2009; Zoschke et al. 2010), which lends further support to the hypothesis of a more general demand for the *mat*K-encoded maturase function.

3.1.3. Structural RNAs.

Reflecting their localization within the IR region, two sets of structural ribosomal RNA species (*rrn*23, *rrn*16, *rrn*5, *rrn*4.5) are encoded in most plastid genomes of green plants studied so far. The few exceptions with only one set occur in lineages that have lost one copy of the IR. The ancient duplication of the plastid ribosomal DNA operon and its conservation throughout plant evolution might be attributed to generally high quantities of rRNA required for ribosome synthesis during early developmental stages (Bendich 1987). The large ribosomal subunit (*rrn*23, cpLSU) is arranged upstream of the smallest ribosomal subunits of 4.5S (*rrn*4.5) and 5S RNA (*rrn*5), which might facilitate expression and delivery of either subunit at equal ratios. Moreover, the existence of two copies facilitates the maintenance of these genes by, e.g., gene conversion (Lemieux and Lee 1987). The small ribosomal subunit (*rrn*16, cpSSU) is separated from the remainder rRNAs by two tRNA genes. Functional domains of either rRNA species are highly conserved and show 65–80% similarity to eubacterial (cyanobacterial) ribosomal RNAs (Palmer 1985; Harris et al. 1994; Stoebe and Kowallik 1999; Zerges 2000).

30 different tRNAs are encoded in a typical angiosperm plastid genome. Recognition of all 61 codons is possible by superwobbling ("two out of three"-mechanism; Lagerkvist 1978; Pfitzinger et al. 1990; Rogalski et al. 2008). Superwobbling allows reading of all possible codons even if there is only *one* tRNA encoded as in the case of alanine, arginine, asparagine, aspartic acid, cysteine, glutamic acid, histidine, lysine, phenylalanine, proline, tryptophan, and tyrosine (Palmer 1991; Sugiura 1992; Bock 2007). In addition to protein biosynthesis, glutamyl tRNA (encoded by the plastid *trn*E gene) plays a prominent role during activation of heme biosynthesis (Smith 1988; Howe and Smith 1991; Jahn et al. 1992). This and the low rates of tRNA import into cell organelles (Dietrich et al. 1992, 1996; Lohan and Wolfe 1998) led Barbrook et al. (2006) to suggest that a minimal plastid genome would at least contain the *trn*E gene. However, experimental data concerning the import machinery for small structural RNAs are rare and evidence for

general tRNA import into plastids is lacking. Therefore, it remains speculative to what extent the plastid genome could eventually be reduced.

Nonphotosynthetic and minimally photosynthetic angiosperms typically retain only a fraction of tRNAs (Morden et al. 1991; Lohan and Wolfe 1998; Funk et al. 2007; McNeal et al. 2007, 2009; Nickrent and García 2009). In Orobanchaceae, the loss of some tRNA-genes, e.g. *trn*C, seems to be correlated with the loss of photosynthesis (Taylor et al. 1991). Because expression analyses of retained genes in the highly reduced plastomes of *Epifagus* (Wolfe et al. 1992) and *Conopholis* (Wimpee et al. 1991, 1992) suggest an intact translation apparatus, the loss of tRNAs from their genomes might be indicative of tRNA import into plastid organelles. Pseudogenization of tRNAs has been reported for the mistletoe *Arceuthobium* (Nickrent and Garcia 2009) and for *Cuscuta* (Funk et al. 2007; McNeal et al. 2007). In non-parasitic plants, the loss of *trn*Kuuu has occurred independently multiple times (*Selaginella:* Tsuji et al. 2007; leptosporangiate ferns: Duffy et al. 2009; Wolf et al. 2010; Gao et al. 2010; Geraniaceae: Guisinger et al. 2010).

3.1.4. Plastid ribosomal proteins and ribosomes.

Plastid protein biosynthesis is carried out at eubacterial-like 70S ribosomes (Zerges 2000). These are assembled from the small 30S ribosomal subunit and the large 50S subunit. The 16S ribosomal RNA builds the backbone of the 30S ribosome subunit, which additionally includes 25 ribosomal proteins (Yamaguchi et al. 2000). The remaining three plastid rRNA species together with 33 ribosomal proteins constitute the 50S ribosome subunit (Yamaguchi and Subramanian 2000). Most genes coding for ribosomal subunit proteins have been transferred to the nuclear genome. However, land plant plastomes commonly encode twelve proteins for the small ribosomal subunits (*rps* genes) and nine large ribosomal subunit proteins (*rpl* genes). Loss of *rps* and *rpl* genes from plastomes is rare, but has been detected in rosids (e.g. *rpl*22, *rpl*23, *rps*16; see Jansen et al. 2007, 2010, Magee et al. 2010 for an overview) and a variety of non-photosynthetic or minimally photosynthetic angiosperms (*Epifagus*: dePamphilis and Palmer 1990; *Conopholis:* Y. Zhang and C. W. dePamphilis, unpublished data; *Cuscuta:* Funk et al. 2007; McNeal et al. 2007; *Arceuthobium*: Nickrent and García 2009). Whether parasitic angiosperms are able to translate proteins with a reduced set of ribosomal proteins or import missing components is still unknown.

Other proteins associated with plastid ribosomes are a nuclear encoded ribosome recycling factor and several plastid ribosome specific proteins (PSRPs) that are unique to plants and show no similarities to bacterial homologs (Yamaguchi et al. 2000, 2003; Yamaguchi and Subramanian 2000; Sharma et al. 2010). The assembly of the eubacterial-

type ribosomes has been studied intensively (reviewed in Moore 1998), but so far no such studies are available for plastid ribosomes. Given the high similarity of ribosomal RNA and most ribosomal proteins between eubacteria and plastids, it can be assumed that plastid ribosome assembly is similar to that of eubacteria. Most of the ribosomal proteins of the 30S ribosome subunit bind to the so-called S7-branch or are dependent on other (plastid encoded) proteins for binding (Grondek and Culver 2004). Thus, through analogy with eubacterial ribosomal proteins, plastid encoded ones might be divided into primary, secondary and tertiary binding components of the 30S and the 50S (Table II-A) ribosome subunit according to their rRNA binding features.

Four proteins that are bound to the 30S ribosome subunits have no homologs in the eubacterial (i.e. *E. coli*-type) ribosome and are nuclear-encoded PSRPs. Two additional PSRP-proteins are bound to the 50S ribosome subunit (Yamaguchi et al. 2000; Yamaguchi and Subramanian 2000). It remains unknown how PSRPs are assembled into functional ribosome complexes. Recent analyses of PSRPs suggest that they play a role in light-dependent regulation of transcription/translation processes (Sharma et al. 2010).

One translation initiation factor assisting in the assembly of the translation initiation complex is encoded by the plastid gene *inf*A (translation initiation factor; a total of three are known from eubacterial translation mechanisms). *Inf*A has been lost multiple times independently during land plant evolution. Although present in all bryophyte and fern lineages, it is pseudogenized in the lycophyte *Isoëtes* (Karol et al. 2010), but appears to be functional in other lycophytes (*Selaginella*: Tsuji et al. 2007; *Huperzia*: Wolf et al. 2005). In angiosperms, multiple losses have been reported (summarized in Jansen et al. 2007; Magee et al. 2010), accumulating in lineages known for their non-canonical plastid genome evolution (e.g. legumes; Millen et al. 2001).

3.1.5. clpP – a protein-modifying enzyme.

High levels of photosynthetic gene expression coincide with an enormous protein turn-over in plastids. Both maturation and protein degradation involve ATP-dependent synthase/protease complexes that act as molecular chaperones restoring or degrading damaged proteins according to the severity of protein denaturation (Wawrzynow et al. 1996; Adam et al. 2001; Adam and Clarke 2002). In plastids, three different protease complexes have been identified: *Fts* (filamentation temperature sensitive protease), *Deg*P/*Htr*A (high temperature requirement protease A) and *Clp* (Caseinolytic protease). Whereas all subunits of the first two complexes are encoded by the nuclear genome, *Clp*P is plastid encoded.

3.2. Plastid genes coding for protein subunits involved in photosynthetic dark reactions and biogenesis

3.2.1. Genes for protochlorophyllide reductase subunits, proteins for CO2 uptake and cytochrome C biogenesis.

Bryophytes, lycophytes, ferns and most gymnosperms harbor genes for three subunits of a light-independent protochlorophyllide reductase (*chl*B, *chl*L, *chl*N) in their plastomes. This enzyme is involved in porphyrin and chlorophyll metabolism (Reinbothe and Reinbothe 1996; Karpinska et al. 1997). In gnetophytes, an aberrant gymnosperm group with still controversial phylogenetic position (e.g. Zhong et al. 2010), *chl*B, *chl*L and *chl*N are lost to different extents (McCoy et al. 2008; Wu et al. 2009). In *Ephedra*, sister group to the remaining Gnetales (Zhong et al. 2010), all three genes are present and intact, whereas *Gnetum* and *Welwitschia* possess pseudogenes of two subunits and have lost the third (McCoy et al. 2008; Wu et al. 2009). Different patterns in pseudogenization and *chl*-gene loss in both genera might indicate relaxation of evolutionary constraints to maintain functional copies, perhaps due to import of as yet unidentified nuclear substitutes.

The gene *ccs*A (*ycf*5) encodes a protein mediating the attachment of heme to c-type cytochromes during cytochrome biogenesis (Xie and Merchant 1996; Saint-Marcoux et al. 2009). The gene is localized in the plastid SSC region, and widely conserved among photosynthetic plants. However, *ccs*A is lost from *Epifagus* (Wolfe et al. 1992), and pseudogenized in *Aneura mirabilis* (Wickett et al. 2008) The reading frame is, however, retained in all *Cuscuta* species sequenced so far (McNeal et al. 2007, Funk et al. 2007).

Land plant plastomes also encode a protein localized in the inner envelope membrane (inner-envelope protein, *cem*A/*ycf*10; Sasaki et al. 1993b). Knockouts of the gene *cem*A in *Chlamydomonas* severely affected the uptake of CO_2, while not affecting photosynthetic reactions (Rolland et al. 1997). *Cem*A is lost from the plastid genome of *Epifagus* (Wolfe et al. 1992) and other Orobanchaceae (S. Wicke et al., unpublished data), but present in *Cuscuta* (Funk et al. 2007; McNeal et al. 2007), and *Aneura* (Wickett et al. 2008a)

3.2.2. rbcL.

The *rbc*L gene encodes the large subunit of the ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO). RuBisCO is estimated to be the most abundant protein on earth (Ellis 1979). With the assistance of chaperones, it is assembled from eight large subunits (RbcL) and eight small subunits (RbcS). In contrast to red algae and Glaucophytes, Chlorophytes and Streptophytes do not possess a functional gene copy for the small RuBisCO subunit (*rbc*S gene) in the plastid genome. Instead, RbcS is encoded as

a nuclear gene family and targeted to the plastid (Clegg et al. 1997). In contrast to many other photosynthesis related genes, *rbc*L is often retained in non-photosynthetic plants. Putatively functional copies of *rbc*L are retained in several representatives of Orobanchaceae, such as *Lathraea* (Delavault et al. 1996; Lusson et al. 1998), *Orobanche corymbosa, O. fasciculata* (Wolfe and dePamphilis 1997; Leebens-Mack and dePamphilis 2002), most species of *Harveya* (Leebens-Mack and dePamphilis 2002; Randle and Wolfe 2005), and the non-photosynthetic liverwort *Aneura mirabilis* (Wickett et al. 2008). In other broomrape species, *rbc*L is only found as a pseudogene (as in *Epifagus*: Wolfe et al. 1992, *O. cernua*: Wolfe and dePamphilis 1997; *Hyobanche*, Randle and Wolfe 2005), or has been completely lost (S. Wicke et al., unpublished data). Retention, expression, and evidence for strong purifying selection in hemiparasitic and some holoparasitic plants have led to the speculation that *rbc*L is involved in another, yet photosynthesis unrelated pathway (Leebens-Mack and dePamphilis 2002; Randle and Wolfe 2005; McNeal et al. 2007; see section II-3.4.1.).

3.3. Plastid genes for thylakoid complexes involved in photosynthetic light reactions

Oxygenic photosynthesis requires efficient light harvesting systems as well as an electron transport chain. The inner (thylakoid) membrane of the plastid contains at least five major protein complexes: photosystem I (PSI), photosystem II (PSII), cytochrome b_6/f complex, ATP synthase and an NAD(P)H-plastoquinone oxido-reductase-complex (summarized in Table II-A; Gounaris et al. 1986; Nixon et al. 1989).

3.3.1. Photosystem I and II (psa and psb genes).

In plants, light is harvested by two photosynthetic reaction centers, PSI and PSII. These are localized in the thylakoid membrane and form supercomplexes, each with its own light harvesting complex that absorbs light via antenna molecules (chlorophyll a/b, and carotenoids). The exact number of proteins and cofactors associated with PSI and PSII supercomplexes is not known. PSII contains at least 17 subunits, 15 of which are encoded by the plastid genome (*psb*A, B, C, D, E, F, H, I, J, K, L, M, N, T, Z). These genes are scattered across the LSC region. All plastid *psb*-gene products form transmembrane helices (Nelson and Yocum 2006) and bind to the subunits PsbA (syn. D1), B, C, and D (syn. D2; Eckardt 2001). The gene products of *psb*N and *psb*Z (syn. *ycf9*) supposedly interact with the chlorophyll-bound subunit *Psb*C that reaches into the thylakoid lumen (Nelson and Yocum 2006). The structure of PSI is less complex than that of PSII, because it contains fewer polypeptides in its reaction center. The genes encoding for its plastid encoded subunits (*psa* genes) are found in the LSC region with the exception of *psa*C, which is

embedded in an operon of ndh-genes in the plastome SSC region. Five subunits of plastid encoded PSI (A, B, C, I, J) are transmembrane proteins. The structurally highly similar apoproteins PsaA and PsaB bind to the iron-sulfur reaction center that mediates the transfer of excitated electrons from plastoquinone to ferrodoxin (Nelson and Yocum 2006). *Psa*C codes for a peripheral subunit on the stromal side of PSI, which is directly involved in ferrodoxin reduction by binding the terminal electron acceptor molecules and linking them to the PSI iron-sulfur center (Fischer et al. 1998). Subunits I and J are not essential for PSI function (Bock 2007).

3.3.2. Photosystem assembly factors (ycf3, ycf4).

Both photosystems have been shown to be assembled with the help of chaperones (Nelson and Yocum 2006). The products of two plastid genes, *ycf*3 (orf62) and *ycf*4 (orf184), function as assembly factors for the photosystem I complex (Boudreau et al. 1997a; Ruf et al. 1997; Naver et al. 2001; Ozawa et al. 2009). Mutations in certain amino acid residues that mediate protein-protein interactions led to decreasing levels of PSI accumulation in the thylakoid membrane, as did gene disruption experiments (Boudreau et al. 1997a). Recently, it has been shown that Ycf3 interacts with at least one nuclear encoded protein during the assembly of PSI (Albus et al. 2010). The naming of both genes is somewhat misleading as it implies that their function is still unknown. However, the transcripts of both ORFs are obviously translated and the resulting polypeptides assist during the assembly of the photosystem I. We therefore suggest renaming both genes to *PSI assembly factor I* (*paf*I, the former *ycf*3) and *II* (*paf*II, the former *ycf*4). The specifications I and II refer to the timing at which they are thought to interact with PSI following the model proposed by Ozawa et al. (2009).

3.3.3. Cytochrome b6f complex (pet genes) and ATP-Synthase complex (atp genes).

PSII and PSI are electrochemically connected in series by the cytochrome *b6/f* complex. This is a functional complex composed of nine different subunits plus several inorganic cofactors, such as chlorophyll a, heme, β-carotene and an iron-sulfur cluster (Baniulis et al. 2008). Six subunits are plastid-encoded (*pet*A, B, D, G, L, N). These participate in electron transfer, generating a proton gradient across the thylakoid membrane (Stroebel et al. 2003). Together with the nuclear encoded Rieske protein (PetC), the gene products of *pet*A (cytochrome f), petB (cytochrome b6) and *pet*D (subunit IV) form the core complex that acts in the linear electron transport (Kurisu et al. 2003). The remaining subunits (PetN, PetG, PetL plus nuclear encoded PetM, PetH) are hydrophobic molecules and are arranged peripherally around the core (Cramer et al. 2006).

Plastid ATP Synthase is a multi-subunit complex composed of nine different proteins generating ATP using the proton gradient. These constitute an integral membrane domain (F0-domain) and an extrinsic catalytic domain (F1-domain) reaching into the stroma (Mccarty 1992). The F1-subunit consists of five different polypeptides (α–ε), three of which are encoded by the plastome (*atp*A, B, E). The F0-domain involved in proton translocation is built from three different polypeptides (a–c) that are exclusively plastid encoded (*atp*F, I, H; Vollmar et al. 2009).

All plastid-encoded genes for the photosynthetic apparatus are highly conserved in land plant plastomes (with the exception of *ndh*A–K, see section 3.3.4.). Loss or pseudogenization have only been reported in non-photosynthetic parasitic (Krause 2008) or myco-heterotrophic (Wickett et al. 2008a, b) plants.

3.3.4. Plastid NAD(P)H-complex (ndh genes).

Electrons are recycled around PSI in different pathways. One of which is carried out by a plastid NAD(P)H-dehydrogenase complex (*Ndh*1-complex) incorporated in the thylakoid membrane (Casano et al. 2000; Nixon 2000). This complex might also be involved in chlororespiration, i.e. the process of respiratory electron transport in addition to and/or in interaction with the photosynthetic electron transport. The plastid *Ndh*-complex non-photochemically reduces and oxidizes plastoquinones. Furthermore, it may also mediate the transport of electrons from PSI-ferrodoxins back to PSII (reverse electron transport; Peltier and Cournac 2002). Subunit composition appear to be highly divergent between cyanobacteria and derived land plants (reviewed in Suorsa et al. 2009). Together with several partly uncharacterized subunits, *Ndh*1 consists of distinct subcomplexes ranging from ca. 500 to over 1000 kDa (Suorsa et al. 2009). Eleven subunits of the *Ndh*1-complex are encoded by the plastid genome (*ndh*A, B, C, D, E, F, G, H, I, J, K). Plastid subunits A-D as well as H-K are homologous to the eubacterial (mitochondrial) proton pumping complex I (Friedrich et al. 1995). Experimental studies have shown that plastid encoded *Ndh*1-subunits might not be essential for plant survival in tobacco, although *ndh*-gene knockouts did cause phenotypic alterations (Peltier and Cournac 2002 and references therein).The plastid encoded genes of the *Ndh*1 are pseudogenized or entirely lost several times during land plant evolution. Given current data, these losses seem to be predominantly connected to a heterotrophic lifestyle in land plants (parasitism, some forms of myco-heterotrophy). This includes the myco-heterotrophic and non-photosynthetic liverwort *Aneura mirabilis* (Wickett et al. 2008), the photosynthetic or partially non-photosynthetic parasitic *Cuscuta* (McNeal et al. 2007, Funk et al. 2007), the non-photosynthetic parasite *Epifagus* (dePamphilis and Palmer 1990), orchid species (Chang et al. 2006, Wu et al. 2010), and some gymnosperms (Wu et al. 2009) as well as

some representatives of carnivorous Lentibulariaceae (B. Schäferhoff, S. Wicke, C. W. dePamphilis and K. Müller, unpublished data), and some species of Geraniaceae (Blazier et al. 2011). *Ndh* genes are also absent from several chlorophyte algae genomes (incl. *Chlamydomonas*), but they are present in plastomes of the closest relatives of land plants (Turmel et al. 2006; see also Martín and Sabater 2010). The *Ndh1* complex may also be associated with other pathways, and might play an important role in adaptation to environmental stress (reviewed in Suorsa et al. 2009). Abiotic stress, such as nutrient starvation (in particular nitrogen starvation), affected and up-regulated *ndh*-gene expression indicating a putative regulating function of *Ndh*1 for the photosynthetic electron flow (Peltier and Schmidt 1991). Due to the presence of nuclear genes of *Arabidopsis* with strong similarities to *ndh* complexes and plastid transit peptide sequences (Peltier and Cournac 2002), the existence of a second, nuclear encoded plastid *ndh* complex (*Nda*2) has long been suspected. Recently, an alternative form of an plastid localized *Ndh*-complex involved in non-photochemical plastoquinone reduction was identified (Sirpiö et al. 2009; Takabayashi et al. 2009; Ishida et al. 2009; Suorsa et al. 2009, 2010). The existence of a second form might explain the multiple losses of *Ndh*1 genes from land plant plastomes. It may be that the function of an alternative *Ndh*-complex, or of fewer or incompletely assembled *Ndh*1-subcomplexes is sufficient for photosynthetic and related pathways in some, yet not all, plants - in particular, if they exhibit a certain degree of heterotrophy (e.g. myco-heterotrophy, parasitism, carnivory). It might be that nutrient supplies could affect the activity of the *Ndh*1 complex in a way that renders it dispensable. In the light of expression analyses under nitrogen starvation (Peltier and Schmidt 1991), the responsible factor may include the type of nitrogen source (nitrate vs. ammonium) or the excess of nitrogen (and/or other nutrients or even assimilates) obtained from a host plant. It is unclear whether this also accounts for the loss of *ndh* genes from the plastomes of Pinaceae, Gnetophytes and some Geraniaceae. As with many land plants, gymnosperms live in close association with mycorrhizae (Wang and Qiu 2006). Thus, it may be possible that fungal associations contribute to the fate of the plastid *ndh* genes. On the other hand, throughout land plants, the presence of mycorrhizae and *ndh* loss appear to be only imperfectly correlated; evidently, more data is necessary before sound conclusions can be drawn, since other reasons such as multiple independent functional gene transfers must be considered as well (see also Blazier et al. 2011).

3.4. Plastid encoded genes for photosynthesis unrelated pathways

Plastid genes for metabolic pathways unrelated to photosynthesis include proteins for fatty acid synthesis, and sulfur metabolism.

3.4.1. AccD and the RuBisCO "shunt".

Acetyl-CoA carboxylase is key enzyme in plastids mediating the irreversible conversion of acetyl-CoA to malonyl-CoA during the biosynthesis of fatty acids (Neuhaus and Emes 2010). The beta subunit of this multimeric enzyme (*acc*D) is encoded in the LSC of the plastome in Streptophytes (Sasaki et al. 1993a) and is considered to be crucial for leaf development (Kode et al. 2005). The *acc*D gene has been lost from the plastid genome several times in angiosperms (Jansen et al. 2007) where its function is fulfilled by nuclear copies (Nakkaew et al. 2008).

Recently, RuBisCO has been found to be involved in a previously unrecognized glycolysis bypassing reaction that converts carbohydrates to fatty acids at low carbon cost in oily seeds of white turnip (*Brassica rapa*, Schwender et al. 2004). This has been proposed as a likely reason for the retention of a photosynthetic pathway in parasitic species of *Cuscuta* that are fully heterotrophic, yet nonetheless would benefit from the RuBisCO "shunt" to enable rapid and efficient lipid synthesis (McNeal et al 2009).

3.4.2. Genes related to sulfur metabolism.

Liverworts contain at least two more protein coding genes absent from most other land plants, *cys*A and *cys*T. *Cys*A (designated *mbp*X in the *Marchantia polymorpha* plastid genome) shows functional domains similar to inner membrane sulfate ABC (*ATP binding cassette*) transporters. Although conservation of amino acid composition drops dramatically towards the N-terminus, similarity searches suggest that both genes might belong to sulfate related transport complexes or sulfate permeases and thus may have a function related to sulfate metabolism (Laudenbach and Grossman 1991). However, both subunits are lacking from most other land plant plastid genomes (mosses, ferns, seed plants) and the green algae *Chlamydomonas* (Sugiura 1992; Maul et al. 2002; Melis and Chen 2005; Lindberg and Melis 2008). In hornworts, a *cys*A-like gene is present in the plastid genome, but it appears to be non-functional (Kugita et al. 2003).

3.5. Plastid genes of unknown function

3.5.1. ycf1 and ycf2.

Green algae, including the closest relatives of Embryophytes, possess an *fts*H reading frame, which encodes a metalloprotease. Predominantly at the carboxyl-terminus, *fts*H exhibits similarities to the largest, yet functionally uncharacterized ORF (*ycf*2) in land plants (Wolfe 1994). Nucleotide sequence similarity among land plant *ycf*2 is

extraordinarily low compared to other plastid-encoded genes, being less than 50% across bryophytes, ferns, and seed plants. *Ycf*2 harbors nucleotide binding sites typical for green algal and eubacterial *fts*H and CDC48 gene families, which are involved in cell division processes, proteolysis, and protein transport (Wolfe 1994). Furthermore, a "DPAL"-motif, shared by CDC48 and *ycf*2, is highly conserved. In several angiosperm plastomes, a smaller ORF, *ycf*15, is present directly downstream of the *ycf*2 gene (Raubeson et al. 2007 and references therein). So far, an exact function has not been assigned to the *ycf*15 gene product. Expression studies in spinach suggested that *ycf*15 might act as a regulator for Ycf2 on the RNA level, but might not function on protein level (Schmitz-Linneweber et al. 2001). Consistent with an RNA-level function, Raubeson et al. (2007) showed that ycf15 is not under purifying selection as expected for most protein coding sequences. A non-protein function might also account for the conservation of the cryptic reading frame *ycf*68 found in the IRs of several angiosperms (Raubeson et al. 2007) and *Aneura mirabilis* (Wickett et al. 2008a). The persistence of both *ycf*15 and *ycf*68 ORFs might be attributable to their localization in the slowly evolving IR region.

*Ycf*1, the second largest gene in plastid genomes, codes for a protein of approximately 1800 amino acids, yet its precise function remains to be determined. Experimental data and comparisons of *Chlamydomonas* and angiosperm *ycf*1 homologs revealed conserved nucleotide binding sites (Boudreau et al. 1997b). Based on these data, functions of *ycf*1 and *ycf*2 have been hypothesized to involve ATPase-related activities, chaperone-function, activity in cell divisions (depicted from similarities with *fts*H) and structural remodeling and/or linkage of plastid chromosomes to protein and/or membrane structures (Wolfe 1994; Boudreau et al. 1997b). Available data on gene expression in tobacco show that, similar to *ycf*2, *ycf1* is expressed in fruits (Drescher et al. 2000). Products of both genes are essential for plant cell survival (Drescher et al. 2000; Boudreau et al. 1997b). In most land plant lineages, *ycf*1 and *ycf*2 genes have elevated substitution rates and may have undergone pseudogenization (Oliver et al. 2010; Wolf et al. 2010). For the most part, however, the 5′ end of both genes are are relatively conserved, whereas other parts seem to evolve more freely. In the case of *ycf*1, this might be due to the co-localization of a replication origin (*ori*B) in this region (Kunnimalaiyaan and Nielsen 1997). This implies that both genes seem to undergo at least weak selective constraints. Analyses regarding differences in d_n/d_s ratios and mutational hotspots within the genic region might corroborate the assignment of a function to both these genes. The losses observed in several photosynthetic lineages, however, raise the question whether they really carry out essential functions in *all* plants. Complete loss of both *ycf*1 and *ycf*2 from the plastomes of some (but not all) derived monocot lineages and putative pseudogenization in other plants (Downie et al. 1994) are in contrast to the high structural conservation in parasites (dePamphilis and Palmer 1990; Wolfe et al. 1992; McNeal et al. 2007). This might in fact point towards a function decoupled from photosynthesis. Nuclear encoded and plastid

targeted proteins similar to Ycf1/Ycf2 were not found in lineages where both genes have been lost from the plastid genome, such as Poaceae (Downie et al. 1994).

4. CONCLUSIONS

In terms of structure, land plant plastid chromosomes evolve much more slowly than their mitochondrial or nuclear counterparts. This structural conservatism might be a result of the common organization of genes in operons that are conserved features between cyanobacteria, green algae and land plants. Other relevant factors include the mode of plastid transmission, the activity of highly effective repair mechanisms, as well as the rarity of plastid fusion and fission. The latter property is one of the major differences relative to mitochondrial genomes that have been shown to frequently fuse, and in doing so, provide opportunities for exchanging divergent genome copies. Most plastome rearrangements appear to be restricted to lineages that show one or more of the following characteristics: (i) aberrant behavior of the inverted repeat region (expansion, contraction, loss), (ii) biparental plastid transmission; (iii) a high frequency of small dispersed repeat sequences, (iv) heterotrophic lifestyle (parasites, myco-heterotrophs). Among land plants, angiosperms show the greatest variation in plastome structure, although distortion of gene synteny by rearrangements and gene loss is still rare compared to the genomes of other cell compartments. Interestingly, plastid chromosome restructuring appears to occur most commonly in the more derived clades of a given lineage (leptosporangiate ferns, Funariales within mosses, Pinaceae and Gnetophytes within gymnosperms, eudicots and Poales within angiosperms). It will be interesting to see whether similar patterns occur in liverwort plastome evolution. The gene content of land plants does not appear to have dramatically changed, and only few gene losses or putative functional transfers (*chl, cys*) might have taken place in the course of land plant evolution. The retention of photosynthetically relevant genes might be attributable to several factors. On the one hand, functional gene transfer is a complex issue since it involves the transfer itself and the evolution of transit peptides; thus, it is expected to be rare. On the other hand, most protein subunits encoded by the plastome (in particular photosynthesis relevant proteins) harbor trans-membrane proteins, and might therefore be difficult to import (as known from mitochondria). Finally, many gene products are required at high expression levels and at early developmental stages (e.g. translation/transcription apparatus, photosynthesis genes) and their retention might be selected for.

5. ACKNOWLEDGMENTS

Thanks are especially due to Monika Ballmann (Bonn) and Ortrun Lepping (Münster) for assistance in retrieving genome annotations from public databases. We are very grateful to Norman Wickett (PSU), Paul Wolf (USU), Yan Zhang (PSU) and Josh Der (PSU) for fruitful and inspiring discussions on plastid genome evolution in land plants. We thank Laura Forrest and Bernard Goffinet (UConn) as well as Bastian Schäferhoff (IEB, Muenster) for sharing unpublished data with us. We appreciate helpful suggestions on an earlier version of this manuscript by two anonymous reviewers. Funding of our own research on parasitic plants and carnivores was obtained from the Austrian Science Fund (FWF grant 19404 to G.M.S), DFG grant MU2875/2, to K.F.M), and N.S.F. grants DEB-0120709 and DBI-0701748 to C.W.D.). Financial support to S.W. from the University of Vienna (Austria) and the Botanical Society of America is gratefully acknowledged.

6. AUTHOR CONTRIBUTIONS

S.W. and D.Q. conceived of the manuscript. S.W. drafted the manuscript; D.Q. critically revised the text. G.M.S., K.F.M. and C.W.D. contributed to the layout of the manuscript and critically revised it. S.W. designed the cover image of the respective journal issue.

This chapter was published in a modified version in a peer-reviewed science journal as:

Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D. 2011. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol 76:273-297.

REFERENCES

Adam Z, Adamska I, Nakabayashi K, Ostersetzer O, Haussuhl K, Manuell A, Zheng B, Vallon O, Rodermel SR, Shinozaki K, Clarke AK (2001) Chloroplast and mitochondrial proteases in *Arabidopsis*. A proposed nomenclature. Plant Physiol 125:1912–1918

Adam Z, Clarke AK (2002) Cutting edge of chloroplast proteolysis. Trends Plant Sci 7:451–456

Albus CA, Ruf S, Schottler MA, Lein W, Kehr J, Bock R (2010) Y3IP1, a nucleus-encoded thylakoid protein, cooperates with the plastid-encoded Ycf3 protein in photosystem I assembly of tobacco and *Arabidopsis*. Plant Cell 22:2838–2855

Ayliffe MA, Scott N, Timmis JN (1998) Analysis of plastid DNA-like sequences within the nuclear genomes of higher plants. Mol Biol Evol 15:738–745

Baniulis D, Yamashita E, Zhang H, Hasan SS, Cramer WA (2008) Structure-function of the cytochrome b_6/f complex. Photochem Photobiol 84:1349–1358

Barbrook AC, Howe CJ, Purton S (2006) Why are plastid genomes retained in non-photosynthetic organisms? Trends Plant Sci 11:101–108

Bell P, Frey-Wyssling A, Mühlethaler K (1966) Evidence for the discontinuity of plastids in the sexual reproduction of a plant. J Ultrastruct Res 15:108–121

Bendich AJ (1987) Why do chloroplasts and mitochondria contain so many copies of their genome? BioEssays 6:279–282

Bendich AJ (1991) Moving pictures of DNA released upon lysis from bacteria, chloroplasts, and mitochondria. Protoplasma 160:121–130

Bendich AJ (2004) Circular chloroplast chromosomes: The grand illusion. Plant Cell 16:1661–1666

Bendich AJ (2007) The size and form of chromosomes are constant in the nucleus, but highly variable in bacteria, mitochondria and chloroplasts. BioEssays 29:474–483

Bendich AJ, Smith SB (1990) Moving pictures and pulsed-field gel electrophoresis show linear DNA molecules from chloroplasts and mitochondria. Curr Genet 17:421–425

Berg S, Krause K, Krupinska K (2004) The *rbc*L genes of two *Cuscuta* species, *C. gronovii* and *C. subinclusa*, are transcribed by the nuclear-encoded plastid RNA polymerase (NEP). Planta 219:541–546

Blazier JC, Guisinger ME, Jansen RK (2011) Recent loss of plastid-encoded *ndh* genes within *Erodium* (Geraniaceae). Plant Mol Biol. doi: 10.1007/s11103-011-9753-5

Bock R (2007) Structure, function, and inheritance of plastid genomes. In Bock R (ed) Cell and Molecular Biology of Plastids, Springer, Berlin Heidelberg,pp 29–63

Bortiri E, Coleman-Derr D, Lazo G, Anderson O, Gu Y (2008) The complete chloroplast genome sequence of *Brachypodium distachyon*: sequence comparison and phylogenetic analysis of eight grass plastomes. BMC Res Notes 1:61

Boudreau E, Takahashi Y, Lemieux C, Turmel M, Rochaix J (1997a) The chloroplast *ycf*3 and *ycf*4 open reading frames of *Chlamydomonas reinhardtii* are required for the accumulation of the photosystem I complex. EMBO J 16:6095–6104

Boudreau E, Turmel M, Goldschmidt-Clermont M, Rochaix J, Sivan S, Michaels A, Leu S (1997b) A large open reading frame (orf1995) in the chloroplast DNA of *Chlamydomonas reinhardtii* encodes an essential protein. Mol Gen Genet 253:649–653

Briat J, Lescure A, Mache R (1986) Transcription of the chloroplast DNA: a review. Biochimie 68:981–990

Cahoon AB, Stern DB (2001) Plastid transcription a menage à trois? Trends Plant Sci 6:45–46

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

Cai Z, Penaflor C, Kuehl JV, Leebens-Mack JH, Carlson JE, dePamphilis CW, Boore JL, Jansen RK (2006) Complete plastid genome sequences of *Drimys, Liriodendron,* and *Piper*: implications for the phylogenetic relationships of magnoliids. BMC Evol Biol 6:77

Casano LM, Zapata JM, Martín M, Sabater B (2000) Chlororespiration and poising of cyclic electron transport. J Biol Chem 275:942–948

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

Chumley TW, Ferraris JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK (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. Mol Biol Evol 23:2175–2190

Clegg MT, Cummings MP, Durbin ML (1997) The evolution of plant nuclear genes. Proc Natl Acad Sci USA 94:7791–7798

Corriveau JL, Coleman AW (1988) Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperm species. Am J Bot 75:1443–1458

Cosner ME, Jansen RK, Palmer JD, Downie SR (1997) The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): Multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. Curr Genet 31:419–429

Cosner ME, Raubeson LA, Jansen RK (2004) Chloroplast DNA rearrangements in *Campanulaceae*: phylogenetic utility of highly rearranged genomes. BCM Evol Biol 4:27

Cui L, Leebens-Mack J, Wang LS, Tang J, Rymarquis L, Stern DB, dePamphilis CW 2006. Adaptive evolution of chloroplast genome structure inferred using a parametric bootstrap approach. BMC Evol Biol 6:13

Cramer WA, Zhang H, Yan J, Kurisu G, Smith JL (2006) Transmembrane traffic in the cytochrome b_6f complex. Annu Rev Biochem 75:769–790

Daniell H, Lee S, Grevich J, Saski C, Quesada-Vargas T, Guda C, Tomkins J, Jansen RK (2006) Complete chloroplast genome sequences of *Solanum bulbocastanum, Solanum lycopersicum* and comparative analyses with other Solanaceae genomes. Theor Appl Genet 112:1503–1518

Day A, Madesis P (2007) DNA replication, recombination, and repair in plastids. In Bock R (ed) Cell and Molecular Biology of Plastids, Springer, Berlin Heidelberg, pp 65–119

Delavault PM, Russo NM, Lusson NA, Thalouarn P (1996) Organization of the reduced plastid genome of *Lathraea clandestina*, an achlorophyllous parasitic plant. Physiol Plant 96:674–682

Deng X, Wing RA, Gruissem W (1989) The chloroplast genome exists in multimeric forms. Proc Natl Acad Sci USA 86:4156–4160

dePamphilis CW, Palmer JD (1990) Loss of photosynthetic and chlororespiratory genes from the plastid genome of a parasitic flowering plant. Nature 348:337–339

Dietrich A, Small I, Cosset A, Weil JH, Maréchal-Drouard L (1996) Editing and import: Strategies for providing plant mitochondria with a complete set of functional transfer RNAs. Biochimie 78:518–529

Dietrich A, Weil JH, Marechal-Drouard L (1992) Nuclear-encoded transfer RNAs in plant mitochondria. Ann Rev Cell Biol 8:115–131

Downie SR, Palmer JD (1992) Restriction site mapping of the chloroplast DNA inverted

repeat - a molecular phylogeny of the Asteridae. Ann Mo Bot Gard 79:266–283

Downie SR, Katz-Downie DS, Wolfe KH, Calie PJ, Palmer JD (1994) Structure and evolution of the largest chloroplast gene (ORF2280): internal plasticity and multiple gene loss during angiosperm evolution. Curr Genet 25:367–378

Drescher A, Ruf S, Calsa T, Carrer H, Bock R (2000) The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Plant J 22:97–104

Duffy AM, Kelchner SA, Wolf PG (2009) Conservation of selection on *mat*K following an ancient loss of its flanking intron. Gene 438:17–25

Eckardt NA (2001) A role for PsbZ in the core complex of photosystem II. Plant Cell 13:1245–1248

Eisen J, Heidelberg J, White O, Salzberg S (2000) Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol 1(6):Research0011

Ellis RJ (1979) The most abundant protein in the world. Trends Biochem Sci 4:241–244

Fan WH, Woelfle MA, and Mosig G (1995) Two copies of a DNA element, 'Wendy', in the chloroplast chromosome *of Chlamydomonas reinhardtii* between rearranged gene clusters. Plant Mol Biol 29:63–80

Fejes E, Engler D, Maliga P (1990) Extensive homologous chloroplast DNA recombination in the *Nicotiana* somatic hybrid. Theor Appl Genet 79:28–32

Fischer N, Hippler M, Setif P, Jacquot J, Rochaix J (1998) The PsaC subunit of photosystem I provides an essential lysine residue for fast electron transfer to ferredoxin. EMBO J 17:849–858

Friedrich T, Steinmüller K, Weiss H (1995) The proton-pumping respiratory complex I of bacteria and mitochondria and its homologue in chloroplasts. FEBS Lett 367:107–111

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

Gao L, Su Y, Wang T (2010) Plastid genome sequencing, comparative genomics, and phylogenomics: Current status and prospects. J Syst Evol 48:77–93

Gao L, Yi X, Yang Y, Su Y, Wang T (2009) Complete chloroplast genome sequence of a tree fern *Alsophila spinulosa*: insights into evolutionary changes in fern chloroplast genomes. BMC Evol Biol 9:130

Gastony GJ, Yatskievych G (1992) Maternal inheritance of the chloroplast and mitochondrial genomes in cheilanthoid ferns. Am J Bot 79:716–722

Goffinet B, Wickett NJ, Werner O, Ros RM, Shaw AJ, Cox CJ (2007) Distribution and

phylogenetic significance of the 71-kb inversion in the plastid genome in Funariidae (Bryophyta). Ann Bot 99:747–753

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

Goulding SE, Olmstead R, Morden CW, Wolfe KH (1996) Ebb and flow of the chloroplast inverted repeat. Mol Gen Genet 252:195–206

Gounaris K, Barber J, Harwood JL (1986) The thylakoid membranes of higher plant chloroplasts. Biochem J 237:313–326

Gray B, Ahner B, Hanson M (2009) Extensive homologous recombination between introduced and native regulatory plastid DNA elements in transplastomic plants. Transgenic Res 18:559–572

Greiner S, Wang X, Rauwolf U, Silber MV, Mayer K, Meurer J, Haberer G, Herrmann RG (2008) The complete nucleotide sequences of the five genetically distinct plastid genomes of *Oenothera,* subsection *Oenothera*: I. Sequence evaluation and plastome evolution. Nucl Acids Res 36:2366–2378

Grewe F, Viehoever P, Weisshaar B, Knoop V (2009) A trans-splicing group I intron and tRNA-hyperediting in the mitochondrial genome of the lycophyte *Isoëtes engelmannii*. Nucl Acids Res 37:5093–5104

Grondek JF, Culver GM (2004) Assembly of the 30S ribosomal subunit: Positioning ribosomal protein S13 in the S7 assembly branch. RNA 10:1861–1866

Guillon J, Raquin C (2000) Maternal inheritance of chloroplasts in the horsetail *Equisetum variegatum* (Schleich.). Curr Genet 37:53–56

Guisinger MM, Kuehl JV, Boore JL, Jansen RK (2010) Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Mol Biol Evol. doi:10.1093/molbev/msq229

Guo XY, Ruan SL, Hu WM, Ca DG, Fan LJ (2008) Chloroplast DNA insertions into the nuclear genome of rice: the genes, sites and ages of insertion involved. Funct Integr Genom 8:101–108

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

Hagemann R (2004) The sexual inheritance of plant organelles. In Daniell H, Chase C (eds) Molecular Biology and Biotechnology of Plant Organelles: Chloroplasts and Mitochondria, Springer, Dordrecht, The Netherlands, pp 104–105

Hajdukiewicz PT, Allison LA, Maliga P (1997) The two RNA polymerases encoded by the nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids. EMBO J 16:4041–4048

Harada T, Soma N, Ishikawa R, Niizeki M (1997) Topological resolution of rice plastid DNA by pulsed-field gel electrophoresis. Bull Fac Agric Hirosaki Univ 61:25–32

Harris EH, Boynton JE, Gillham NW (1994) Chloroplast ribosomes and protein synthesis. Microbiol Mol Biol Rev 58:700–754

Hausner G, Olsen R, Johnson I, Simone D, Sanders ER, Karol KG, McCourt RM, Zimmerly S (2006) Origin and evolution of the chloroplast *trn*K (*mat*K) intron: a model for evolution of group II intron RNA structures. Mol Biol Evol 23:380–391

Hilu KW, Liang HP (1997) The *mat*K gene: Sequence variation and application in plant systematics. Am J Bot 84:830–839

Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun CR, Meng BY (1989a) The complete sequence of the rice (*Oryza sativa*) chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Mol Gen Genet 217:185–194

Hirao T, Watanabe A, Kurita M, Kondo T, Takata K (2008) Complete nucleotide sequence of the *Cryptomeria japonica* D. Don. chloroplast genome and comparative chloroplast genomics: diversified genomic structure of coniferous species. BMC Plant Biol 8:70

Howe CJ, Smith AG (1991) Plants without chlorophyll. Nature 349:109

Hupfer H, Swiatek M, Hornung S, Herrmann RG, Maier RM, Chiu WL, Sears BB (2000) Complete nucleotide sequence of the *Oenothera elata* plastid chromosome, representing plastome I of the five distinguishable Euoenothera plastomes. Mol Gen Genet 263:581–585

Jahn D, Verkamp E, Söll D (1992) Glutamyl-transfer RNA: a precursor of heme and chlorophyll biosynthesis. Trends Biochem Sci 17:215–218

Jankowiak K, Rybarczyk A, Wyatt R, Odrzykoski I, Pacak A, Szweykowska-Kulinska Z (2005) Organellar inheritance in the allopolyploid moss *Rhizomnium pseudopunctatum*. Taxon 54:383–388

Jankowiak-Siuda K, Pacak A, Odrzykoski I, Wyatt R, Szweykowska-Kulińska Z (2008) Organellar inheritance in the allopolyploid moss *Plagiomnium curvatulum*. Taxon 57:145–152

Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, Leebens-Mack JH, Müller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, Chumley TW, Lee SB, Peery R,

McNeal JR, Kuehl JV, Boore JL (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA 104:19369–19374

Jansen RK, Palmer JD (1987) A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). Proc Natl Acad Sci USA 84:5818–5822

Jansen RK, Saski C, Lee S, Hansen AK, Daniell H (2010) Complete plastid genome sequences of three rosids (*Castanea, Prunus, Theobroma*): evidence for at least two independent transfers of *rpl*22 to the nucleus. Mol Biol Evol. doi: 10.1093/molbev/msq261

Jansen RK, Wojciechowski MF, Sanniyasi E, Lee S, Daniell H (2008) Complete plastid genome sequence of the chickpea (*Cicer arietinum)* and the phylogenetic distribution of *rps*12 and *clp*P intron losses among legumes (Leguminosae). Mol Phylogenet Evol 48:1204–1217

Karol KG, Arumuganathan K, Boore JL, Duffy AM, Everett KDE, Hall JD, Hansen SK, Kuehl JV, Mandoli D, Mishler BD, Olmstead RG, Renzaglia K, Wolf PG (2010) Complete plastome sequences of *Equisetum arvense* and *Isoetes flaccida*: implications for phylogeny and plastid genome evolution of early land plant lineages. BMC Evol Biol 10:321

Karpinska B, Karpinski S, Hällgren J (1997) The *chl*B gene encoding a subunit of light-independent protochlorophyllide reductase is edited in chloroplasts of conifers. Curr Genet 31:343–347

Keeling PJ (2010) The endosymbiotic origin, diversification and fate of plastids. Phil Trans R Soc B 365:729–748

Keeling PJ (2004) Diversity and evolutionary history of plastids and their hosts. Am J Bot 91:1481–1493

Kim KJ, Lee HL (2005) Widespread occurrence of small inversions in the chloroplast genomes of land plants. Mol Cell 19:104–113

Kim K, Choi K, Jansen RK (2005) Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae). Mol Biol Evol 22:1783–1792

Knoop V (2004) The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. Curr Genet 46:123–139

Knoop V, Unseld M, Marienfeld J, Brandt P, Sunkel S, Ullrich H, Brennicke A (1996) *Copia-*, *Gypsy-* and Line-like retrotransposon fragments in the mitochondrial genome of *Arabidopsis thaliana*. Genetics 142:579–585

Knox EB, Palmer JD (1999) The chloroplast genome arrangement of *Lobelia thuliniana*

(Lobeliaceae): expansion of the inverted repeat in an ancestor of the Campanulales. Plant Syst Evol 214:49–64

Kobayashi Y, Dokiya Y, Sugita M (2001) Dual targeting of phage-type RNA polymerase to both mitochondria and plastids is due to alternative translation initiation in single transcripts. Biochem Biophys Res Commun 289:1106–1113

Kode V, Mudd EA, Iamtham S, Day A (2005) The tobacco plastid *acc*D gene is essential and is required for leaf development. Plant J 44: 237–244

Kolodner R, Tewari KK (1979) Inverted repeats in chloroplast DNA from higher plants. Proc Natl Acad Sci USA 76:41–45

Kozik A, Kochetkova E, Michelmore R (2002) GenomePixelizer – a visualization program for comparative genomics within and between species. Bioinformatics 18:335–336

Krause K (2008) From chloroplasts to "cryptic" plastids: evolution of plastid genomes in parasitic plants. Curr Genet 54:111–121

Krause K, Berg S, Krupinska K (2003) Plastid transcription in the holoparasitic plant genus *Cuscuta*: parallel loss of the *rrn*16 PEP-promoter and of the *rpo*A and *rpo*B genes coding for the plastid-encoded RNA polymerase. Planta 216:815–823

Kubo T, Nishizawa S, Sugawara A, Itchoda N, Estiati A, Mikami T (2000) The complete nucleotide sequence of the mitochondrial genome of sugar beet (*Beta vulgaris* L.) reveals a novel gene for tRNACys(GCA). Nucl Acids Res 28:2571–2576

Kugita M, Kaneko A, Yamamoto Y, Takeya Y, Matsumoto T, Yoshinaga K (2003) The complete nucleotide sequence of the hornwort (*Anthoceros formosae*) chloroplast genome: insight into the earliest land plants. Nucl Acids Res 31:716–721

Kühn K, Weihe A, Börner T (2005) Multiple promoters are a common feature of mitochondrial genes in *Arabidopsis*. Nucl Acids Res 33:337–346

Kunnimalaiyaan M, Nielsen BL (1997) Fine mapping of replication origins (ori A and ori B) in *Nicotiana tabacum* chloroplast DNA. Nucl Acids Res 25:3681–3686

Kurisu G, Zhang H, Smith JL, Cramer WA (2003) Structure of the cytochrome b_6f complex of oxygenic photosynthesis: Tuning the cavity. Science 302:1009–1014

Lagerkvist U (1978) "Two out of three": an alternative method for codon reading. Proc Natl Acad Sci USA 75:1759–1762

Lambowitz AM, Zimmerly S (2004) Mobile group II introns. Annu Rev Genet 3388:1–35

Laudenbach DE, Grossman AR (1991) Characterization and mutagenesis of sulfur-regulated genes in a cyanobacterium: evidence for function in sulfate transport. J Bacteriol 173:2739–2750

Lee H, Jansen RK, Chumley TW, Kim K (2007) Gene relocations within chloroplast

genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. Mol Biol Evol 24:1161–1180

Leebens-Mack JH, dePamphilis CW (2002) Power analysis of tests for loss of selective constraint in cave crayfish and nonphotosynthetic plant lineages. Mol Biol Evol 19:1292–1302

Lemieux C, Lee RW (1987) Nonreciprocal recombination between alleles of the chloroplast 23S rRNA gene in interspecific *Chlamydomonas* crosses. Proc Natl Acad Sci USA 8844:4166–4170

Lewis LA, McCourt RM (2004) Green algae and the origin of land plants. Am J Bot 9911:1535–1556

Liere K, Link G (1995) RNA binding activity of the *mat*K protein encoded by the chloroplast trnK intron from mustard (*Sinapis alba*). Nucl Acids Res 23:917–921

Lilly JW, Havey MJ, Jackson SA, Jiang J (2001) Cytogenomic analyses reveal the structural plasticity of the chloroplast genome in higher plants. Plant Cell 13:245–254

Lindberg P, Melis A (2008) The chloroplast sulfate transport system in the green alga *Chlamydomonas reinhardtii*. Planta 228:951–961

Lohan AJ, Wolfe KH (1998) A subset of conserved tRNA genes in plastid DNA of nongreen plants. Genetics 150:425–433

Lu Z, Kunnimalaiyaan M, Nielsen BL (1996) Characterization of replication origins flanking the 23S rRNA gene in tobacco chloroplast DNA. Plant Mol Biol 32:693–706

Lusson NA, Delavault PM, Thalouarn P (1998) The *rbc*L gene from the non-photosynthetic parasite *Lathraea clandestina* is not transcribed by a plastid-encoded RNA polymerase. Curr Genet 34:212–215

Mackiewicz P, Mackiewicz D, Kowalczuk M, Cebrat S (2001) Flip-flop around the origin and terminus of replication in prokaryotic genomes. Genome Biol 2:Interactions1004

Magee AM, Aspinall S, Rice DW, Cusack BP, Sémon M, Perry AS, Stefanović S, Milbourne D, Barth S, Palmer JD, Gray JC, Kavanagh TA, Wolfe KH (2010) Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Res 20:1700–1710

Maréchal A, Brisson N (2010) Recombination and the maintenance of plant organelle genome stability. New Phytol 186:299–317

Maréchal A, Parent J, Véronneau-Lafortune F, Joyeux A, Lang BF, Brisson N (2009) Whirly proteins maintain plastid genome stability in *Arabidopsis*. Proc Natl Acad Sci USA 106:14693–14698

Mariotti R, Cultrera N, Munoz Diez C, Baldoni L, Rubini A (2010) Identification of new polymorphic regions and differentiation of cultivated olives (*Olea europaea* L.) through plastome sequence comparison. BMC Plant Biology 10:211

Martin W (2003) Gene transfer from organelles to the nucleus: Frequent and in big chunks. Proc Natl Acad Sci USA 100:8612–8614

Martin W, Kowallik KV (1999) Annotated English translation of Mereschkowsky's 1905 paper 'Über Natur und Ursprung der Chromatophoren im Pflanzenreiche'. Eur J Phycol 34:287–295

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

Masood S, Nishikawa T, Fukuoka S, Njenga PK, Tsudzuki T, Kadowaki K (2004) The complete nucleotide sequence of wild rice (*Oryza nivara*) chloroplast genome: first genome wide comparative sequence analysis of wild and cultivated rice. Gene 340:133–139

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

Maul JE, Lilly JW, Cui L, dePamphilis CW, Miller W, Harris EH, Stern DB (2002) The *Chlamydomonas reinhardtii* plastid chromosome: Islands of genes in a sea of repeats. Plant Cell 14:2659–2679

Mccarty RE (1992) A plant biochemist's view of H^+ ATPases and ATP synthases. J Exp Biol 172:431–441

McCoy S, Kuehl J, Boore J, Raubeson L (2008) The complete plastid genome sequence of *Welwitschia mirabilis*: an unusually compact plastome with accelerated divergence rates. BMC Evol Biol 8:130

McFadden GI, van Dooren GG (2004) Red algal genome affirms a common origin of all plastids. Curr Biol 14:R514–R516

McNeal JR, Kuehl J, Boore J, Leebens-Mack JH, dePamphilis CW (2009) Parallel loss of plastid introns and their maturase in the genus *Cuscuta*. PLoS ONE 4:e5982

McNeal JR, Kuehl J, Boore J, dePamphilis C (2007) Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus *Cuscuta*. BMC Plant Biol 7:57

Melis A, Chen H (2005) Chloroplast sulfate transport in green algae - genes, proteins and effects. Photosynth Res 86:299–307

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

Milligan BG, Hampton JN, Palmer JD (1989) Dispersed repeats and structural reorganization in subclover chloroplast DNA. Mol Biol Evol 6:355–368

Mohr G, Lambowitz AM (2003) Putative proteins related to group II intron reverse transcriptase/maturases are encoded by nuclear genes in higher plants. Nucl Acids Res 31:647–652

Mohr G, Perlman PS, Lambowitz AM (1993) Evolutionary relationships among group-II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. Nucl Acids Res 21:4991–4997

Moore PB (1998) The three-dimensional structure of the ribosome and its components. Ann Rev Biophys Biomol Struct 27:35–58

Morden CW, Wolfe KH, dePamphilis CW, Palmer JD (1991) Plastid translation and transcription genes in a nonphotosynthetic plant- Intact, missing and pseudogenes. EMBO J 10:3281–3288

Müller K, Borsch T, Hilu KW (2006) Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: Contrasting *mat*K, *trn*T-F, and *rbc*L in basal angiosperms. Mol Phylogenet Evol 41:99–117

Nakazono M, Hira A (1993) Identification of the entire set of transferred chloroplast DNA sequences in the mitochondrial genome of rice. Mol Gen Genet 236:341–346

Nakkaew A, Chotigeat W, Eksomtramage T, Phongdara A (2008) Cloning and expression of a plastid-encoded subunit, beta-carboxyltransferase gene (*acc*D) and a nuclear-encoded subunit, biotin carboxylase of acetyl-CoA carboxylase from oil palm (*Elaeis guineensis* Jacq.). Plant Sci 175:497–504

Natcheva R, Cronberg N (2007) Maternal transmission of cytoplasmic DNA in interspecific hybrids of peat mosses, *Sphagnum* (Bryophyta). J Evol Biol 20:1613–1616

Naver H, Boudreau E, Rochaix J (2001) Functional studies of Ycf3: Its role in assembly of photosystem I and interactions with some of its subunits. Plant Cell 13:2731–2745

Nelson N, Yocum CF (2006) Structure and function of photosystems I and II. Annu Rev Plant Biol 57:521–565

Neuhaus HE, Emes MJ (2010) Nonphotosynthetic metabolism in plastids. Annu Rev Plant Physiol Plant Mol Biol 51:111–140

Nickrent D, García M (2009) On the brink of holoparasitism: Plastome evolution in dwarf mistletoes (*Arceuthobium*, Viscaceae). J Mol Evol 68:603–615

Nixon PJ (2000) Chlororespiration. Philos Trans R Soc B 355:1541–1547

Nixon PJ, Gounaris K, Coomber SA, Hunter CN, Dyer TA, Barber J (1989) *psb*G is not a photosystem two gene but may be an *ndh* gene. J Biol Chem 264:14129–14135

Notsu Y, Masood S, Nishikawa T, Kubo N, Akiduki G, Nakazono M, Hirai A, Kadowaki K (2002) The complete sequence of the rice (*Oryza sativa* L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. Mol Genet Genomics 268:434–445

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

Odom OW, Baek K, Dani RN, Herrin DL (2008) *Chlamydomonas* chloroplasts can use short dispersed repeats and multiple pathways to repair a double-strand break in the genome. Plant J 53:842–853

Ogihara Y, Terachi T, Sasakuma T (1988) Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species. Proc Natl Acad Sci USA 85:8573–8577

Ohyama K (1996) Chloroplast and mitochondrial genomes from a liverwort, *Marchantia polymorpha* - gene organization and molecular evolution. J Mol Evol 60:16–24

Oliver M, Murdock A, Mishler BD, Kuehl J, Boore J, Mandoli D, Everett K, Wolf PG, Duffy A, Karol KG (2010) Chloroplast genome sequence of the moss *Tortula ruralis*: gene content, polymorphism, and structural arrangement relative to other green plant chloroplast genomes. BMC Genomics 11:143

Ozawa S, Nield J, Terao A, Stauber EJ, Hippler M, Koike H, Rochaix J, Takahashi Y (2009) Biochemical and structural studies of the large Ycf4-photosystem I assembly complex of the green alga *Chlamydomonas reinhardtii*. Plant Cell 21:2424–2442

Pacak A, Szweykowska-Kulińska Z (2002) Organellar inheritance in liverworts: An example of *Pellia borealis*. J Mol Evol 56:11–17

Palmer JD (1983) Chloroplast DNA exists in two orientations. Nature 301:92–93

Palmer JD (1985) Comparative organization of chloroplast genomes. Annu Rev Genet 1199:325–354

Palmer, JD (1991) Plastid chromosomes: structure and evolution. In: Bogorad L, Vasil IK (eds) Cell Culture and Somatic Genetics of Plant, Vol. 7A, Molecular Biology of Plastids, Academic Press, San Diego, pp 5–53

Palmer JD (2000) Molecular evolution: A single birth of all plastids? Nature 405:32–33

Palmer JD, Herbon LA (1988) Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. J Mol Evol 28:87–97

Palmer JD, Nugent JM, Herbon LA (1987a) Unusual structure of geranium chloroplast DNA: a triple-sized inverted repeat, extensive gene duplications, multiple inversions, and two repeat families. Proc Natl Acad Sci USA 84:769–773

Palmer JD, Osorio B, Aldrich J, Thompson WF (1987b) Chloroplast DNA evolution among legumes: Loss of a large inverted repeat occurred prior to other sequence rearrangements. Curr Genet 11:275–286

Palmer JD, Soltis DE, Chase MW (2004) The plant tree of life: An overview and some points of view. Am J Bot 91:1437–1445

Park J, Manen J, Schneeweiss GM (2007) Horizontal gene transfer of a plastid gene in the non-photosynthetic flowering plants *Orobanche* and *Phelipanche* (Orobanchaceae). Mol Phylogenet Evol 43:974–985

Peltier G, Cournac L (2002) Chlororespiration. Annu Rev Plant Biol 53:523–550

Peltier G, Schmidt GW (1991) Chlororespiration: an adaptation to nitrogen deficiency in *Chlamydomonas reinhardtii*. Proc Natl Acad Sci USA 88:4791–4795

Peltier J, Ripoll DR, Friso G, Rudella A, Cai Y, Ytterberg J, Giacomelli L, Pillardy J, van Wijk KJ (2004) Clp protease complexes from photosynthetic and non-photosynthetic plastids and mitochondria of plants, their predicted three-dimensional structures, and functional implications. J Biol Chem 279:4768–4781

Perry AS, Brennan S, Murphy DJ, Kavanagh TA, Wolfe KH (2002) Evolutionary re-organisation of a large operon in Adzuki bean chloroplast DNA caused by inverted repeat movement. DNA Res 99:157–162

Perry AS, Wolfe KH (2002) Nucleotide substitution rates in legume chloroplast DNA depend on the presence of the inverted repeat. J Mol Evol 55:501–508

Pfitzinger H, Weil JH, Pillay DTN, Guillemaut P (1990) Codon recognition mechanisms in plant chloroplasts. Plant Mol Biol 14**:** 805–814

Plunkett GM, Downie SR (2000) Expansion and contraction of the chloroplast inverted repeat in Apiaceae subfamily Apioideae. Syst Bot 25:648–667

Pyle AM, Lambowitz AM (2006) Group II introns: Ribozymes that splice RNA and invade DNA. In: Gesteland RF, Cech TR, Atkins JF (eds) The RNA World, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, USA, pp 449–505

Qiu Y, Li L, Wang B, Chen Z, Knoop V, Groth-Malonek M, Dombrovska O, Lee J, Kent L, Rest J, Estabrook GF, Hendry TA, Taylor DW, Testa CM, Ambros M, Crandall-

Stotler B, Duff RJ, Stech M, Frey W, Quandt D, Davis CC (2006) The deepest divergences in land plants inferred from phylogenomic evidence. Proc Natl Acad Sci USA 103:15511–15516

Quandt D, Müller K, Huttunen S (2003) Characterisation of the chloroplast DNA *psb*T-H region and the influence of dyad symmetrical elements on phylogenetic reconstructions. Plant Biol 55:400–410

Randle CP, Wolfe AD (2005) The evolution and expression of RBCL in holoparasitic sister-genera *Harveya* and *Hyobanche* (Orobanchaceae). Am J Bot 92:1575–1585

Raubeson LA, Stein DB (1995) Insights into fern evolution from mapping chloroplast genomes. Am Fern J 85:193–204

Raubeson LA, Jansen RK (1992) Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. Science 255:1697–1699

Raubeson LA, Jansen RK (2005) Chloroplast genomes of plants. In: Henry R (ed) Diversity and Evolution of Plants - Genotypic and Phenotypic Variation in Higher Plants, CABI Publishing, Wallingford, UK, pp 45–68

Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM, Boore JL, Jansen RK (2007) Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. BMC Genomics 88:174

Reinbothe S, Reinbothe C (1996) The regulation of enzymes involved in chlorophyll biosynthesis. Eur J Biochem 237:323–343

Revill MJW, Stanley S, Hibberd JM (2005) Plastid genome structure and loss of photosynthetic ability in the parasitic genus *Cuscuta*. J Exp Bot 56:2477–2486

Rogalski M, Karcher D, Bock R (2008) Superwobbling facilitates translation with reduced tRNA sets. Nat Struct Mol Biol 15:192–198

Rolland N, Dorne A, Amoroso G, Sultemeyer DF, Joyard J, Rochaix J (1997) Disruption of the plastid *ycf*10 open reading frame affects uptake of inorganic carbon in the chloroplast of *Chlamydomonas*. EMBO J 16:6713–6726

Roper JM, Kellon Hansen S, Wolf PG, Karol KG, Mandoli DF, Everett KDE, Kuehl J, Boore JL (2007) The complete plastid genome sequence of *Angiopteris evecta* (G. Forst.) Hoffm. (Marattiaceae). Am Fern J 97:95–106

Rowan BA, Oldenburg DJ, Bendich AJ (2010) RecA maintains the integrity of chloroplast DNA molecules in *Arabidopsis*. J Exp Bot 61: 2575–2588

Ruf S, Kössel H, Bock R (1997) Targeted inactivation of a tobacco intron–containing open reading frame reveals a novel chloroplast-encoded photosystem I–related gene. J Cell Biol 139:95–102

San Filippo J, Lambowitz AM (2002) Characterization of the C-terminal DNA-binding/DNA endonuclease region of a group II intron-encoded protein. J Mol Biol 324:933–951

Saint-Marcoux D, Wollman F, de Vitry C (2009) Biogenesis of cytochrome b_6 in photosynthetic membranes. J Cell Biol 185:1195–1207

Sasaki Y, Hakamada K, Suama Y, Nagano Y, Furusawa I, Matsuno R (1993a) Chloroplast-encoded protein as a subunit of acetyl-CoA carboxylase in pea plant. J Biol Chem 268:25118–25123

Sasaki Y, Sekiguchi K, Nagana Y, Matsumo R (1993b) Chloroplast envelope protein encoded by the chloroplast genome. FEBS Lett 316:93–98

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

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

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

Schwender J, Goffman F, Ohlrogge JB, Shachar-Hill Y (2004) Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds. Nature 432:779–782

Sears BB (1980) Elimination of plastids during spermatogenesis and fertilization in the plant kingdom. Plasmid 44:233–255

Shahmuradov IA, Akbarova YY, Solovyev VV, Aliyev JA (2003) Abundance of plastid DNA insertions in nuclear genomes of rice and *Arabidopsis*. Plant Mol Biol 52:923–934

Sharma MR, Dönhöfer A, Barat C, Marquez V, Datta PP, Fucini P, Wilson DN, Agrawal RK (2010a) PSRP1 is not a ribosomal protein, but a ribosome-binding factor that is recycled by the ribosome-recycling factor (RRF) and elongation factor G (EF-G). J Biol Chem 285:4006–4014

Sheppard AE, Timmis JN (2009) Instability of plastid DNA in the nuclear genome. PLoS Genet 55:e1000323

Shiina T, Tsunoyama Y, Nakahira Y, Khan MS (2005) Plastid RNA polymerases, promoters, and transcription regulators in higher plants. Int Rev Cytol 244:1–68

Smith AG (1988) Subcellular localization of two porphyrin-synthesis enzymes in *Pisum sativum* (pea) and *Arum* (cuckoo-pint) species. Biochem J 249:423–428

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

Stein DB, Conant DS, Ahearn ME, Jordan ET, Kirch SA, Hasebe M, Iwatsuki K, Tan MK, Thomson JA (1992) Structural rearrangements of the chloroplast genome provide an important phylogenetic link in ferns. Proc Natl Acad Sci USA 89:1856–1860

Stern DB, Astwood JD (1986) Tripartite mitochondrial genome of spinach: physical structure, mitochondrial gene mapping, and locations of transposed chloroplast DNA sequences. Nucl Acids Res 14:5651–5666

Stern DB, Goldschmidt-Clermont M, Hanson MR (2010) Chloroplast RNA metabolism. Annu Rev Plant Biol 61:125–155

Stern DB, Lonsdale DM (1982) Mitochondrial and chloroplast genomes of maize have a 12-kilobase DNA sequence in common. Nature 299:698–702

Stoebe B, Kowallik KV (1999) Gene-cluster analysis in chloroplast genomics. Trends Genet 15:344–347

Stroebel D, Choquet Y, Popot J, Picot D (2003) An atypical haem in the cytochrome b_6f complex. Nature 426:413–418

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

Sugiura M (1992) The chloroplast genome. Plant Mol Biol 19:149–168

Swiatecka-Hagenbruch M, Emanuel C, Hedtke B, Liere K, Borner T (2008) Impaired function of the phage-type RNA polymerase RpoTp in transcription of chloroplast genes is compensated by a second phage-type RNA polymerase. Nucl Acids Res 36:785–792

Tangphatsornruang S, Sangsrakru D, Chanprasert J, Uthaipaisanwong P, Yoocha T, Jomchai N, Tragoonrung S (2010) The chloroplast genome sequence of mungbean (*Vigna radiata*) determined by high-throughput pyrosequencing: Structural organization and phylogenetic relationships. DNA Res 17:11–22

Taylor GW, Wolfe KH, Morden CW, dePamphilis CW, Palmer JD (1991) Lack of a

functional plastid $tRNA^{Cys}$ gene is associated with loss of photosynthesis in a lineage of parasitic plants. Curr Genet 20:515–518

Thompson WF, Stein DB, Palmer JD (1986) Structural evolution and flip-flop recombination of chloroplast DNA in the fern genus *Osmunda*. Curr Genet 10:835–841

Tilney-Bassett RAE, Almouslem AB (1989) Variation in plastid inheritance between pelargonium cultivars and their hybrids. Heredity 63:145–153

Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet 55:123–135

Tsuji S, Ueda K, Nishiyama T, Hasebe M, Yoshikawa S, Konagaya A, Nishiuchi T, Yamaguchi K (2007) The chloroplast genome from a lycophyte (microphyllophyte), *Selaginella uncinata*, has a unique inversion, transpositions and many gene losses. J Plant Res 120:281–290

Turmel M, Otis C, Lemieux C (2002) The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: Insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. Proc Natl Acad Sci USA 99:11275–11280

Turmel M, Otis C, Lemieux C (2006) The chloroplast genome sequence of *Chara vulgaris* sheds new light into the closest green algal relatives of land plants. Mol Biol Evol 23:1324–1338

Vogel JC, Russel SJ, Rumsey FJ, Barrett JA, Gibby M (1998) Evidence for maternal transmission of chloroplast DNA in the genus *Asplenium* (Aspleniaceae, Pteridophyta). Bot Acta 111:247–249.

Vollmar M, Schlieper D, Winn M, Büchner C, Groth G (2009) Structure of the c14 rotor ring of the proton translocating chloroplast ATP synthase. J Biol Chem 284:18228–18235

Wakasugi T, Nishikawa A, Yamada K, Sugiura M (1998) Complete nucleotide sequence of the plastid genome from a fern, *Psilotum nudum*. Endocytobiosis Cell Res 13(Suppl):147

Wakasugi T, Tsudzuki J, Ito S, Nakashima K, Tsudzuki T, Sugiura M (1994) Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the Black Pine *Pinus thunbergii*. Proc Natl Acad Sci USA 91:9794–9798

Wang B, Qiu Y (2006) Phylogenetic distribution and evolution of mycorrhizas in land plants. Mycorrhiza 16:299–363

Wang R, Cheng C, Chang C, Wu C, Su T, Chaw S (2008) Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. BMC Evol Biol 8:36

Wawrzynow A, Banecki B, Zylicz M (1996) The Clp ATPases define a novel class of molecular chaperones. Mol Microbiol 21:895–899

Weihe A, Börner T (1999) Transcription and the architecture of promoters in chloroplasts. Trends Plant Sci 4:169–170

Wellburn FAM, Wellburn AR (1979) Conjoined mitochondria and plastids in the barley mutant 'albostrians'. Planta 147:178–179

Wicke S, Quandt D (2009) Universal primers for the amplification of the plastid *trn*K/*mat*K region in land plants. Anales Jard Bot Madrid 66:285–288

Wickett NJ, Zhang Y, Hansen SK, Roper JM, Kuehl JV, Plock SA, Wolf PG, dePamphilis CW, Boore JL, Goffinet B (2008a) Functional gene losses occur with minimal size reduction in the plastid genome of the parasitic liverwort *Aneura mirabilis*. Mol Biol Evol 25:393–401

Wickett NJ, Fan Y, Lewis P, Goffinet B (2008b) Distribution and evolution of pseudogenes, gene losses, and a gene rearrangement in the plastid genome of the nonphotosynthetic liverwort, *Aneura mirabilis* (Metzgeriales, Jungermanniopsida). J Mol Evol 67:111–122

Wimpee C, Wrobel R, Garvin D (1991) A divergent plastid genome in *Conopholis americana*, an achlorophyllous parasitic plant. Plant Mol Biol 17:161–166

Wimpee CF, Morgan R, Wrobel RL (1992) Loss of transfer RNA genes from the plastid 16S–23S ribosomal RNA gene spacer in a parasitic plant. Curr Genet 21:417–422

Wolf PG, Rowe CA, Sinclair RB, Hasebe M (2003) Complete nucleotide sequence of the chloroplast genome from a leptosporangiate fern, *Adiantum capillus-veneris* L. DNA Res 10:59–65

Wolf PG, Rowe CA, Hasebe M (2004) High levels of RNA editing in a vascular plant chloroplast genome: analysis of transcripts from the fern Adiantum capillus-veneris. Gene 339:89–97

Wolf PG, Karol KG, Mandoli DF, Kuehl J, Arumuganathan K, Ellis MW, Mishler BD, Kelch DG, Olmstead RG, Boore JL (2005) The first complete chloroplast genome sequence of a lycophyte, *Huperzia lucidula* (Lycopodiaceae). Gene 350:117–128

Wolf PG, Der J, Duffy A, Davidson J, Grusz A, Pryer KM (2010) The evolution of chloroplast genes and genomes in ferns. Plant Mol Biol. doi: 10.1007/s11103-010-9706-4

Wolf PG, Roper JM, Duffy AM (2010) The evolution of chloroplast genome structure in ferns. Genome 53:731–738

Wolfe AD, dePamphilis CW (1997) Alternate paths of evolution for the photosynthetic

gene *rbc*L in four nonphotosynthetic species of *Orobanche*. Plant Mol Biol 3:965–977

Wolfe KH (1994) Similarity between putative ATP-binding sites in land plant plastid ORF2280 proteins and the FtsH/CDC48 family of ATPases. Curr Genet 25:379–383

Wolfe KH, Morden CW, Palmer JD (1992) Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc Natl Acad Sci USA 89:10648–10652

Woodhouse MR, Pedersen B, Freeling M (2010) Transposed genes in *Arabidopsis* are often associated with flanking repeats. PLoS Genet 66:e1000949

Wu C, Lai Y, Lin C, Wang Y, Chaw S (2009) Evolution of reduced and compact chloroplast genomes (cpDNAs) in gnetophytes: Selection toward a lower-cost strategy. Mol Phylogenet Evol 52:115–124

Wu CS, Wang YN, Liu SM, Chaw SM (2007) Chloroplast genome (cpDNA) of *Cycas taitungensis* and 56 cp protein-coding genes of *Gnetum parvifolium*: Insights into cpDNA evolution and phylogeny of extant seed plants. Mol Biol Evol 24:1366–1379

Wu F, Chan M, Liao D, Hsu C, Lee Y, Daniell H, Duvall M, Lin C (2010) Complete chloroplast genome of *Oncidium* Gower Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. BMC Plant Biol 10:68

Xie Z, Merchant S (1996) The plastid-encoded *ccs*A gene is required for heme attachment to chloroplast c-type cytochromes. J Biol Chem 271:4632–4639

Yamaguchi K, Beligni MV, Prieto S, Haynes PA, McDonald WH, Yates JR, Mayfield SP (2003) Proteomic characterization of the *Chlamydomonas reinhardtii* chloroplast ribosome. J Biol Chem 278:33774–33785

Yamaguchi K, von Knoblauch K, Subramanian AR (2000) The plastid ribosomal proteins-Identification of all the proteins in the 30 S subunit of organelle ribosome (chloroplast). J Biol Chem 275:28455–28465

Yamaguchi K, Subramanian AR (2000) The plastid ribosomal proteins - Identification of all the proteins in the 50 S subunit of organelle ribosome (chloroplast*).* J Biol Chem 275:28466–28482

Yin C, Richter U, Börner T Weihe A (2010) Evolution of plant phage-type RNA polymerases: The genome of the basal angiosperm *Nuphar advena* encodes two mitochondrial and one plastid phage-type RNA polymerases. BMC Evol Biol 10:379

Young ND, dePamphilis CW (2000) Purifying selection detected in the plastid gene *mat*K and flanking ribozyme regions within a group II intron of nonphotosynthetic plants. Mol Biol Evol 17:1933–1941

Zerges W (2000) Translation in chloroplasts. Biochimie 82:583–601

Zhang Q, Sodmergen (2010) Why does biparental plastid inheritance revive in angiosperms? J Plant Res 123:201–206

Zhong B, Yonezawa T, Zhong Y, Hasegawa M (2010) The position of gnetales among seed plants: Overcoming pitfalls of chloroplast phylogenomics. Mol Biol Evol 27:2855–2863

Zoschke R, Nakamura M, Liere K, Sugiura M, Börner T, Schmitz-Linneweber C (2010) An organellar maturase associates with multiple group II introns. Proc Natl Acad Sci USA 107:3245–3250

– CHAPTER III –

ASSEMBLY AND RECONSTRUCTION OF SPECIFIC GENOMIC SEGMENTS USING WHOLE GENOME SHOTGUN PYROSEQUENCING

ABSTRACT. Next generation sequencing technologies (NGS) have revolutionized genome research. The number of fully or partially sequenced animal and plant organellar and nuclear genomes has increased remarkably during the past few years. Nevertheless, there is no study available focusing on requirements for confident reconstruction of distinct genomic regions from un-enriched whole-genomic DNA. According to the sequenced sample or tissue, the abundance of a given genomic element may vary dramatically (e.g. mitochondrial, plastid DNA). Thus, its reconstruction from locus-unenriched DNA depends on several (usually) unknown parameters such as genome size, ratio and size of the desired region to total genomic DNA, and the complexity and abundance of small and large repetitive elements. In the present study, we investigate quality and accuracy aspects for the reconstruction of specific genomic segments from total genomic DNA extracts by evaluating the impact of different assembly strategies and 454-datasets of varying complexities. Using the plastid chromosome as an exemplary genomic locus and a resampling scheme for quality assessment for assemblies from simulated and empirical 454-datasets, we analyze and discuss the minimum requirements in terms of sequencing efforts for high-quality assembly and reconstruction of genomic regions. We are able to demonstrate that the successful reconstruction of a genomic region strongly depends on the size of the assembled read pool. Overall best results for *in silico* extraction and reconstruction of plastid regions are obtained at an average depth of 20-25X where contigs reach their optimal length, and number and length of gaps to a reference sequence drop to a local minimum. For those parameters, we show here that the assembly substantially benefits from read clustering. Optimal assembly read pools could be estimated a priori from the abundance of the particular region in the original dataset. This study uses these results to elaborate and provide a method for *a priori* assessment of the optimal read pool size for the assembly of genomic subsequences. Considering the rapid advances in sequencing technologies and the consistently increasing amounts of data that can be generated in one single run (or parts thereof), this scenario might affect future works on organellar genomics.

KEYWORDS. Locus-specific assembly, assembly statistics, assembly quality, sequencing effort, dataset-specific assembly parameters, *a priori* estimation

CONTENTS.

This chapter contains approx. 18,650 words, 14 figures and 14 tables plus 12 pages of supplemental material.

1. INTRODUCTION

The number of organellar genomes deposited in public databases has increased radically during the past few years. Mitochondrial genome sequencing has become a standard technique in animal research. Similarly, sequencing of chloroplast (plastid) genomes has increased dramatically in plant sciences (Gao et al. 2010). Several factors contribute to this trend. Consistently dropping prices and the growing amounts of data generated by non-Sanger sequencing techniques offer novel research approaches. High-throughput data from next-generation sequencing (NGS) facilitate screening of a great diversity of different organisms within short time and allow analyzing several aspects simultaneously. A single NGS sequencing run typically generates far more data than required for locus-specific reconstruction of a certain genome segment, which bears the potential of invaluable insight into other genome portions as a "side-effect". Moreover, commercial sequencing centers take over essential wet lab work, allowing researchers to concentrate on data analysis.

Regarding the field of plant sciences, plastid genomes are the best-characterized cellular genome. Land plant plastid chromosomes evolve in a highly conservative manner, rendering them an ideal tool for general use (as partial or complete genomes) in plant phylogenetics, comparative genomic studies, and as basis for biotechnical research (reviewed in Wicke et al. 2011). Sequencing of plastid chromosomes using traditional approaches with Sanger-sequencing as part of it is often laborious in terms of lab work, although generally (still) considered a safe way to obtain plastid genomes (Jansen et al. 2005; McNeal et al. 2006; Guisinger et al. 2010). Regarding performance and accuracy, *pyrosequencing* of plastid-enriched DNA extracts has been demonstrated to be comparable to that of traditional Sanger-sequencing and adequately applicable a more cost-efficient alternative for complete plastid genome reconstruction (Moore et al. 2006). Irrespective of the sequence technology used, chloroplast enrichment is often part of the DNA preparation procedure in order to minimize the number of contaminating sequences from other genome portions. On the one hand side, this step simplifies rapid genome assembly and dramatically reduces the number of necessary base pairs to be sequenced in total (Wakasugi et al. 1997; Moore et al. 2006; Raubeson et al. 2007; Cattolico et al. 2008; Oliver et al. 2010). On the other hand side, the enrichment procedures often require vast amounts of fresh and preferentially young plant material. In addition, downstream applications (NGS; BAC- or fosmid-libraries) usually demand an amount of pure and high molecular weight DNA, further increasing the need for large quantities of starting material. Different methods for separating chloroplasts (or plastid DNA) from the remaining cell compartments (or cellular genomes) have been established (Maniatis et al. 1982; Jansen et

al. 2005; Wolf et al. 2005; Truernit & Hibberd 2007; Sandberg et al. 2009), none of which is suitable for sparse or older and dried tissue. In addition, chloroplast enrichment procedures are virtually impossible for several plant groups such as parasitic plants or myco-heterotrophs (own works) due to supposedly altered plastid densities, or lowered abundance of plastids or plastid DNA (ptDNA) in the tissue, lacking fluorescing structures or extremely reduced plastomes. Besides, secondary metabolites or cellular structures (e.g. liverwort oil bodies) present in many plant lineages interfere with many molecular techniques (e.g. Forrest et al. 2011). Plastid enrichment may be circumvented via long (range) PCR. However, this approach often requires data from a close relative for primer design. Above that, it is highly sensitive to unexpected genome rearrangements and does not provide high coverage. Alternatively, large amounts of DNA may be obtained by whole genome amplification of DNA samples via rolling circle amplification/multiple displacement amplification (Dean et al. 2001; Hawkins et al. 2002). However, as with every intermediate amplification step, this holds the risk of introducing biases or artifacts due to e.g. heterogeneous template quality.

A promising alternative is organellar genome sequencing from *whole genomic DNA* (gDNA), i.e. un-enriched extracts. This became feasible only recently due to NGS generating the necessary amount of sequence data with reasonable effort. The success of this approach depends, however, on several (partly unknown) parameters such as overall genome size, stoichiometry and the amount of plastid DNA compared to total genomic DNA, number and similarity of plastid-like DNA nuclear and mitochondrial fractions, and complexity and abundance of repetitive elements (mini-, microsatellites). In contrast to mitochondrial DNA (mtDNA) that typically constitutes about one percent of gDNA extracts (Leaver & Gray 1982), the amount of ptDNA varies considerably. Normally, ptDNA represents around five to ten percent of gDNA extracts (Pascoe & Ingle 1978), but can be much less in old tissue or low-light (light-unexposed) tissue. In view of these percentages it is evident that the amount of base pairs for plastid or mitochondrial assemblies from total gDNA has to exceed the amount required from organelle enriched DNA considerably. The extent of this difference is, however, still controversial. Although the minimum amount of reads is largely determined by overall genome size, an *a priori* approximation of the adequate number of reads has remained inaccessible so far. Simply producing a number of reads that is way beyond any reasonable assumption of the required number ("brute force approach") would ensure sufficient coverage of organellar genomes and may have the benefit of providing valuable insights also into the nuclear genome. However, it is burdened with a greater risk of assembling chimeric plastomes due to the faulty inclusion of nuclear and mitochondrial plastid insertions. Additionally, simply increasing the overall number of reads implies increasing computational demands and bioinformatic expertise in the context of data handling and contig assembly.

Consequently, such an approach does not ensure the most efficient use of research funds. On the other hand, underestimation of the necessary amount of sequencing data entails the risk of large assembly gaps preventing the confident reconstruction of the entire desired region.

Different strategies are at the core of any *de novo* reconstruction of specific genomic loci (e.g. the plastid chromosome) from NGS-data: (i) Raw reads are assembled from the entire read pool; subsequently, the desired fragments are identified and extracted and, as far as necessary, subjected to further manual or automated post-assembly procedures (Fig. III-1); (ii), raw reads are sorted according to their similarity, which we will refer to as "CP-clustering" or "pre-clustering"); those sequence clusters are subsequently automatically assembled. Each of the two approaches has its advantages and disadvantages. Pre-clustering and assembly of single clusters may reduce the overall computational burden (in terms of RAM consumption and CPU-time) because only a small proportion of the data will be assembled at once. However, pre-clustering might not include all reads of the genomic region within one cluster. This would require searching other clusters or even modifying the clustering options. Assembling the entire read pool circumvents this, but at the expense of RAM and CPU-time. Pre-clustering may contribute to more stringent assemblies and reduce the overall risk of chimeric contigs, because read presorting putatively filters divergent copies localized in other genome segments.

Besides the assembly strategy, it has to be considered that sequencing errors from NGS-projects do not occur randomly. The shearing step of the shotgun-procedure as such is not a random process, but accumulates at certain portions in the genome, such as repetitive DNA and/or mononucleotide stretches where DNA breaks occur more frequently. Accumulation of such errors also affects the genome assembly and might therefore result in biased assemblies where errors are treated as truly existing genomic fragments (Moore et al. 2006; Meader et al. 2010). Assembly of individual reads can be performed in different ways and carried out by a number of various software packages (reviewed in Narzisi & Mishra 2011). The impact of a certain strategy becomes even more important with increasing amounts of sequence data and the entailed increasing need for automation (Schwartz & Waterman 2010). Thus, we may assume that excessive sequencing is only beneficial up to a certain point, and thereafter likely to cause problems due to e.g. assembly errors from biases during DNA preparation processes. Moreover, datasets providing "reasonable coverage" for the reconstruction of genomic segments produce most confident assemblies implying that both underrepresentation and overrepresentation is disadvantageous (Chevreux et al. 1999; Chevreux 2011). In particular, plant genomes with extreme genome sizes and large proportions of repetitive elements require assembly optimization.

This study specifically explores high-throughput whole-genome shotgun

pyrosequencing for the reconstruction of specific genomic loci. Using the plastid chromosome as an exemplary region, we aim to evaluate the pros and contras of *de novo* reconstruction of genomic loci from total genomic and non-enriched DNA from pyrosequencing data. We will examine two major assembly strategies and evaluate the extent to which high-quality locus-specific assemblies depend on read number and/or species-specific genomic peculiarities. This will allow us to illuminate adequate sequencing efforts or assembly requirements for the *de novo* reconstruction of genomic regions using whole-genome shotgun sequencing approaches. We hypothesize that there is a trade-off between plastome assembly accuracy and -quality and at the same time increasing assembly problems due to larger sequence pools. In order to investigate this, we simulated pyrosequencing of thale cress (*Arabidopsis thaliana*) given a range of different ratios of plastid DNA and different read pool sizes. In repeated and independent runs, we evaluate the subsequent *in-silico* extraction. We compare the results from simulated *Arabidopsis*-pyrosequencing runs to four different plant species that differ in (at least) genome size and the complexity of the plastome. Using this approach, we assess here the minimum sequencing effort that is required to obtain an adequate and homogenous coverage for *de novo* reconstruction of the plastid genome. We will test to what extent read number (data deficit and excess) influences the overall performance of plastome reconstruction as reflected in assembly gaps, putative chimeras and inhomogeneous coverage. Furthermore, we will analyze and discuss the impact on performance in the different *de novo* assembly strategies for pyrosequencing data and assembly quality per cost and effort.

2. MATERIAL AND METHODS

2.1. Taxon sampling for empirical data

Three different parasitic but photosynthetically active species from the mostly parasitic broomrape family (Orobanchaceae) have been chosen according to their plastome features. *Lindenbergia*, the sole non-parasitic genus of this family, exhibits a conserved angiosperm plastome structure that does not differ significantly in gene content and synteny from other angiosperms. The parasitic *Schwalbea* and *Striga* exhibit some gene losses (4 and 5 ndh-genes, respectively) as well as an elevated ratio of short dispersed repeats compared to other angiosperm plastid chromosomes (S. Wicke et al., unpublished data). While the structure of the *Schwalbea* plastome is widely co-linear to that of *Lindenbergia*, gene synteny exhibits several changes in *Striga* including aberrant structural features in the plastid large inverted repeat region (Downie & Palmer 1992; S. Wicke et al.,

unpublished data). In order to test the extent to which genomic peculiarities influence the overall assembly process, we additionally included the non-photosynthetic parasitic species *Phelipanche ramosa* that possesses a highly derived plastome structure with many genes lost, long dispersed repeats and an extraordinarily high amount of plastid-like sequences localized in the nuclear/mitochondrial genome (NUPTs).

2.2. Shotgun-sequencing of plastid genomes from total genomic DNA

Total genomic DNA was isolated from one fully autotrophic plant (*Lindenbergia philippensis*); two photosynthetic hemiparasites (*Striga hermonthica; Schwalbea americana*) as well as a non-photosynthetic parasitic plant using a modified CTAB-based protocol (McNeal et al., 2006). DNA was additionally spin-column purified employing the NucleoSpin® Plant II kit (Macherey-Nagel). For *Lindenbergia*, an additional plastid-enrichment step (following the protocol described by Jansen et al., 2005) was carried out. All DNA extracts have been digested with RNase A. Between 1 and 5 μg of purified DNA was employed to generate DNA-shotgun libraries of ca. 600-750 bp that were then used for sequencing on a Roche 454-Titanium platform at the ZMF Center for Medical Research, Core Facility Molecular Biology of the Medical University of Graz/Austria. The plastid-enriched *Lindenbergia* was sequenced at the core facility of the University of Connecticut/USA. 454-sequence data (i.e. the complete read pool) was assembled *de novo* using MIRA v3 (Chevreux et al., 1999) under the "accurate" 454-assembly mode with few modifications (minimal overlap = 70 bp, minimum overlap identity = 95; for *Schwalbea*: nasty-repeat masking was required hiding *k-mers* repeated 15 or more times in the read pool). Plastid-like regions were extracted with the local BLAST suite (available from the NCBI ftp server) using custom plastid sequence databases, retrieving only BLASTx-hits with a minimal E-value of 10^{-30} (10^{-10} for *Phelipanche*), and a minimal alignment length of 100 bp. The plastid genomes were manually reconstructed *gene-by-gene*. That is, contigs matching a specific gene were post-assembled employing the Lasergene® SeqMan II (DNA Star) assembler mode at high stringency (min. identity = 97%, min. overlap = 80 bp; max. 10 gaps per contig; max. 5 gaps per sequence, gap length penalty = 0.95). Construction of the plastid supercontig by overlapping gene-contigs was assisted by preliminary annotation using the DOGMA-web server (Wyman et al., 2004) setting the BLASTx identity cutoff for protein coding genes to 40%. Identity cutoff for structural RNAs was set to 90 % and an E-value of 5. Contig joints including regions of uncertainty (e.g. microsatellites, homopolymeric stretches >15bp) as well as junctions of plastome single-copy regions to the large inverted repeat regions were PCR-amplified and verified by Sanger sequencing. Plastid genomes were annotated with the assistance of DOGMA, followed by manual refinement. Physical maps, comparative plastome analyse as well as detailed information regarding finishing are provided in chapters IV and V.

The obtained datasets were of different size and slightly different in average length of sequences. The *Lindenbergia* dataset contained 348.563 reads with an average of 251 bp (totaling 87 Mbp). The *Schwalbea* dataset contained 509.838 reads of on average 257 bp length (131 Mbp); 674.688 reads for *Striga* with on average 316 bp in length (213 Mbp). Two halfs of a picotiter plate were sequenced for *Phelipanche* yielding a total of 1.516.862 reads with an average length of 344 bp (261.5 Mbp). A separate dataset for *Lindenbergia* was sequenced from a plastid-enriched and whole-genome amplified DNA extract containing a total of 10.109 reads with an average length of 394 bp (3.9 Mbp).

2.3. Simulation of Arabidopsis whole-genome shotgun datasets.

454-sequence data was generated using MetaSim v. 0.9.5 (Richter et al. 2008), which allows manipulating ratios of genomic sequences to one another by adjusting copy numbers of individual regions. Whole genome sequence data (i.e. nuclear chromosomal, mitochondrial and plastid sequences) from *Arabidopsis thaliana* were retrieved from the NCBI BioProject Database (Accession: PRJNA116, ID: 116). Following Doležel et al. (Dolezel et al. 2003), we estimated the amount of totally required base pairs to be 978×10^6 Mb assuming a total of 1μg DNA as requirement for whole genome shotgun pyrosequencing (i.e. 454-seqeuncing). According to its total genome size of ca. 188.8 Mb (chromosome sizes: 1 – 30 Mb, 2 – 19.7 Mb, 3 – 23.5, 4 – 18.6 Mb, 5 – 27 Mb, plastid – 0.155 Mb, mitochondrial – 0.367 Mb), we calculated the number of required copies for 1μg DNA of nine different ptDNA ratios (1-9%) and a constant proportion of mitochondrial DNA (0.5% mtDNA). Thus, for simulation of 1% ptDNA in a 978×10^6 Mb-dataset, this corresponds to a fraction of ca. 9.78 Mb plastid DNA (i.e. 63,096,774 plastid genome copies), 4.89 Mb mitochondrial DNA (13,216,216 copies), and 963.3 Mb nuclear DNA (8,108,838 copies). MetaSim generated a 454-shotgun dataset of 1.1 million reads, which matches the average dataset size obtained from empirical 454 runs at the ZMF Core Facility Molecular Biology of the Medical University, Graz/Austria (see section 2.2., last paragraph). *MetaSim* was run using the 454-error model assuming 200 flow cycles; lognormal distribution of mean read length and standard deviation as well as the proportionality constant was used at default settings of MetaSim v. 0.9.5. Clone parameters were set to a normal mean distribution of 350.0 for the first parameter, and 250.0 for the second parameter. Under these settings, nine different 454-datasets were generated with read lengths of an average 360 bp. Due to long computational time required for assemblies without previous read clustering, we omitted direct assemblies for this set of analyses.

2.4. Analysis of 454 data

2.4.1. Assembly strategy for plastome reconstruction from pyrosequencing data.

In order to test which assembly strategy is more suitable for locus-specific assembly, we assessed the assembly quality of clustering and non-clustering approaches for two different assemblers (CAP3: Huang & Madan 1999; MIRA v3: Chevreux et al. 1999). CAP3 uses a time-consuming greedy assembly algorithm, whereas MIRA is based upon an overlap-layout approach including an iterative repeat handling.

Test runs revealed that error rates and sample variances per read pool size were not significantly different between 50 and 100 replicates per read number (data not shown). In view of this and the extraordinary computational time required in particular for larger sequence pools and in the absence of initial read clustering, only 50 replicates were run for each read pool size.

2.4.2. Automated testing and statistical evaluation of assembly confidence.

A resampling scheme was used for assessing the minimum requirements for high quality plastome reconstruction from pyrosequencing data (Fig III-1). Using custom Perl scripts, *n* reads were randomly selected from the total read pool, using a starting number of n=100,000 for simulated data. The number was stepwise increased after 50 replicates by another 100,000 reads up to one million reads. We analyzed assembly aspects in more detail using the empirical datasets, where we sampled data every 50,000 reads from 50,000 to 500,000. For a given n, the remainder of the pipeline executed the two different assembly approaches, (i) an initial read-clustering step (where reads were sorted according to the similarity of reads to a custom plastid sequence archive) followed by CAP3 and MIRA assembly, (ii) CAP3/MIRA assembly without the initial clustering. Either of those was subjected to an automated quality assessment procedure as follows. First, general assembly statistics were inferred that included the computational effort (CPU time), the total amount of contigs built, number of debris/singlets as well as the average length and quality of contigs. Second, locus-specific statistics were calculated after aligning all created contigs against the species-specific verified plastome reference sequence. In order to separate ptDNA of a true plastid origin from the remainder, we automatically aligned the total set of contigs to a verified plastid chromosome sequence using highly stringent MEGABLAST. Using the obtained Evalue as a measure for stringency, we defined results retrieving values smaller than $1e^{-100}$ threshold as a highly significant hit. Low complexity masking was employed to allow for small stretches of low-complexity in the query/subject region. In order to account for potential duplicate hits due to the presence of a large repeat region, one of these segments was removed from the reference sequence. After the alignment step, the truly observed ptDNA was inferred from the amount of reads

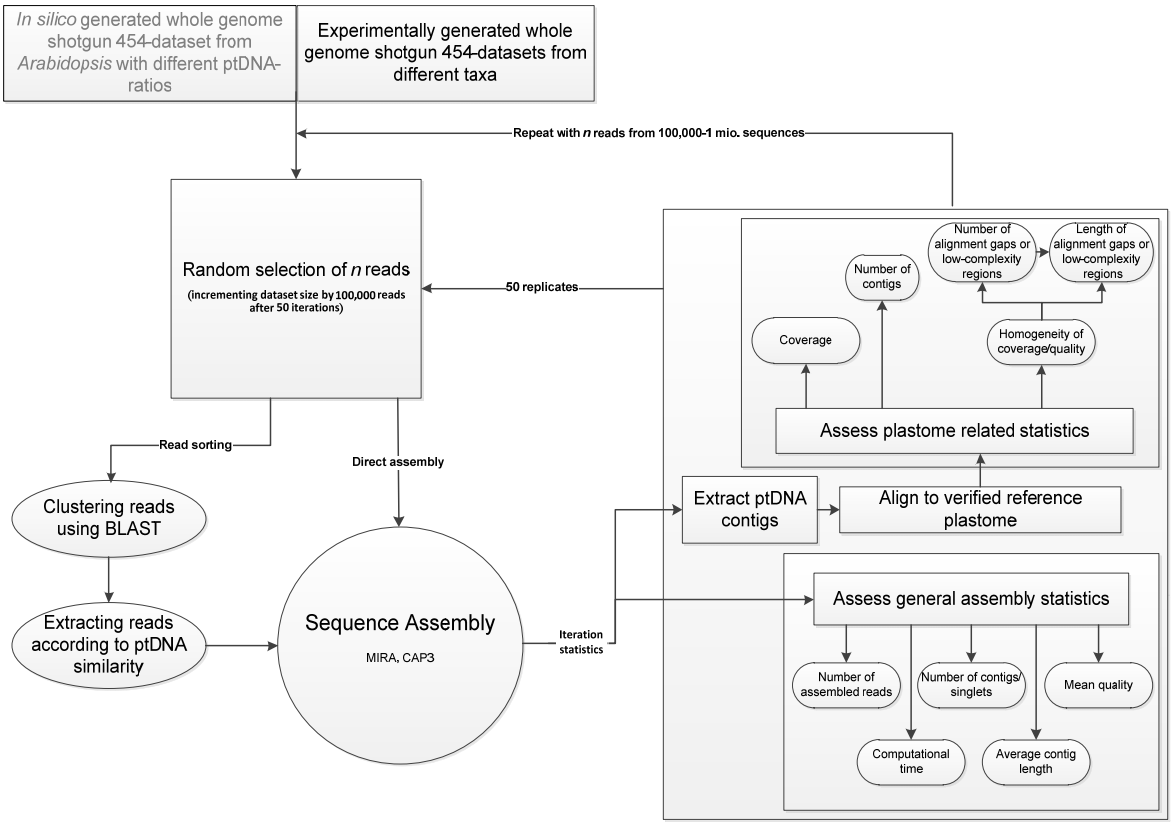


Fig. III-1 **Flowchart illustrating the resampling scheme for assembly quality assessment.** Reads are extracted from different 454 datasets originally generated from four different plant taxa or nine different in-silico datasets of Arabidopsis, respectively. A single iteration of the assembly process consists of randomly selecting a number of reads. Either these reads are directly handed over to the automated assembler, or only plastid-like reads are assembled that have previously been sorted (clustered) using BLAST-algorithms. The general assembly quality and performance is subsequently assessed as well as plastome specific statistics are automatically inferred. The random selection step is repeated 50 times for each n.

assembled into MEGABLAST-verified plastid contigs, and read depth (coverage) of the entire region and a statistic for the homogeneity of coverage were computed. The latter was inferred from the number of alignment breaks or alignment gaps. Furthermore, the number and length of potential gaps and/or low-complexity regions were calculated from alignments gaps to the reference genome. CAP3 and MIRA-assemblies were computed on the ZIVSMP high-performance computing cluster at the University of Münster/Germany. In order to assess computational time on desktop computers, we measured computational time independently for two empirical datasets on a quad core desktop PC (i7 CPU 860 @2.80 GHz, 8GB RAM). Time used for the clustering (if applied), assembly and alignment to reference was averaged over ten independent runs for read pool sizes of 100.000, and 500.000 sequences for the *Striga* and *Phelipanche* dataset.

Raw results obtained for all iterations per read pool were further analyzed using custom R scripts. Beyond descriptive statistics for every pool size per species, we tested the measured quality parameters regarding the normality of distribution (with QQ-plots,

skewness and kurtosis and the Shapiro-Wilk test; results are available upon request from the first author), regression analyses, as well as non-parametric tests to analyze differences across the different read pools per taxon. Tests for normality revealed that only a fraction of the obtained data per read pool and taxon could be accepted as normally distributed. Thus, nonparametric Mann-Whitney–U tests was conservatively given preference over parametric tests since test conditions for those were not or not known to be fulfilled. In order to assess the increment/slope of coverage per dataset and assembly method, generalized linear models were fitted across the coverage data per taxon and read pool. Correlation of data was re-confirmed by the Spearman rank and Kendall correlation test to statistically assure the data correlation. Nonlinear regression estimates for the relation of coverage slope increase and the ptDNA/gDNA ratio (thereafter referred to only as ptDNA-ratio) was carried out in the statistics software program *STATISTICA* (StatSoft, Inc.). We fitted an exponential decline function of type $y(x) \cong a \times e^{(-bx)}$,where x represents the ptDNA ratio, and y stands for the slope to be estimated *a priori*, through all *Arabidopsis* data points, where we assumed coverage slope to depend on ptDNA-ratio (i.e. *x*). Observed vs. predicted-value plots, the correlation coefficient R as well as an F-test (data did not depart from normality) were computed to test the goodness of fit.

All scripts and *Arabidopsis* raw datasets will be provided by the first author upon request. Empirical dataset will be deposited to the NCBI Sequence Read Archive.

3. RESULTS AND DISCUSSION

3.1. General assembly statistics

3.1.1. Preclustering saves immense computational effort for locus specific assembly of large sequence pools.

Computational saving is immense when using clustering methods, in particular when working with high read numbers. The number of reads identified with similarity to ptDNA ranges around similar percentages, regardless of ptDNA amount (Table III-A). Filtering becomes prominently effective at around 500,000 reads, which corresponds to sequencing about one half of a picotiter plate. At these and larger initial read pool sizes, approximately 80% of reads were excluded from the assembly (Fig. III-2). The ptDNA-like clusters built from the remaining 20 % of reads can thus be assembled on a desktop computer with 6-8 GB of RAM for both CAP3 and MIRA. Direct assemblies (i.e. without previous sorting) of read pools >500.000 with an average length of >350bp can already be difficult to achieve, and may require virtual memory on an 8 GB desktop computer. Unlike MIRA, CAP3 cannot parallelize parts of the assembly process. On average MIRA analyses

Table III-A **Computational savings using a read clustering for the reconstruction of plastid genome sequences.** The average proportion of assembled sequences after a read sorting is summarized for *simulated* datasets as well as all empirical datasets. Time required for the assembly process on a quadcore PC is presented below for two dataset and two read pool sizes, sorted according to the assembler. Standard deviation is provided in brackets. [Abbr.: k - ×1000, M – ×10^6, NA – not sampled].

	Read pool size									
Dataset	**100k**	**200k**	**300k**	**400k**	**500k**	**600k**	**700k**	**800k**	**900k**	**1 M**
Proportion of reads assembled during assemblies with previous read sorting (in %)										
Arath – 1%	51.81 (0.16)	31.58 (0.08)	26.84 (3.13)	21.65 (3.62)	23.18 (0.04)	20.73 (0.03)	19.27 (0.02)	18.33 (0.02)	17.52 (0.02)	16.70 (0.01)
Arath – 2%	47.92 (0.12)	31.36 (0.07)	23.07 (0.04)	18.48 (0.03)	23.49 (0.04)	20.98 (0.03)	19.45 (0.03)	18.58 (0.02)	17.68 (0.02)	17.03 (0.01)
Arath – 3%	51.33 (0.13)	36.01 (0.07)	29.97 (0.06)	26.12 (0.05)	23.70 (0.04)	21.19 (0.04)	19.68 (0.02)	18.77 (0.02)	17.90 (0.02)	17.25 (0.01)
Arath – 4%	51.56 (0.12)	36.26 (0.08)	30.23 (0.06)	26.35 (0.05)	23.89 (0.04)	21.35 (0.04)	19.86 (0.03)	18.97 (0.02)	18.12 (0.01)	17.49 (0.01)
Arath – 5%	51.81 (0.13)	36.60 (0.08)	30.55 (0.07)	26.60 (0.05)	24.15 (0.03)	21.61 (0.03)	20.11 (0.03)	19.23 (0.02)	18.39 (0.01)	17.76 (0.01)
Arath – 6%	51.99 (0.15)	36.75 (0.09)	30.68 (0.06)	26.78 (0.05)	24.31 (0.03)	21.77 (0.03)	20.27 (0.03)	19.39 (0.02)	18.57 (0.01)	17.95 (0.01)
Arath – 7%	52.26 (0.13)	37.02 (0.09)	30.93 (0.06)	26.99 (0.04)	24.56 (0.04)	22.00 (0.03)	20.49 (0.02)	19.61 (0.02)	18.81 (0.02)	18.20 (0.01)
Arath – 8%	52.48 (0.12)	37.23 (0.08)	31.13 (0.06)	27.23 (0.05)	24.72 (0.04)	22.20 (0.03)	20.69 (0.02)	19.82 (0.02)	19.04 (0.01)	18.41 (0.01)
Arath – 9%	52.74 (0.12)	37.51 (0.08	31.41 (0.06)	27.47 (0.05)	24.99 (0.03)	22.44 (0.04)	20.97 (0.02)	20.13 (0.02)	19.34 (0.01)	18.71 (0.01)
Lindenbergia	45.80 (0.11)	34.89 (0.01)	23.02 (0.01)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	44.02 (0.02)	29.46 (<0.01)	22.27 (<0.01)	17.53 (<0.01)	14.86 (<0.01)	NA	NA	NA	NA	NA
Striga	37.47 (<0.01)	25.58 (<0.01)	18.67 (<0.01)	15.45 (<0.01)	13.25 (<0.01)	NA	NA	NA	NA	NA
Phelipanche	44.84 (<0.01)	28.79 (<0.01)	21.10 (<0.01)	16.88 (<0.01)	14.09 (<0.01)	NA	NA	NA	NA	NA
Time to Assemble incl. Alignment (in hh:mm)										
CAP3 with previous read sorting										
Striga	00:18 (<00:00)				03:22 (00:05)					
Phelipanche	00:27 (00:08)				05:22 (00:31)					
CAP3 without previous read sorting										
Striga	00:46 (00:01)				07:20 (00:01)					
Phelipanche	01:06 (00:45)				05:26 (00:03)					
MIRA with previous read sorting										
Striga	00:04 (<00:00)				00:05 (00:01)					
Phelipanche	00:11 (<00:00)				00:05 (00:03)					
MIRA without previous read sorting										
Striga	00:06 (00:01)				01:28 (00:11)					
Phelipanche	00:15 (<00:00)				01:13 (00:02)					

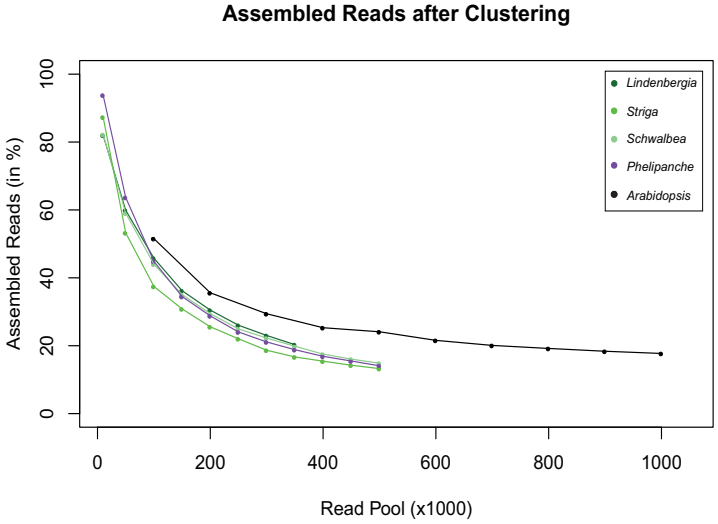


Fig. III-2 **The proportion of reads retained after read clustering of simulated and empirical datasets.** 100% corresponds to the assembly of the unfiltered initial read pool.

are therefore much faster. The difference of required computational time becomes most evident in assemblies without previous read clustering. On an quadcore (i7) PC, MIRA finishes the assembly of half a million unsorted reads in ca. 73 min (*Phelipanche*) and 88 min. (*Striga*), respectively, whereas CAP3 takes 326 min (*Phelipanche*) and 440 min for the *Striga* dataset (Table III-A). Regarding computational time, read sorting before the assembly is highly efficient in combination with both the CAP3 and the MIRA assembler.

3.1.2. Number of contigs and unused reads varies greatly between different assemblers.

The number of contigs increases proportionally with the amount of plastid DNA in the gDNA extract. In assemblies from the largest read pools (1 mio. reads) employed in a presorting strategy, CAP3 generates up to 11.000 contigs (Table III-B). Although the overall number of contigs increases depending on the available sequence pool, the correlation of read pool size and contig number is not linear, but for CAP3 bears similarity to a saturation curve. The MIRA assembler produces much fewer contigs - only up to approx. 6000 contigs from pools of one million *Arabidopsis* reads. Differences in the quantity of contigs are marginal between datasets differing in the amount of the desired genomic region (ptDNA content). This observation is especially evident in assemblies from small sequence pools. The amount of contigs deviates slightly more when assembled from a large sequence pool. MIRA creates the largest number of contigs from datasets with a ptDNA content of 2-6%, whereas the number of contigs CAP3 remain nearly equal. Similar to CAP3, the number of MIRA-contig increases not linearly with read pool size. However, the incline is dissimilar from a saturation curve, but tends to increase slightly exponentially (Fig. III-3). The quantity of created contigs from empirical datasets deviates from that of the simulated 454 datasets in that both assemblers reach a saturation point

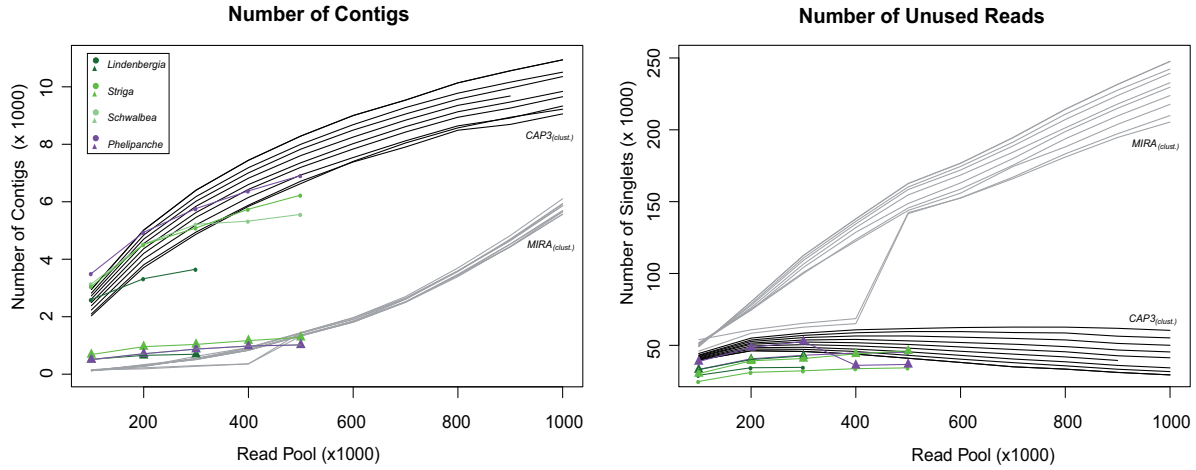


Fig. III-3 **Number of contigs and unassembled reads using a read-presorting assembly strategy.** The left hand side illustrates the number of contigs created in assemblies from different read pools of simulated and experimentally generated 454 data. The right hand graphic provides an overview of the quantity of singlets after assemblies of differently sized read pools. Black lines and colored dots represent data obtained from the CAP3 assembler; gray lines and colored triangles illustrate MIRA data.

Table III-B **Number of contigs.** The number of contiguous sequences from CAP3 and MIRA-assemblies using read-presorting datasets has been averaged over 50 samples per read pool size for *in-silico* 454-datasets as well as all experimental generated datasets. The standard deviation is provided in brackets. [Abbr.: k - ×1000, M – ×10⁶, NA – not sampled; # – not available due to technical problems].

	Read pool size									
Dataset	**100k**	**200k**	**300k**	**400k**	**500k**	**600k**	**700k**	**800k**	**900k**	**1 M**
CAP3 with previous read sorting										
Arath – 1%	2030 (40)	3710 (51)	4870 (52)	5840 (57)	6633 (60)	7405 (48)	8046 (62)	8571 (60)	8940 (46)	9221 (47)
Arath – 2%	2095 (40)	3807 (58)	4945 (63)	5880 (56)	6716 (70)	7384 (63)	7905 (56)	8489 (62)	8692 (53)	9057 (49)
Arath – 3%	2232 (44)	4008 (46)	5196 (71)	6150 (55)	6931 (58)	7539 (63)	8120 (56)	8641 (53)	8913 (47)	9331 (41)
Arath – 4%	2380 (41)	4201 (50)	5458 (56)	6430 (51)	7198 (58)	7835 (64)	8425 (60)	8940 (60)	9260 (51)	9656 (45)
Arath – 5%	2491 (40)	4366 (52)	5628 (66)	6599 (58)	7356 (68)	8022 (64)	8614 (65)	9132 (55)	9471 (39)	9840 (39)
Arath – 6%	2613 (47)	4545 (51)	5845 (51)	6820 (54)	7608 (52)	8251 (53)	8840 (51)	9354 (53)	9680 (44)	NA#
Arath – 7%	2722 (44)	4699 (50)	6020 (49)	7006 (70)	7830 (50)	8493 (47)	9053 (57)	9569 (53)	9956 (41)	10356 (42)
Arath – 8%	2819 (39)	4841 (55)	6181 (65)	7186 (53)	8000 (56)	8683 (55)	9278 (53)	9779 (53)	10162 (48)	10509 (49)
Arath – 9%	2949 (43)	5013 (51)	6403 (53)	7443 (49)	8286 (59)	9001 (46)	9538 (554)	10137 (54)	10558 (54)	10939 (39)
Lindenbergia	2593 (41)	3324 (46)	3654 (31)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	3142 (47)	4526 (47)	5216 (43)	5326 (40)	5561 (23)	NA	NA	NA	NA	NA
Striga	3055 (39)	4510 (57)	5090 (49)	5741 (52)	6231 (38)	NA	NA	NA	NA	NA
Phelipanche	3501 (44)	4939 (62)	5763 (54)	6377 (62)	6911 (53)	NA	NA	NA	NA	NA
MIRA with previous read sorting										
Arath – 1%	125 (8)	223 (11)	298 (10)	359 (16)	1329 (26)	1820 (31)	2510 (37)	3444 (42)	4522 (44)	5657 (36)
Arath – 2%	150 (8)	186 (11)	273 (20)	357 (17)	1447 (31)	1968 (28)	2701 (37)	3696 (36)	4825 (40)	6105 (39)
Arath – 3%	154 (8)	284 (12)	539 (19)	899 (25)	1439 (28)	1903 (36)	2613 (33)	3573 (34)	4702 (46)	5938 (35)
Arath – 4%	137 (9)	267 (13)	544 (14)	922 (23)	1418 (32)	1921 (29)	2640 (38)	3596 (51)	4678 (41)	5925 (39)
Arath – 5%	123 (9)	276 (29)	621 (28)	929 (30)	1443 (33)	1962 (31)	2650 (34)	3596 (40)	4670 (55)	5862 (47)
Arath – 6%	116 (9)	334 (15)	537 (19)	857 (28)	1364 (32)	1842 (31)	2606 (104)	3586 (49)	4642 (40)	5867 (42)
Arath – 7%	122 (9)	285 (14)	550 (20)	841 (26)	1337 (26)	1814 (36)	2515 (42)	3477 (34)	4532 (46)	5696 (37)
Arath – 8%	127 (8)	293 (13)	526 (20)	829 (25)	1332 (30)	1809 (37)	2499 (36)	3395 (46)	4445 (51)	5652 (36)
Arath – 9%	129 (8)	311 (13)	507 (18)	831 (21)	1344 (32)	1845 (36)	2517 (39)	3443 (41)	4430 (41)	5567 (41)
Lindenbergia	516 (15)	655 (17)	695 (12)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	508 (17)	712 (14)	873 (17)	979 (16)	1020 (12)	NA	NA	NA	NA	NA
Striga	684 (17)	958 (21)	1036 (22)	1169 (22)	1282 (19)	NA	NA	NA	NA	NA
Phelipanche	508 (17)	712 (14)	873 (17)	979 (16)	1020 (12)	NA	NA	NA	NA	NA

Table III-C **Number of unused reads.** The number of contiguous sequences from CAP3 and MIRA-assemblies using a read-presorting assembly strategy has been averaged over 50 samples per read pool size for in-silico 454-datasets as well as all experimental generated datasets. The standard deviation is provided in brackets. [Abbr.: k - ×1000, M – ×106, NA – not sampled; # – not available due to technical problems].

	Read pool size									
Dataset	**100k**	**200k**	**300k**	**400k**	**500k**	**600k**	**700k**	**800k**	**900k**	**1 M**
CAP3 with previous read sorting										
Arath – 1%	44,314 (155)	55,040 (177)	58,501 (151)	60,796 (150)	61,658 (184)	62,385 (133)	62,700 (147)	62,669 (119)	61,846 (126)	60,419 (100)
Arath – 2%	43,239 (155)	53,717 (138)	56,929 (151)	58,764 (136)	59,596 (148)	59,567 (149)	58,916 (179)	58,620 (143)	56,369 (148)	55,184 (119)
Arath – 3%	42,797 (121)	52,711 (179)	55,393 (188)	56,517 (182)	56,519 (180)	55,704 (169)	54,817 (141)	53,773 (156)	51,335 (121)	50,041 (81)
Arath – 4%	42,239 (127)	51,570 (157)	53,733 (148)	54,145 (162)	53,286 (165)	51,958 (180)	50,655 (161)	49,289 (122)	46,752 (122)	45,412 (103)
Arath – 5%	41,800 (140)	50,747 (156)	52,298 (209)	51,818 (194)	50,628 (143)	48,880 (151)	47,145 (231)	45,522 (137)	42,740 (105)	41,330 (100)
Arath – 6%	41,223 (138)	49,527 (174)	50,498 (175)	49,648 (176)	48,083 (156)	45,878 (145)	43,914 (170)	41,938 (136)	39,490 (110)	NA#
Arath – 7%	40,781 (134)	48,354 (176)	48,645 (160)	47,357 (155)	45,503 (178)	43,041 (142)	40,489 (151)	38,557 (156)	36,032 (116)	34,324 (90)
Arath – 8%	40,323 (126)	47,378 (162)	47,217 (159)	45,715 (160)	43,055 (173)	40,370 (160)	37,966 (150)	35,751 (125)	33,371 (108)	31,725 (88)
Arath – 9%	39,788 (144)	46,228 (175)	45,742 (156)	43,866 (196)	41,013 (174)	38,101 (148)	35,075 (2382)	33,452 (119)	31,151 (99)	29,431 (97)
Lindenbergia	29,186 (130)	34,420 (124)	34,680 (79)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	30,277 (144)	36,065 (137)	36,958 (125)	35,450 (110)	34,634 (45)	NA	NA	NA	NA	NA
Striga	24,875 (127)	31,195 (153)	32,273 (120)	33,775 (123)	34,328 (107)	NA	NA	NA	NA	NA
Phelipanche	33,225 (140)	40,635 (145)	43,088 (148)	44,661 (177)	45,179 (130)	NA	NA	NA	NA	NA
MIRA with previous read sorting										
Arath – 1%	53,910 (174)	60,813 (169)	65,345 (167)	68,615 (173)	142,355 (315)	152,483 (272)	165,983 (279)	181,032 (253)	194,684 (282)	205,401 (229)
Arath – 2%	45,970 (119)	58,421 (154)	62,645 (141)	65,093 (171)	141,688 (315)	152,635 (285)	167,042 (391)	183,280 (311)	197,261 (245)	209,907 (246)
Arath – 3%	51,800 (165)	75,206 (151)	100,834 (245)	122,523 (307)	143,900 (304)	155,899 (380)	174,441 (2,655)	188,103 (313)	203,835 (300)	217,713 (210)
Arath – 4%	51,313 (151)	74,528 (214)	99,999 (185)	123,727 (256)	145,604 (288)	158,699 (414)	175,322 (315)	193,751 (329)	209,673 (271)	223,831 (225)
Arath – 5%	50,890 (143)	74,683 (942)	104,227 (341)	128,698 (1,766)	148,685 (976)	163,039 (328)	179,997 (284)	198,757 (317)	214,994 (332)	229,654 (280)
Arath – 6%	50,375 (168)	76,208 (228)	106,560 (245)	131,977 (294)	154,706 (315)	168,493 (263)	184,407 (683)	201,354 (749)	217,959 (222)	232,761 (244)
Arath – 7%	49,948 (147)	77,727 (195)	109,086 (264)	134,503 (266)	158,420 (247)	171,996 (337)	188,752 (460)	207,140 (260)	223,855 (302)	239,279 (272)
Arath – 8%	49,484 (123)	78,765 (264)	110,747 (283)	136,255 (264)	160,211 (286)	174,532 (314)	191,389 (327)	210,831 (323)	227,599 (361)	242,290 (224)
Arath – 9%	49,085 (133)	80,323 (228)	112,545 (238)	138,031 (238)	162,618 (248)	176,745 (382)	194,545 (345)	214,256 (313)	231,630 (332)	247,619 (240)
Lindenbergia	33,105 (161)	40,090 (129)	42,658 (122)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	NA#	NA#	NA#	NA#	NA#	NA	NA	NA	NA	NA
Striga	30,379 (149)	39,221 (180)	40,830 (177)	44,085 (185)	46,172 (188)	NA	NA	NA	NA	NA
Phelipanche	38,724 (161)	48,551 (178)	52,699 (222)	36,067 (165)	36,731 (200)	NA	NA	NA	NA	NA

with respect to the overall contig number (Fig. III-3). Besides, both assemblers create slightly more contigs per read pool for empirical data than from the simulated 454-datasets. The variation in overall contig number between the different assemblers directly relates to the number of unassembled reads in both simulated and empirical datasets (Table III-C, Fig. III-3). The steadily growing number of CAP3 contigs results in a continuously decreasing proportion of unused reads in assemblies employing previous preclustering. Similarly, the proportion of singlet reads follows a saturation curve that reflects the exponential incline of MIRA contigs at large sequence pools. Compared to *Arabidopsis* data, the mean number of unused reads is slightly lower in the empirical datasets due to the on average higher number of generated contigs in both MIRA and CAP3 assemblies.

3.1.3. Average contig length reaches a taxon- and strategy-specific local maximum in both simulated and empirical data.

CAP3 contigs generated from simulated data with and without preclustering are on average between 600 and 900 bp in length. MIRA contigs are generally longer ranging from 800-2000 bp. Using the CAP3 assembler, differences in contig size are minimal between the different plastid ratios of *Arabidopsis* data (Table III-D, Fig III-4). In contrast, MIRA contig sizes show more distinctive differences in response to different ptDNA ratios. Above that, the average length of contigs tends to decrease in assemblies from large read pools. This is particularly evident in datasets with a low ptDNA amount (1-2%). Contig size increases until a specific read pool size after which contigs length decreases again. Contigs from plastid-unenriched Orobanchaceae samples are on average shorter than those of simulated 454 runs data. This may in part be due to a more complex genomic structure (e.g., repetitive DNA) in Orobanchaceae compared to *Arabidopsis*. In contrast, contigs from plastid-enriched *Lindenbergia* data range around 1,500 bp in CAP3-

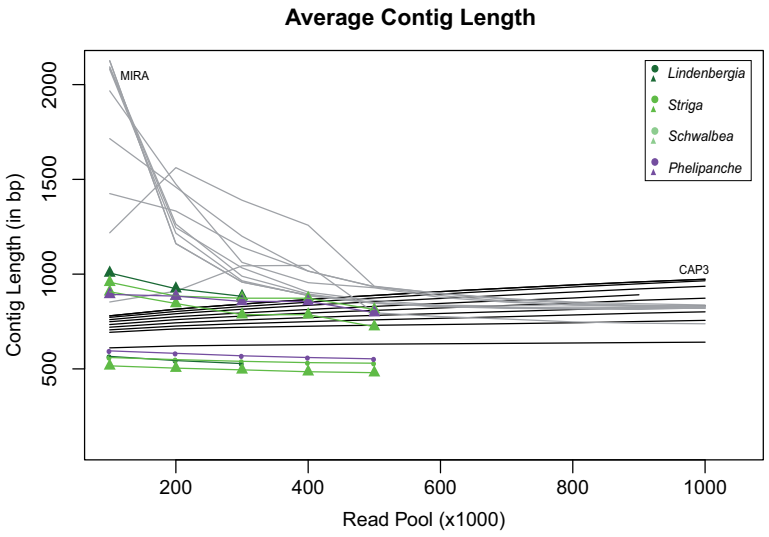


Fig. III-4 **Average contig length.** Average contig length in assemblies from different read pool sizes is illustrated for CAP3 using previous read sorting for simulated (black lines) and experimental datasets (colored dots). Dark gray lines and colored triangles represent MIRA-data.

assemblies. MIRA even generated contigs of an average length up to 7,000 bp (Supplemental Material: Table SIII-A). Similar to the *Arabidopis*-MIRA contigs, we observe shorter contigs when they are assembled from larger sequence pools in all photosynthetic and nonphotosynthetic taxa (Table III-D, Fig. III-4). Even more, it seems as if contig length reaches a local maximum. Figure III-5 provides a more detailed overview of contig differences in experimentally generated datasets and compares the trends in contig lengths between direct and presorting assemblies. The optimum seems to be dataset specific. Longest contigs are generated around 50,000 reads in the plastid-unenriched *Lindenbergia* dataset with both CAP3 and MIRA assembler; the other datasets require larger read pools. A different pattern, however, can be observed in the *Striga* dataset, where contig length is rather stable in CAP3 assemblies tending towards a slight upward trend (Fig. III-5). A local minimum can be observed in the *Schwalbea* dataset assembled with MIRA and without read presorting. Passing a minimal contig length occurring after 150,000 reads,

Table III-D **Average Contig Length.** Contig sizes obtained from CAP3 and MIRA assemblies using previous read sorting are averaged over 50 samples per read pool for simulated and empirical datasets. Standard deviation is given in brackets. [Abbr.: k - ×1000, M – ×10^6, NA – not sampled; # – not available due to technical problems].

	Read pool size									
Dataset	**100k**	**200k**	**300k**	**400k**	**500k**	**600k**	**700k**	**800k**	**900k**	**1 M**
CAP3 with previous read sorting										
Arath – 1%	611 (4)	621 (3)	625 (2)	628 (2)	631 (2)	633 (2)	634 (2)	637 (2)	639 (1)	641 (1)
Arath – 2%	693 (5)	711 (4)	720 (3)	725 (3)	730 (3)	734 (2)	739 (2)	743 (2)	750 (2)	756 (2)
Arath – 3%	706 (4)	726 (3)	740 (4)	749 (3)	758 (3)	768 (3)	776 (3)	784 (2)	794 (2)	801 (2)
Arath – 4%	720 (5)	743 (4)	758 (3)	771 (3)	782 (3)	793 (3)	803 (3)	813 (2)	824 (2)	833 (2)
Arath – 5%	734 (5)	760 (3)	778 (4)	794 (3)	809 (3)	823 (3)	834 (3)	846 (3)	862 (2)	873 (2)
Arath – 6%	749 (5)	775 (4)	796 (3)	813 (3)	830 (3)	845 (3)	861 (3)	874 (3)	891 (2)	NA[#]
Arath – 7%	760 (5)	791 (4)	816 (4)	835 (4)	854 (3)	872 (3)	890 (3)	905 (3)	922 (3)	936 (2)
Arath – 8%	772 (5)	803 (4)	829 (4)	852 (4)	874 (3)	895 (3)	914 (3)	932 (3)	949 (2)	965 (2)
Arath – 9%	780 (5)	815 (4)	842 (4)	866 (3)	888 (4)	908 (3)	926 (4)	943 (3)	959 (3)	973 (2)
Lindenbergia	566 (3)	544 (3)	528 (2)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	516 (3)	504 (2)	495 (2)	485 (1)	480 (1)	NA	NA	NA	NA	NA
Striga	560 (3)	549 (2)	540 (2)	533 (2)	530 (2)	NA	NA	NA	NA	NA
Phelipanche	595 (3)	582 (3)	569 (2)	560 (2)	553 (2)	NA	NA	NA	NA	NA
MIRA with previous read sorting										
Arath – 1%	853 (30)	910 (26)	1044 (26)	1046 (24)	796 (7)	776 (6)	763 (5)	748 (4)	741 (3)	738 (2)
Arath – 2%	1218 (51)	1562 (54)	1390 (58)	1258 (33)	936 (9)	898 (7)	871 (6)	853 (4)	844 (3)	837 (2)
Arath – 3%	1425 (58)	1333 (30)	1142 (19)	1016 (13)	934 (6)	898 (7)	867 (5)	846 (4)	836 (3)	830 (3)
Arath – 4%	1715 (73)	1459 (36)	1200 (17)	1016 (10)	933 (9)	887 (7)	858 (6)	837 (4)	830 (3)	827 (3)
Arath – 5%	1967 (98)	1475 (138)	1062 (16)	955 (52)	928 (16)	877 (7)	852 (4)	838 (4)	833 (3)	832 (3)
Arath – 6%	2088 (116)	1246 (29)	1033 (14)	905 (12)	861 (7)	836 (6)	832 (7)	840 (6)	833 (3)	829 (3)
Arath – 7%	2094 (116)	1264 (29)	989 (15)	895 (9)	850 (6)	832 (6)	823 (6)	826 (4)	824 (3)	826 (3)
Arath – 8%	2080 (80)	1216 (29)	965 (15)	886 (12)	846 (7)	828 (7)	820 (5)	816 (3)	814 (3)	822 (2)
Arath – 9%	2126 (87)	1161 (24)	958 (16)	889 (9)	848 (7)	830 (6)	823 (5)	819 (3)	819 (3)	820 (2)
Lindenbergia	1006 (15)	923 (10)	883 (7)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	907 (17)	845 (10)	786 (7)	786 (7)	723 (5)	NA	NA	NA	NA	NA
Striga	957 (15)	884 (11)	873 (8)	873 (8)	813 (6)	NA	NA	NA	NA	NA
Phelipanche	893 (12)	884 (11)	856 (11)	856 (11)	797 (9)	NA	NA	NA	NA	NA

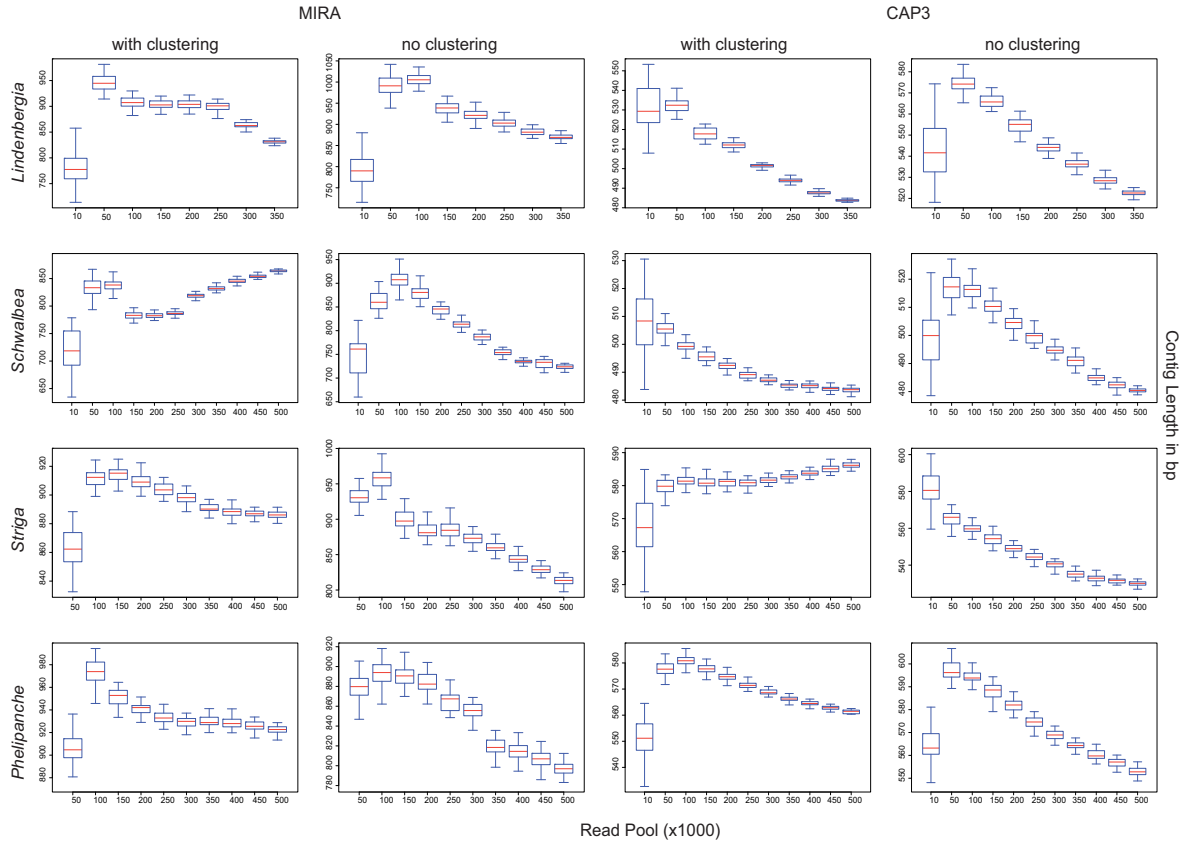


Fig. III-5 **Average contig length in four experimentally generated datasets.** Boxplots for CAP3 and MIRA-assemblies with and without previous read clustering illustrate the difference in contig length depending on the assembled read pool size. Red bars represent the median from 50 samples.

contig length increases again after 250,000 reads. This local minimum was most likely due to the much higher amount of repetitive DNA in *Schwalbea* (compared to other datasets) that bogged down the MIRA assembly process by collision with the critical *megahub* ratio (Chevreau et al. 1999). In assemblies without previous clustering, average contig length reaches its maximum at a read number of 10,000 reads in *Schwalbea*, and thereafter slightly, but significantly decreases by ca. 10-30 bp per read number (Supplemental material: Table SIII-B and SIII-C). In *Lindenbergia*, the largest contigs are produced in assemblies working with a 50,000 reads pool. We measured decreasing contig sizes of up to 150 bp in MIRA assemblies between neighboring sample points, whereas those from CAP3 assemblies hardly ever exceeded 50 bp. However, those differences in the median contig lengths across pool sizes are highly significant based upon 50 samples in Wilcoxon tests (Supplemental Material SIII-B, and SIII-C). Length differences are slightly more pronounced (with both the MIRA and CAP3-assembler) when read sorting is applied (Fig. III-5, Supplemental Material: Tables SIIIB-SIII-C).

3.2. Plastid-specific assembly statistics

3.2.1. Number of plastid contigs varies up to five orders of magnitude between datasets of different ptDNA amounts.

Given low ptDNA amounts in the read pool, our set of analyses in which we separated true plastid contigs from NUPTs and mitochondrial DNA yielded less than 10 % of all CAP3-contigs (Table III-E). The quantity of plastid contigs from CAP3 assemblies increases with the ptDNA-amount. This observation also applies to MIRA assemblies, although it is less obvious here due the overall smaller number of contigs created. The amount of ptDNA-originating contigs is prominently lower in the experimentally generated data sets, in which on average less than 10% of the original contig number are retained, even though a read presorting strategy was employed for the assembly. Reflecting the quantity of CAP3- and MIRA-contigs, respectively, the number of plastid

Table III-E **Number of plastid contigs.** The average number of highly significant MEGABLAST-based alignment hits is summarized according to per assembler and assembled read pool size. Standard deviation is given in brackets. [Abbr.: k - ×1000, M – ×10^6, NA – not sampled; # – not available due to technical problems].

	Read pool size									
Dataset	**100k**	**200k**	**300k**	**400k**	**500k**	**600k**	**700k**	**800k**	**900k**	**1 M**
CAP3 with previous read sorting										
Arath – 1%	175 (11)	377 (18)	566 (17)	741 (16)	903 (17)	1078 (19)	1233 (20)	1393 (23)	1554 (20)	1715 (16)
Arath – 2%	363 (12)	718 (16)	1047 (18)	1353 (23)	1649 (25)	1937 (23)	2218 (31)	2482 (26)	2739 (23)	2995 (21)
Arath – 3%	548 (13)	1062 (22)	1522 (27)	1966 (26)	2376 (26)	2775 (28)	3167 (30)	3527 (31)	3869 (26)	4206 (23)
Arath – 4%	739 (18)	1386 (25)	1995 (27)	2549 (29)	3071 (30)	3562 (40)	4029 (33)	4458 (39)	4881 (28)	5272 (26)
Arath – 5%	903 (17)	1685 (27)	2391 (35)	3052 (29)	3651 (41)	4218 (39)	4736 (37)	5221 (40)	5675 (30)	6103 (30)
Arath – 6%	1072 (25)	1981 (27)	2803 (43)	3538 (31)	4214 (41)	4832 (40)	5402 (34)	5933 (38)	6404 (36)	NA
Arath – 7%	1233 (24)	2267 (36)	3199 (33)	4019 (46)	4779 (40)	5464 (40)	6076 (38)	6650 (39)	7182 (29)	7657 (32)
Arath – 8%	1382 (27)	2537 (31)	3538 (42)	4431 (45)	5238 (46)	5961 (41)	6614 (38)	7221 (46)	7745 (37)	8208 (34)
Arath – 9%	1542 (28)	2822 (40)	3919 (36)	4877 (43)	5751 (41)	6522 (41)	7173 (427)	7836 (50)	8408 (43)	8890 (37)
Lindenbergia	259 (12)	369 (11)	475 (10)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	225 (11)	299 (10)	371 (13)	442 (10)	508 (3)	NA	NA	NA	NA	NA
Striga	178 (9)	219 (9)	263 (11)	310 (12)	347 (9)	NA	NA	NA	NA	NA
Phelipanche	59 (4)	69 (5)	77 (7)	81 (6)	90 (6)	NA	NA	NA	NA	NA
MIRA with previous read sorting										
Arath – 1%	39 (4)	94 (6)	85 (7)	57 (6)	79 (6)	74 (5)	82 (6)	86 (6)	89 (6)	91 (5)
Arath – 2%	84 (5)	51 (7)	58 (16)	59 (7)	90 (7)	109 (9)	131 (9)	136 (8)	131 (7)	131 (8)
Arath – 3%	69 (6)	70 (6)	87 (6)	109 (8)	135 (9)	123 (8)	128 (8)	154 (9)	189 (9)	217 (11)
Arath – 4%	54 (7)	58 (6)	94 (6)	138 (7)	126 (8)	159 (9)	199 (12)	214 (12)	218 (12)	239 (9)
Arath – 5%	42 (5)	69 (21)	162 (18)	145 (19)	166 (14)	213 (11)	225 (10)	246 (13)	278 (10)	323 (9)
Arath – 6%	41 (6)	126 (10)	118 (10)	121 (9)	137 (9)	171 (10)	294 (75)	389 (13)	440 (15)	505 (16)
Arath – 7%	46 (4)	86 (7)	132 (8)	119 (8)	159 (9)	219 (11)	292 (14)	369 (14)	446 (17)	516 (17)
Arath – 8%	51 (6)	102 (8)	121 (10)	136 (9)	196 (12)	273 (11)	353 (16)	429 (17)	505 (19)	606 (17)
Arath – 9%	55 (5)	122 (9)	116 (8)	161 (11)	241 (13)	336 (17)	430 (16)	524 (18)	599 (19)	693 (19)
Lindenbergia	68 (4)	99 (24)	117 (5)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	56 (5)	78 (7)	78 (6)	97 (6)	109 (6)	NA	NA	NA	NA	NA
Striga	NA#	NA#	NA#	NA#	NA#	NA	NA	NA	NA	NA
Phelipanche	34 (3)	23 (3)	17 (3)	16 (3)	14 (4)	NA	NA	NA	NA	NA

hits increases with the size of the read pool saturating towards the largest ones (Table III-E, Fig. III-6). The difference between the two assemblers mirrors the different amounts of generated contigs.

3.2.2. Risk of creating potentially chimeric contigs strongly depends on the employed assembly strategy, and increases with higher read numbers after passing a ptDNA-specific critical sequence pool size.

We would expect a true plastid contig to be unlikely to yield more than one single hit to the reference. More than one hit per contig points to the presence of long repeats such as the large inverted repeat segment of the verified plastid chromosome. Duplicate hits may also occur if contigs harbor a long unalignable sequence stretch (e.g. unknown or low-complexity region). Those stretches should however not occur when the contigs are aligned to a verified species-specific reference sequence. However, such long stretches do occur in divergent plastid-like DNA copies localized in either the nuclear or mitochondrial genome. During the assembly process, those plastid-like sequences can falsely be assembled into true plastid contigs resulting in chimeric contigs. Given a high MEGABLAST stringency, we may thus use the number of duplicate hits as evidence for the amount of potentially chimeric contigs in an assembly. Our results imply that the amount of potential contig chimeras is severely influenced by the number of assembled raw sequences in data sets with high ptDNA ratios, and strongly varies between the employed assemblers. The benefit of greater read pool sizes is counterbalanced by an increasing risk for chimeras, especially in CAP3-assemblies.

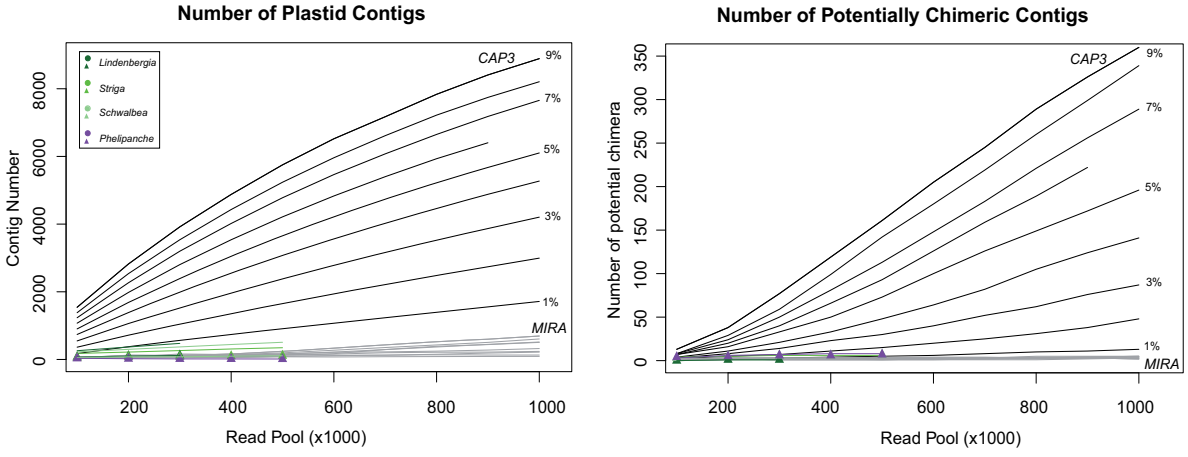


Fig. III-6 **Number of plastid contigs and potentially chimeric contigs using a read-presorting assembly strategy.** The left hand side of this figure illustrates the number of plastid contigs created in assemblies from different read pools of simulated and experimentally generated 454-data. The right hand diagram shows an overview of the amount of contigs with more than one hit against a verified reference sequence evidencing potentially chimeric plastid contigs. Data obtained from the CAP3-assembler is shown by black lines; gray lines represent MIRA-data. Colored dots (CAP3) or triangles (MIRA), respectively, illustrate data from empirical datasets.

Table III-F **Number of putative contig chimera.** Contig sizes obtained from CAP3 and MIRA assemblies using previous read sorting are averaged over 50 samples per read pool for all Arabidopsis dataset as well as for *Lindenbergia*, *Schwalbea*, *Striga* and *Phelipanche*. Standard deviation is given in brackets. [Abbr.: k - ×1000, M – $\times 10^6$, NA – not sampled; # – not available due to technical problems].

	Read pool size									
Dataset	**100k**	**200k**	**300k**	**400k**	**500k**	**600k**	**700k**	**800k**	**900k**	**1 M**
CAP3 with previous read sorting										
Arath – 1%	1 (1)	2 (1)	3 (2)	4 (1)	5 (2)	6 (2)	8 (2)	10 (2)	11 (2)	13 (2)
Arath – 2%	2 (1)	5 (2)	7 (3)	11 (3)	15 (3)	20 (3)	25 (4)	31 (5)	38 (4)	48 (4)
Arath – 3%	3 (2)	8 (1)	14 (3)	23 (4)	30 (4)	40 (5)	52 (7)	62 (6)	76 (6)	87 (7)
Arath – 4%	4 (2)	11 (3)	21 (4)	33 (5)	48 (5)	64 (7)	82 (7)	105 (9)	124 (8)	141 (9)
Arath – 5%	7 (2)	16 (3)	33 (5)	50 (7)	73 (8)	100 (9)	126 (10)	149 (10)	172 (10)	196 (10)
Arath – 6%	7 (3)	20 (4)	40 (6)	66 (7)	93 (8)	126 (11)	159 (10)	189 (12)	222 (12)	NA#
Arath – 7%	8 (2)	24 (5)	50 (6)	81 (10)	113 (9)	148 (10)	183 (11)	221 (12)	256 (10)	289 (10)
Arath – 8%	8 (3)	29 (6)	59 (6)	99 (8)	142 (12)	180 (11)	219 (15)	260 (13)	299 (13)	339 (14)
Arath – 9%	13 (3)	38 (5)	77 (8)	119 (9)	161 (10)	205 (10)	245 (21)	289 (13)	326 (14)	360 (13)
Lindenbergia	0 (1)	1 (0)	1 (0)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	1 (1)	2 (1)	2 (1)	3 (1)	3 (0)	NA	NA	NA	NA	NA
Striga	4 (1)	5 (1)	6 (1)	6 (1)	6 (1)	NA	NA	NA	NA	NA
Phelipanche	3 (1)	6 (2)	7 (1)	8 (1)	8 (1)	NA	NA	NA	NA	NA
MIRA with previous read sorting										
Arath – 1%	0 (0)	1 (1)	2 (1)	2 (1)	2 (1)	2 (1)	3 (1)	3 (1)	3 (1)	2 (1)
Arath – 2%	1 (1)	2 (1)	2 (1)	3 (1)	3 (1)	3 (1)	3 (1)	3 (1)	4 (1)	4 (1)
Arath – 3%	2 (1)	3 (1)	3 (1)	3 (1)	3 (1)	3 (1)	3 (1)	3 (1)	3 (1)	3 (1)
Arath – 4%	2 (1)	3 (1)	3 (1)	3 (1)	4 (1)	4 (1)	4 (1)	3 (1)	3 (1)	2 (1)
Arath – 5%	2 (1)	3 (1)	3 (1)	3 (1)	3 (1)	3 (1)	3 (1)	4 (1)	4 (1)	4 (1)
Arath – 6%	2 (0)	2 (1)	2 (1)	2 (1)	2 (1)	2 (1)	3 (2)	5 (1)	5 (2)	5 (1)
Arath – 7%	2 (1)	3 (1)	2 (1)	2 (1)	1 (1)	1 (1)	2 (1)	2 (1)	3 (1)	2 (1)
Arath – 8%	2 (0)	2 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	2 (1)	3 (1)
Arath – 9%	2 (1)	2 (1)	1 (1)	1 (1)	1 (1)	2 (1)	2 (1)	3 (1)	4 (1)	5 (1)
Lindenbergia	0 (0)	1 (0)	1 (0)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	3 (1)	5 (2)	6 (1)	6 (2)	7 (1)	NA	NA	NA	NA	NA
Striga	NA#	NA#	NA#	NA#	NA#	NA	NA	NA	NA	NA
Phelipanche	4 (1)	5 (1)	6 (1)	6 (1)	7 (1)	NA	NA	NA	NA	NA

Unlike CAP3, MIRA checks for potential chimeras during the assembly process (Chevreux et al. 1999; Chevreux 2011). For that reason, the number of duplicate hits is very low in MIRA-assemblies compared to CAP3. In direct comparison to MIRA, potential chimeras occur extremely frequent in CAP3-assemblies being up to nearly 20% in simulated *Arabidopsis* 454-runs of 5-9% ptDNA. Despite chimeric contig check, the number of potential chimeric MIRA contigs also increases slightly for simulated datasets, starting from read pools of more than 600,000 sequences in case of abundant ptDNA (>4%, table III-F). An explanation for extreme chimera-rate in CAP3 assemblies may be based in frequently occurring alignment overflows during the assembly process of large reads pools. These overflows can lead to artificially elongated low-complexity regions, such as e.g. AT-rich microsatellite stretches that are abundant in non-coding plastid DNA. In the end, this might contribute to a generally higher number and length of microsatellite

regions, especially in the high ptDNA-datasets where additional base calling errors of normally identical fragments introduce further differences, particularly at sequence ends. The latter problem may be overcome by stringent and generous sequence-end trimming of raw reads. Much fewer putative chimeras are created in assemblies of experimentally generated datasets when assembled with CAP3, although MIRA does show in minimal increase in the number of duplicate hits in assemblies from larger read pools (Table III-F). Interestingly, dataset specific amounts of such putative chimeras are not divergent between the two assemblers in case of empirical data. Among the different experimental data, the *Lindenbergia* dataset suffers least from the risk to produce chimeras, whereas *Striga* and *Phelipanche* datasets show the highest risk.

3.2.3. The ratio of desired ptDNA to the remainder genomic sequences is generally underestimated in datasets of different read pool sizes.

We measured the amount of ptDNA in the assembled datasets and compared them to the true (i.e. simulated) ratio. Our analysis of the ptDNA ratio in *Arabidopsis* datasets shows that plastid DNA is slightly underestimated in CAP3-assemblies using read presorting (Table III-G, Fig. III-7). Underestimation increases continuously with a greater abundance of ptDNA in the dataset. Underestimation is more extreme with MIRA than in CAP3 assemblies, which may be due to the generally smaller number of contigs - particularly plastid contigs - created during the assembly. For datasets with high ptDNA-abundance (5-9%), a very small deviation of the simulated ratios from the recovered ones is only obtained in assemblies from 100,000 reads. Lower ptDNA-amounts are estimated nearly correctly in lower ptDNA-dataset until reaching a dataset-specific optimal read pool size. The turning point at which the underestimation takes severe effects in the MIRA-assemblies of empirical data is approximately in line with the results obtained for general contig length that reached maxima between 150,000 and 250,000 reads in MIRA assemblies. CAP3 results for the *Arabidopsis* datasets allow us to approximate *a posteriori* the ptDNA amount of our experimentally generated datasets. Our analyses infers a wide range of ptDNA ratio among the analyzed species. Among photosynthetic species, the lowest ratio was inferred for *Striga hermonthica* with on average 3-4% ptDNA, whereas *Lindenbergia philippensis* shows a mean proportion of likely more than 10% (Table 1) in the un-enriched dataset. The holoparasitic, i.e. non- photosynthetic species *Phelipanche* shows a substantially lower amount of ptDNA, which accounts for about 1 % in the sequence pool. Similarly low ratios have been inferred for another broomrape species, *Orobanche crenata*, in an extra run to reconfirm the low holoparasite ptDNA-amount. Using a dataset of 500,000 reads of this species, we inferred its mean proportion of ptDNA from CAP3 runs (with previous sorting) to range around 0.8%.

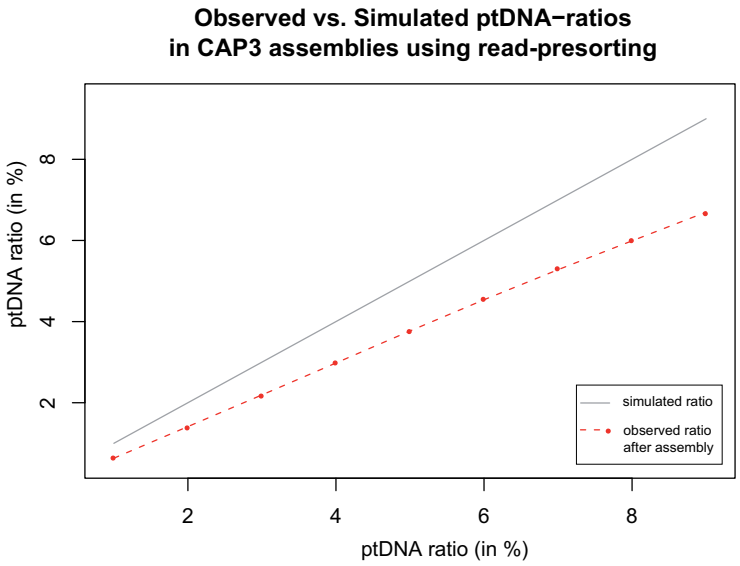


Fig. III-7 **Inference of ptDNA ratio after CAP3 assembly using read pre-sorting.** The mean proportion of plastid DNA has been inferred for all in-silico generated datasets after CAP3 assembly (with read-sorting). The results are illustrated here as a red dashed line in direct comparison to the ptDNA amounts (gray line).

Table III-G **Inferred ptDNA ratio after CAP3 and MIRA assemblies.** Plastid DNA ratio was inferred after aligning contigs to the reference genome for all CAP3 and MIR. ptDNA-abundance was averaged over 50 samples per read pool for all simulated datasets as well as for *Lindenbergia, Schwalbea, Striga* and *Phelipanche*. Standard deviation is given in brackets. [Abbr.: k - ×1000, M – ×10⁶, NA – not sampled; # – not available due to technical problems].

	Read pool size									
Dataset	**100k**	**200k**	**300k**	**400k**	**500k**	**600k**	**700k**	**800k**	**900k**	**1 M**
CAP3 with previous read sorting										
Arath – 1%	0.5 (0.0)	0.6 (0.0)	0.6 (0.0)	0.7 (0.0)	0.7 (0.0)	0.7 (0.0)	0.7 (0.0)	0.7 (0.0)	0.7 (0.0)	0.7 (0.0
Arath – 2%	1.1 (0.0)	1.3 (0.0)	1.4 (0.0)	1.4 (0.0)	1.4 (0.0)	1.5 (0.0)	1.5 (0.0)	1.5 (0.0)	1.5 (0.0)	1.5 (0.0
Arath – 3%	1.9 (0.0)	2.1 (0.0)	2.1 (0.0)	2.2 (0.0)	2.2 (0.0)	2.3 (0.0)	2.3 (0.0)	2.3 (0.0)	2.3 (0.0)	2.3 (0.0
Arath – 4%	2.6 (0.1)	2.9 (0.0)	3.0 (0.0)	3.0 (0.0)	3.1 (0.0)	3.1 (0.0)	3.1 (0.0)	3.1 (0.0)	3.1 (0.0)	3.1 (0.0
Arath – 5%	3.4 (0.1)	3.6 (0.0)	3.8 (0.0)	3.8 (0.0)	3.9 (0.0)	3.9 (0.1)	3.9 (0.0)	3.8 (0.0)	3.8 (0.0)	3.7 (0.0
Arath – 6%	4.2 (0.1)	4.5 (0.1)	4.6 (0.1)	4.7 (0.1)	4.7 (0.1)	4.7 (0.1)	4.6 (0.0)	4.6 (0.1)	4.5 (0.1)	4.5 (0.0
Arath – 7%	5.0 (0.1)	5.3 (0.1)	5.4 (0.1)	5.5 (0.1)	5.4 (0.1)	5.4 (0.1)	5.4 (0.1)	5.3 (0.1)	5.2 (0.1)	5.2 (0.1
Arath – 8%	5.7 (0.1)	6.1 (0.1)	6.2 (0.1)	6.2 (0.1)	6.2 (0.1)	6.1 (0.1)	6.0 (0.1)	5.9 (0.1)	5.8 (0.1)	5.7 (0.1
Arath – 9%	6.5 (0.1)	6.9 (0.1)	7.0 (0.1)	6.9 (0.1)	6.8 (0.1)	6.7 (0.1)	6.5 (0.4)	6.4 (0.1)	6.3 (0.1)	6.1 (0.1
Lindenbergia	8.1 (0.1)	8.1 (0.1)	7.8 (0.0)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	4.8 (0.1)	4.8 (0.0)	4.8 (0.0)	4.8 (0.0)	NA#	NA	NA	NA	NA	NA
Striga	2.3 (0.0)	2.3 (0.0)	2.3 (0.0)	2.4 (0.0)	2.3 (0.0)	NA	NA	NA	NA	NA
Phelipanche	0.7 (0.1)	0.7 (0.0)	0.7 (0.0)	0.7 (0.0)	0.7 (0.0)	NA	NA	NA	NA	NA
MIRA with previous read sorting										
Arath – 1%	0.4 (0.0)	0.6 (0.0)	0.7 (0.1)	0.8 (0.1)	0.7 (0.1)	0.7 (0.0)	0.7 (0.0)	0.7 (0.1)	0.7 (0.1)	0.7 (0.0
Arath – 2%	1.4 (0.1)	1.7 (0.1)	1.8 (0.2)	1.9 (0.2)	1.3 (0.1)	1.1 (0.0)	1.0 (0.0)	0.9 (0.0)	0.8 (0.0)	0.7 (0.0
Arath – 3%	2.1 (0.1)	2.3 (0.3)	2.1 (0.1)	1.7 (0.0)	1.4 (0.0)	1.2 (0.0)	1.1 (0.0)	0.9 (0.0)	0.7 (0.0)	0.6 (0.0
Arath – 4%	3.1 (0.2)	3.1 (0.2)	2.9 (0.1)	1.9 (0.0)	1.6 (0.0)	1.2 (0.0)	0.9 (0.0)	0.7 (0.0)	0.6 (0.0)	0.5 (0.0
Arath – 5%	4.3 (0.6)	3.6 (0.7)	1.9 (0.1)	1.2 (0.5)	1.5 (0.2)	1.0 (0.0)	0.7 (0.0)	0.6 (0.0)	0.6 (0.0)	0.6 (0.0
Arath – 6%	5.5 (0.9)	2.9 (0.1)	1.4 (0.0)	0.7 (0.0)	0.6 (0.0)	0.5 (0.0)	0.6 (0.1)	0.8 (0.1)	0.7 (0.0)	0.7 (0.0
Arath – 7%	6.2 (0.6)	2.6 (0.1)	1.1 (0.0)	0.6 (0.0)	0.5 (0.0)	0.5 (0.0)	0.5 (0.0)	0.6 (0.0)	0.6 (0.0)	0.7 (0.0
Arath – 8%	6.9 (0.7)	2.4 (0.1)	0.9 (0.0)	0.6 (0.0)	0.5 (0.0)	0.6 (0.0)	0.6 (0.0)	0.7 (0.0)	0.8 (0.0)	0.9 (0.0
Arath – 9%	7.9 (0.6)	2.1 (0.1)	0.8 (0.0)	0.6 (0.0)	0.6 (0.0)	0.7 (0.0)	0.8 (0.0)	0.9 (0.0)	1.0 (0.0)	1.1 (0.0
Lindenbergia	7.7 (0.1)	7.7 (0.0)	7.7 (0.0)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	5.3 (0.2)	5.1 (0.3)	4.0 (0.1)	3.3 (0.1)	3.2 (0.0)	NA	NA	NA	NA	NA
Striga	2.6 (0.2)	2.6 (0.1)	2.7 (0.1)	2.6 (0.1)	2.4 (0.1)	NA	NA	NA	NA	NA
Phelipanche	0.6 (0.1)	1.2 (0.3)	1.4 (0.4)	1.1 (0.2)	1.5 (0.7)	NA	NA	NA	NA	NA

3.2.4. Assembly from large sequence pools bears a high risk to produce suboptimal contig length for plastid genome reconstruction.

Plastid contig length develops similar to the remainder contig sizes in CAP3 and MIRA assemblies of simulated 454 datasets of *Arabidopsis*. Using a presorting assembly strategy, the length of CAP3-generated ptDNA contigs steadily increases with read pool size (Table III-I, Fig. III-9). PtDNA contigs are slightly longer than non-ptDNA contigs (Tables III-B and III-I). At their optimal length, they range in size around 800-1100bp. Results obtained from our experimental dataset are mostly in line with *Arabidopsis* CAP3-data, although ptDNA contigs differ less remarkably in length from the contigs of non-plastid origin. In contrast, MIRA generates contiguous sequences of up to 4000 bp from the *in-silico* datasets. The read pool size at which MIRA creates the longest plastid contigs depends substantially on the abundance of the desired genomic region in the sequence pool. In datasets with high ptDNA-content (4-9%), ptDNA contigs are largest with up to 4,700 bp at the smallest sampled read pool of 100,000 sequences (Fig. III-8, Table III-H). For greater sequence pool sizes, we observe a continuously decreasing contig length in assemblies with more than 100,000 reads of datasets with more than 4% ptDNA content (Fig. III-8; also see supplemental material: Tables SIII-D and SIII-E). In contrast, MIRA creates contigs of 15 kb on average from as little as 10,000 reads of the plastid-enriched *Lindenbergia* dataset (Supplemental material: Tables SIII-A). In datasets that are less rich in ptDNA, contig size increases up to a certain read pool size. Similar to high ptDNA-content datasets, MIRA-contig length decreases with more reads. According to the abundance of ptDNA, between 200,000 and 500,000 reads are necessary to reach the maximum contig sequence length. Such local maxima do not only occur in simulated datasets, but also in our experimentally generated ones. Plastid contig length reaches

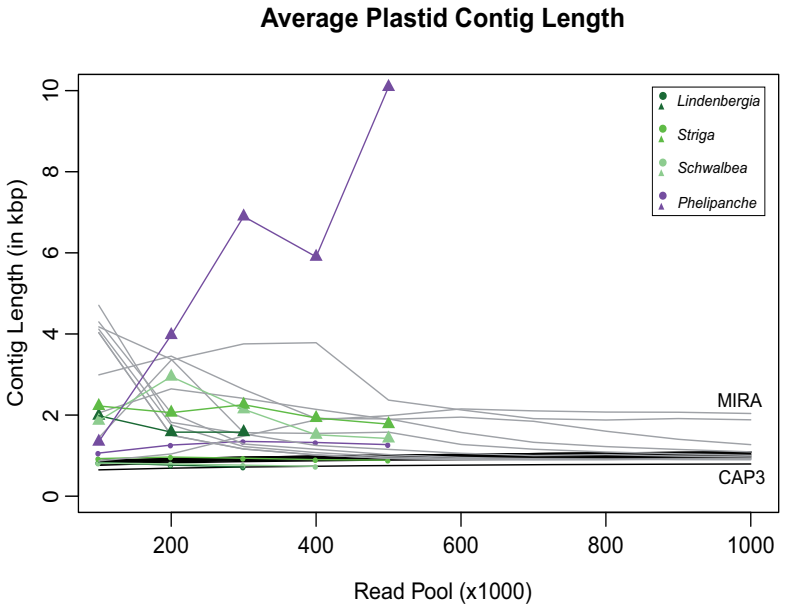


Fig. III-8 **Plastid contig length.** Average plastid contig length in assemblies from different readpool sizes is illustrated for CAP3 with read sorting for *in-silico* data (black lines) and experimental data (colored points). Dark gray lines or colored triangles represent data from the MIRA. Triangles mark direct assemblies.

Table III-H **Length of plastid contigs.** Plastid contig sizes obtained from CAP3 and MIRA assemblies using previous read sorting are averaged over 50 samples per read pool for simulated and empirical datasets. Standard deviation is given in brackets. [Abbr.: k - ×1000, M – ×10⁶, NA – not sampled; # – not available due to technical problems].

	Read pool size									
Dataset	**100k**	**200k**	**300k**	**400k**	**500k**	**600k**	**700k**	**800k**	**900k**	**1 M**
CAP3 with previous read sorting										
Arath – 1%	649 (20)	691 (13)	719 (11)	738 (11)	757 (9)	767 (9)	776 (7)	783 (8)	789 (5)	794 (4)
Arath – 2%	761 (17)	822 (12)	853 (9)	873 (10)	890 (9)	902 (6)	914 (6)	923 (7)	934 (6)	944 (5)
Arath – 3%	792 (12)	845 (9)	879 (9)	898 (9)	916 (7)	931 (7)	940 (6)	953 (7)	963 (4)	972 (4)
Arath – 4%	814 (12)	869 (8)	896 (7)	919 (7)	936 (6)	949 (7)	962 (6)	977 (7)	985 (5)	996 (5)
Arath – 5%	833 (11)	885 (7)	916 (8)	938 (8)	958 (7)	975 (7)	989 (7)	1003 (7)	1018 (6)	1029 (5)
Arath – 6%	852 (10)	901 (9)	931 (8)	954 (6)	974 (8)	990 (6)	1011 (6)	1023 (6)	1039 (5)	1047 (4)
Arath – 7%	861 (9)	912 (8)	946 (7)	969 (8)	989 (6)	1009 (6)	1027 (6)	1043 (7)	1058 (5)	1072 (6)
Arath – 8%	872 (10)	922 (8)	956 (7)	984 (7)	1007 (6)	1030 (6)	1049 (7)	1065 (5)	1081 (6)	1096 (5)
Arath – 9%	877 (7)	927 (7)	962 (6)	988 (6)	1010 (7)	1031 (5)	1048 (7)	1065 (6)	1077 (5)	1088 (5)
Lindenbergia	829 (26)	767 (15)	719 (8)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	815 (27)	796 (18)	765 (15)	736 (11)	NA#	NA	NA	NA	NA	NA
Striga	927 (35)	965 (27)	936 (27)	901 (18)	888 (15)	NA	NA	NA	NA	NA
Phelipanche	1063 (87)	1261 (85)	1350 (109)	1324 (93)	1271 (92)	NA	NA	NA	NA	NA
MIRA with previous read sorting										
Arath – 1%	870 (64)	1041 (56)	1482 (100)	1882 (181)	1985 (179)	2152 (137)	2104 (119)	2074 (112)	2070 (118)	2039 (86)
Arath – 2%	1442 (89)	3355 (386)	3755 (1070)	3784 (647)	2369 (133)	2128 (117)	1912 (78)	1878 (74)	1914 (70)	1883 (76)
Arath – 3%	2052 (189)	2646 (274)	2412 (137)	2141 (103)	1895 (82)	1949 (81)	1847 (70)	1602 (48)	1403 (37)	1271 (29)
Arath – 4%	2991 (335)	3453 (292)	2633 (153)	1915 (71)	1906 (83)	1572 (53)	1329 (35)	1223 (30)	1152 (29)	1093 (16)
Arath – 5%	4175 (514)	3375 (898)	1567 (62)	1547 (415)	1578 (143)	1276 (32)	1158 (18)	1090 (18)	1045 (17)	1007 (10)
Arath – 6%	4706 (810)	1820 (84)	1535 (65)	1250 (39)	1154 (28)	1058 (21)	1002 (18)	1025 (28)	988 (13)	966 (9)
Arath – 7%	4298 (453)	2054 (119)	1286 (43)	1157 (26)	1025 (22)	955 (16)	923 (18)	913 (12)	901 (11)	898 (8)
Arath – 8%	4118 (461)	1758 (86)	1224 (46)	1074 (25)	973 (17)	928 (16)	913 (12)	908 (11)	913 (11)	916 (10)
Arath – 9%	4032 (362)	1504 (61)	1163 (34)	1018 (22)	935 (17)	906 (15)	900 (12)	903 (10)	913 (9)	920 (9)
Lindenbergia	1985 (112)	1579 (70)	1579 (47)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	1860 (134)	2947 (394)	2142 (139)	1514 (50)	1423 (16)	NA	NA	NA	NA	NA
Striga	2224 (260)	2060 (214)	2256 (171)	1928 (117)	1777 (86)	NA	NA	NA	NA	NA
Phelipanche	1344 (152)	3973 (939)	6897 (3370)	5903 (1434)	10090 (8618)	NA	NA	NA	NA	NA

an optimum at ca. 100,000 reads in *Lindenbergia*. In contrast, 150,000-200,000 reads are necessary to obtain largest plastid contigs for *Schwalbea* and *Striga*, respectively. In both CAP3 and MIRA assemblies for our experimental data, contig length continuously and significantly drops after reaching a local optimum in *Lindenbergia, Schwalbea* and *Striga* – irrespective of the applied strategy (Fig III-9, Table III-H; Supplemental material: Table SIII-D and SIII-E). In *Lindenbergia, Schwalbea* and *Striga,* plastid-read clustering contributes to increase average plastid contig length by ca. 500-600bp compared to a direct assembly. We observed a steadily increasing plastid-contig size in MIRA-assemblies for the *Phelipanche*-dataset. Irrespective of the applied assembly strategy, read pools of 500,000 produce plastid contigs of up to 10kb in length. (Table III-H; Supplemental material: Table SIII-E). Contrasting all other datasets, the range of plastid contig lengths steadily increases in *Phelipanche* when MIRA-assembled without read clustering (Fig. III-9). The effect nearly

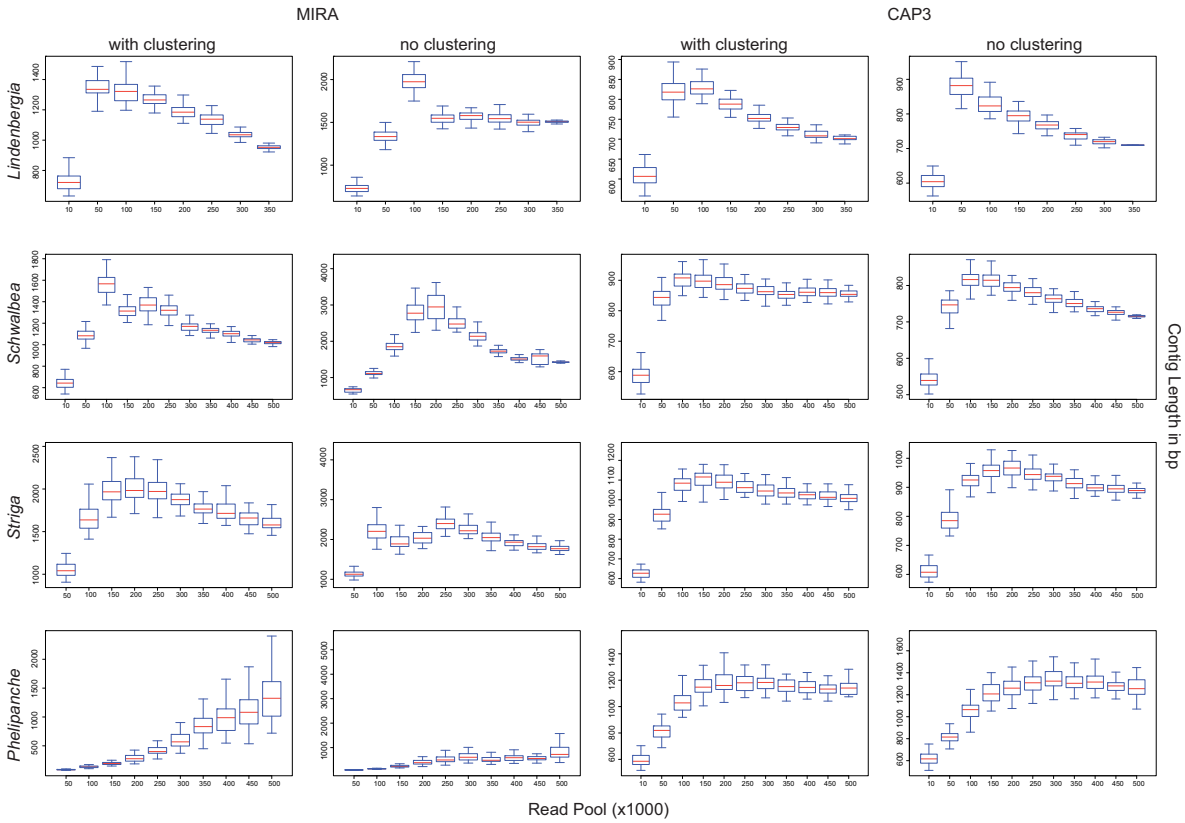


Fig. III-9 **Plastid contig length in four experimental datasets.** Box-whisker blots illustrate plastid contig length in assemblies from differently sized sequence pools employing the MIRA or CAP3 assembler, with and without read presorting. Red bars represent the median of 50 samples per read number.

vanishes in assemblies employing previous read clustering. However, longest contigs are not of a *true* plastid origin as revealed by later analyses. Besides chimeras from plastid and long mitochondrial and/or nuclear sequences, we detected several contigs that were erroneously elongated at regions that contained repeated elements of more than 200 bp in length (refer to Chapter IV for details.). The unusual high amount of plastid-like sequences in the genome of *Phelipanche* provides another possible explanation for the aberrant behavior of this specific dataset. Our analysis shows that previous read sorting partially eliminates contaminations by plastid-like DNA stretches that may lead to chimeric contigs in direct assemblies. Thus, assembly of a desired genomic region harboring longer duplicated elements requires refinement of assembly options. The assembly of regions with unexpected complexity would supposedly benefit from paired-end information rather than additional unpaired reads.

3.2.5. Mean contig quality reaches saturation at dataset-specific read pool sizes.

Besides average read depth (coverage), the mean per-base quality score of a contiguous sequence provides another measure for the confidence of its assembly. Plastid contigs from MIRA assemblies show a marginally higher average per-base quality than

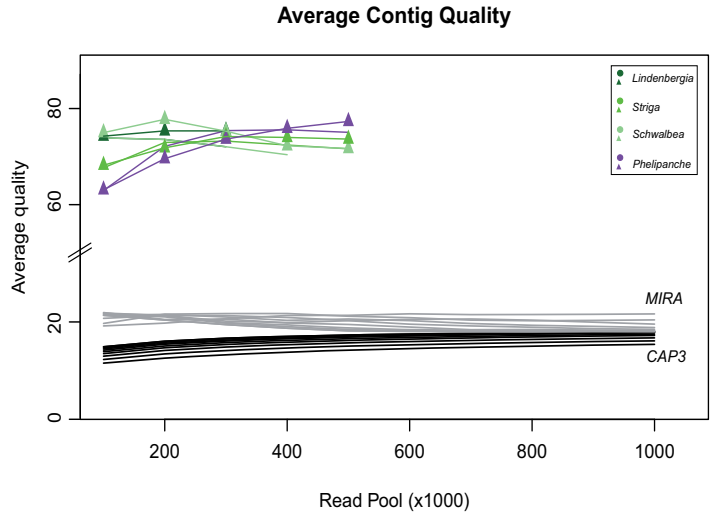


Fig. III-10 Quality of plastid contigs. Average per-base quality of plastid contigs of assemblies from different read pool sizes is illustrated for CAP3 with read pre-sorting for simulated 454-data (black lines) and all experimental datasets (colored points). Dark gray lines or colored triangles represent data from the MIRA-assembler.

Table III-I Average per-base quality of plastid contigs. Contig sizes obtained from CAP3 and MIRA assemblies using previous read sorting are averaged over 50 samples per read pool for all Arabidopsis dataset as well as for *Lindenbergia*, *Schwalbea*, *Striga* and *Phelipanche*. Standard deviation is given in brackets. [Abbr.: k - ×1000, M – ×10⁶, NA – not sampled; # – not available due to technical problems].

	Read pool size									
Dataset	**100k**	**200k**	**300k**	**400k**	**500k**	**600k**	**700k**	**800k**	**900k**	**1 M**
CAP3 with previous read sorting										
Arath – 1%	12.69 (0.15)	13.43 (0.10)	13.93 (0.08)	14.32 (0.10)	14.61 (0.07)	14.84 (0.07)	15.05 (0.08)	15.21 (0.07)	15.36 (0.06)	15.47 (0.04)
Arath – 2%	13.23 (0.09)	14.06 (0.11)	14.58 (0.09)	14.94 (0.08)	15.19 (0.07)	15.42 (0.07)	15.59 (0.07)	15.77 (0.05)	15.89 (0.04)	15.99 (0.03)
Arath – 3%	13.72 (0.10)	14.57 (0.10)	15.08 (0.08)	15.43 (0.07)	15.70 (0.07)	15.90 (0.06)	16.07 (0.05)	16.21 (0.05)	16.33 (0.04)	16.44 (0.03)
Arath – 4%	14.10 (0.11)	14.97 (0.09)	15.46 (0.07)	15.80 (0.07)	16.05 (0.06)	16.27 (0.06)	16.43 (0.04)	16.59 (0.04)	16.70 (0.04)	16.80 (0.03)
Arath – 5%	14.39 (0.11)	15.23 (0.08)	15.74 (0.07)	16.08 (0.05)	16.32 (0.06)	16.50 (0.05)	16.66 (0.04)	16.78 (0.04)	16.89 (0.03)	16.98 (0.02)
Arath – 6%	14.62 (0.10)	15.50 (0.08)	15.97 (0.07)	16.32 (0.06)	16.55 (0.05)	16.75 (0.05)	16.88 (0.04)	17.00 (0.04)	17.11 (0.03)	17.21 (0.02)
Arath – 7%	14.80 (0.09)	15.65 (0.08)	16.10 (0.07)	16.42 (0.06)	16.62 (0.04)	16.80 (0.05)	16.95 (0.04)	17.05 (0.04)	17.13 (0.02)	17.21 (0.02)
Arath – 8%	14.95 (0.11)	15.80 (0.07)	16.27 (0.07)	16.57 (0.06)	16.77 (0.04)	16.92 (0.04)	17.04 (0.04)	17.14 (0.03)	17.25 (0.02)	17.33 (0.02)
Arath – 9%	15.11 (0.09)	15.94 (0.07)	16.38 (0.06)	16.66 (0.04)	16.85 (0.04)	17.01 (0.04)	17.11 (0.06)	17.24 (0.03)	17.32 (0.03)	17.39 (0.02)
Lindenbergia	73.76 (0.73)	70.58 (0.58)	70.58 (0.39)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	73.88 (0.64)	73.57 (0.60)	72.04 (0.55)	70.39 (0.40)	NA#	NA	NA	NA	NA	NA
Striga	67.67 (0.79)	72.94 (0.67)	73.23 (0.63)	72.41 (0.62)	71.61 (0.54)	NA	NA	NA	NA	NA
Phelipanche	62.95 (1.56)	72.13 (1.39)	75.40 (1.09)	75.57 (1.04)	75.05 (1.19)	NA	NA	NA	NA	NA
MIRA with previous read sorting										
Arath – 1%	18.21 (0.31)	18.63 (0.19)	19.29 (0.13)	19.76 (0.20)	19.79 (0.20)	20.01 (0.15)	19.90 (0.17)	19.91 (0.13)	19.94 (0.15)	20.00 (0.13)
Arath – 2%	18.59 (0.18)	19.97 (0.22)	20.05 (0.29)	20.06 (0.21)	19.66 (0.15)	19.44 (0.17)	19.08 (0.18)	18.98 (0.16)	19.03 (0.14)	19.11 (0.15)
Arath – 3%	19.34 (0.17)	19.77 (0.17)	19.75 (0.16)	19.47 (0.17)	19.03 (0.15)	19.18 (0.16)	19.24 (0.16)	19.07 (0.14)	18.78 (0.13)	18.49 (0.13)
Arath – 4%	19.83 (0.21)	19.92 (0.17)	19.54 (0.15)	19.04 (0.16)	19.25 (0.15)	19.00 (0.13)	18.57 (0.14)	18.31 (0.14)	18.15 (0.14)	17.98 (0.12)
Arath – 5%	20.13 (0.21)	19.76 (0.23)	18.93 (0.20)	18.71 (0.25)	18.99 (0.17)	18.48 (0.15)	18.22 (0.10)	18.02 (0.15)	17.81 (0.11)	17.71 (0.10)
Arath – 6%	20.20 (0.21)	19.17 (0.15)	19.22 (0.18)	18.65 (0.16)	18.43 (0.15)	18.05 (0.18)	17.66 (0.13)	17.68 (0.14)	17.60 (0.09)	17.54 (0.07)
Arath – 7%	20.06 (0.18)	19.76 (0.16)	18.70 (0.15)	18.37 (0.18)	17.91 (0.16)	17.68 (0.14)	17.55 (0.14)	17.49 (0.12)	17.47 (0.10)	17.46 (0.09)
Arath – 8%	19.95 (0.20)	19.56 (0.18)	18.63 (0.14)	18.12 (0.16)	17.68 (0.16)	17.53 (0.09)	17.45 (0.11)	17.44 (0.09)	17.44 (0.09)	17.44 (0.08)
Arath – 9%	19.80 (0.18)	19.10 (0.17)	18.42 (0.18)	17.88 (0.18)	17.49 (0.13)	17.41 (0.14)	17.37 (0.10)	17.35 (0.08)	17.40 (0.09)	17.38 (0.07)
Lindenbergia	74.25 (0.33)	75.35 (0.40)	75.35 (0.43)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	74.99 (0.37)	77.75 (0.58)	75.28 (0.44)	72.24 (0.56)	71.68 (0.21)	NA	NA	NA	NA	NA
Striga	68.20 (0.68)	71.82 (0.62)	74.17 (0.52)	73.99 (0.47)	73.64 (0.56)	NA	NA	NA	NA	NA
Phelipanche	63.09 (1.22)	69.56 (0.90)	73.63 (0.64)	75.89 (0.63)	77.31 (0.76)	NA	NA	NA	NA	NA

those of CAP3-assemblies (Table III-I). We observe that the use of large read pools does not significantly improve the mean per-base quality of contigs over that found for smaller read pools in the case of MIRA (Fig. III-10). In CAP3 assemblies, the slopes with which mean quality increases depending on the read pool size saturates around 300,000 reads. This is more pronounced in low-ptDNA content datasets of *Arabidopsis*. Although to a smaller extent, saturation is also evident in MIRA assemblies for ptDNA amounts of up to 4%. Unlike those performed with MIRA, CAP3 assemblies exhibit a slight upwards trend in mean quality. In case of high ptDNA-ratios in the simulated *Arabidopsis* datasets, average per-base quality even decreases in assemblies from large sequence pools. Quality scores are different between simulated and the empirical 454-datasets. For that reason, we observe threefold higher average values in the latter. Nevertheless, mean per-base quality saturates as well between 200,000 and 300,000 reads for both the CAP3 and MIRA assembler (Fig. III-10).

3.2.6. Number and length of assembly gaps reach a local minimum at a dataset-specific read pool size and thereafter increase notably with additional reads.

If assembled with confident and homogeneous coverage, aligning all contigs to a reference genomic region should not leave any regions uncovered in the reference, especially in case of *in-silico* shotgun datasets created by using this particular region. Thus, the number and extension of alignment gaps is another measure for success and quality of (or confidence in) the assembly. Although masked during automatic alignment, long

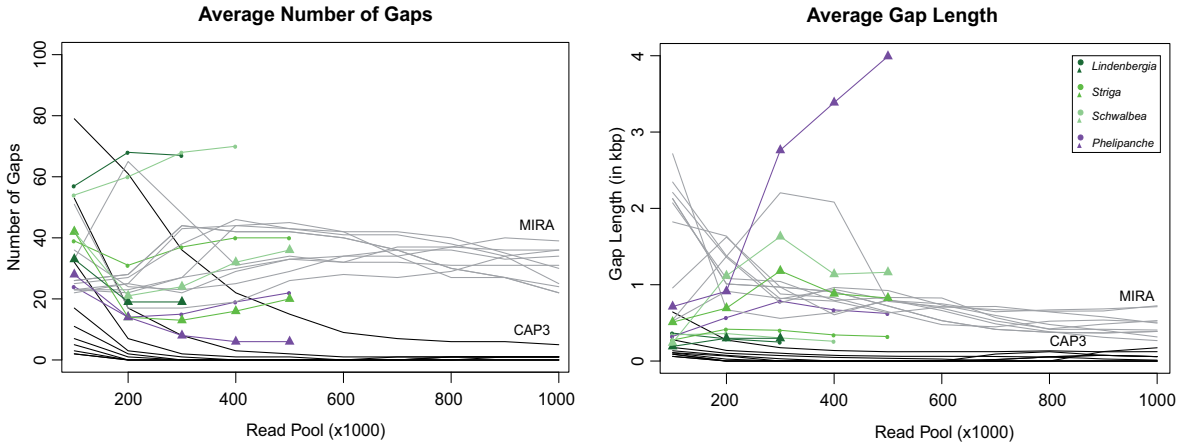


Fig. III-11 **Number and length of alignment gaps.** The left hand side of the figure illustrates the number of alignment gaps between all plastid contigs and a reference sequence. Average length of alignment gaps is shown on the right hand side. Results for *Arabidopsis* data and read presorting are shown using black lines for CAP3-assemblies and gray lines for MIRA-assemblies. Experimental data are represented by colored dots (CAP3/read clustering) or triangles (MIRA/read presorting).

Table III-J **Number of alignment gaps.** The number of gaps in the alignments of a reference sequence to all contigs is summarized for in-silico and experimental data assembled with CAP3 and MIRA under a read presorting strategy. [Abbr.: k - ×1000, M – ×10⁶, NA – not sampled; # – not available due to technical problems]

	Read pool size									
Dataset	**100k**	**200k**	**300k**	**400k**	**500k**	**600k**	**700k**	**800k**	**900k**	**1 M**
CAP3 with previous read sorting										
Arath – 1%	79 (5)	61 (5)	36 (5)	22 (5)	15 (4)	9 (2)	7 (2)	6 (2)	6 (2)	5 (1)
Arath – 2%	53 (6)	17 (3)	8 (3)	3 (2)	2 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)
Arath – 3%	32 (5)	7 (3)	2 (1)	1 (1)	1 (1)	0 (1)	0 (0)	0 (0)	0 (0)	0 (0)
Arath – 4%	17 (4)	3 (2)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0)	1 (1)
Arath – 5%	11 (3)	2 (1)	0 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (1)	0 (1)
Arath – 6%	7 (2)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (1)	0 (0)	0 (0)
Arath – 7%	5 (2)	0 (1)	0 (0)	0 (0)	0 (0)	0 (1)	0 (1)	0 (1)	0 (1)	0 (0)
Arath – 8%	3 (2)	0 (1)	0 (0)	0 (0)	0 (0)	0 (0)	1 (1)	1 (0)	1 (1)	1 (0)
Arath – 9%	2 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (1)	1 (1)	1 (1)	1 (1)
Lindenbergia	57 (6)	68 (5)	67 (5)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	54 (6)	60 (6)	68 (5)	70 (5)	NA#	NA	NA	NA	NA	NA
Striga	39 (5)	31 (5)	37 (5)	40 (5)	40 (5)	NA	NA	NA	NA	NA
Phelipanche	24 (3)	14 (3)	15 (4)	19 (3)	22 (4)	NA	NA	NA	NA	NA
MIRA with previous read sorting										
Arath – 1%	33 (4)	65 (5)	48 (4)	31 (8)	34 (3)	32 (3)	32 (3)	31 (3)	31 (3)	31 (4)
Arath – 2%	51 (4)	17 (4)	17 (4)	19 (4)	26 (4)	28 (3)	27 (4)	29 (3)	34 (4)	36 (4)
Arath – 3%	36 (5)	24 (3)	23 (3)	25 (3)	29 (4)	34 (4)	36 (3)	36 (4)	33 (4)	34 (4)
Arath – 4%	23 (3)	23 (4)	27 (4)	30 (4)	33 (3)	34 (4)	34 (5)	37 (4)	40 (3)	39 (3)
Arath – 5%	22 (3)	25 (3)	22 (5)	29 (4)	33 (4)	32 (3)	37 (3)	37 (4)	36 (3)	36 (4)
Arath – 6%	23 (4)	22 (4)	27 (4)	44 (3)	45 (4)	42 (3)	34 (10)	27 (4)	27 (4)	24 (4)
Arath – 7%	25 (4)	27 (4)	38 (4)	46 (4)	43 (4)	42 (4)	42 (4)	40 (4)	35 (4)	30 (3)
Arath – 8%	23 (4)	25 (5)	43 (3)	44 (4)	43 (4)	41 (4)	41 (4)	38 (4)	34 (3)	25 (4)
Arath – 9%	26 (4)	28 (3)	44 (4)	42 (3)	42 (4)	40 (5)	36 (3)	30 (3)	27 (4)	22 (3)
Lindenbergia	33 (3)	19 (3)	19 (2)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	42 (5)	21 (4)	24 (3)	32 (3)	36 (2)	NA	NA	NA	NA	NA
Striga	42 (4)	14 (3)	13 (2)	16 (3)	20 (2)	NA	NA	NA	NA	NA
Phelipanche	28 (3)	14 (2)	8 (2)	6 (1)	6 (1)	NA	NA	NA	NA	NA

stretches of low-complexity regions (e.g. microsatellite regions, long mononucleotide stretches) may still cause small gaps. The frequency of such gaps is expected to correlate negatively with the amount of sequenced data and its sufficiency to reconstruct the entire region. Hence, the total number as well as the average length of uncovered regions should decrease with additional raw data.

Indeed, the results of our CAP3 assemblies (with read presorting) of our simulated 454-data show the expected pattern (Tables III-J and III-K, Fig. III-11). Gap number and length steadily decreases and reaches zero at a dataset-specific read pool size. Reflecting the abundance of the desired region in the dataset, *Arabidopsis* data with 7-9% ptDNA require at least 200,000 reads for a gapless reconstruction of the plastid chromosome, whereas datasets with 3-6% require between 300,000 and 400,000 reads. A low number of (short) gaps remain, however, in datasets with less than 2% ptDNA abundance. In contrast, a gap-free reconstruction of the entire desired region was not possible with MIRA, although gap number and length tend to decrease for large read pools. Gaps also

Table III-K **Length of gaps.** The length of gaps in the alignments (if any) of a reference sequence to all contigs is summarized for nine simulated and four empirical datasets assembled with CAP3 and MIRA under a read presorting strategy. [Abbr.: k - ×1000, M – ×10⁶, NA – not sampled; # – not available due to technical problems]

	Read pool size									
Dataset	**100k**	**200k**	**300k**	**400k**	**500k**	**600k**	**700k**	**800k**	**900k**	**1 M**
CAP3 with previous read sorting										
Arath – 1%	643 (67)	275 (34)	175 (30)	138 (24)	122 (26)	123 (44)	123 (47)	134 (47)	121 (39)	123 (35)
Arath – 2%	279 (33)	139 (41)	103 (37)	73 (48)	63 (46)	65 (57)	58 (71)	51 (64)	27 (40)	12 (24)
Arath – 3%	178 (32)	102 (38)	72 (55)	48 (65)	36 (74)	26 (60)	None	None	None	None
Arath – 4%	135 (35)	75 (45)	30 (62)	None	None	None	None	None	124 (175)	174 (212)
Arath – 5%	110 (40)	66 (60)	None	None	None	None	None	None	None	None
Arath – 6%	108 (57)	25 (38)	None	None	None	None	None	None	None	None
Arath – 7%	92 (40)	None	None	None	None	None	None	None	None	None
Arath – 8%	89 (66)	None	None	None	None	None	92 (123)	122 (118)	75 (104)	58 (87)
Arath – 9%	61 (48)	None	None	None	None	None	22 (40)	57 (65)	75 (63)	58 (52)
Lindenbergia	366 (46)	291 (40)	248 (26)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	344 (71)	360 (50)	303 (28)	259 (27)	NA#	NA	NA	NA	NA	NA
Striga	275 (56)	417 (91)	403 (66)	342 (49)	321 (34)	NA	NA	NA	NA	NA
Phelipanche	323 (80)	569 (216)	785 (207)	670 (134)	623 (151)	NA	NA	NA	NA	NA
MIRA with previous read sorting										
Arath – 1%	2716 (380)	672 (84)	559 (195)	636 (206)	792 (168)	750 (142)	646 (112)	670 (98)	683 (103)	712 (129)
Arath – 2%	539 (80)	1370 (500)	2204 (1109)	2082 (583)	782 (186)	712 (165)	674 (133)	653 (135)	656 (113)	724 (95)
Arath – 3%	524 (98)	915 (228)	822 (196)	785 (152)	631 (128)	708 (114)	715 (110)	644 (78)	579 (92)	497 (68)
Arath – 4%	958 (362)	1624 (370)	969 (201)	606 (95)	802 (150)	693 (133)	525 (101)	422 (70)	484 (70)	530 (71)
Arath – 5%	1823 (551)	1637 (549)	879 (272)	831 (140)	646 (114)	477 (73)	448 (62)	522 (77)	537 (78)	510 (87)
Arath – 6%	2072 (494)	1082 (312)	1042 (212)	793 (82)	833 (111)	826 (119)	569 (209)	415 (67)	416 (60)	406 (69)
Arath – 7%	2209 (519)	1360 (250)	763 (124)	964 (107)	924 (129)	734 (107)	597 (109)	479 (78)	403 (71)	316 (60)
Arath – 8%	2345 (571)	1378 (305)	815 (124)	939 (119)	802 (135)	662 (96)	487 (84)	384 (60)	311 (48)	268 (44)
Arath – 9%	2128 (345)	1010 (221)	969 (124)	909 (112)	722 (135)	529 (83)	415 (70)	377 (59)	368 (64)	388 (67)
Lindenbergia	192 (50)	298 (152)	298 (171)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	239 (87)	1117 (380)	1631 (262)	1138 (188)	1163 (123)	NA	NA	NA	NA	NA
Striga	510 (178)	693 (357)	1181 (400)	883 (214)	826 (139)	NA	NA	NA	NA	NA
Phelipanche	714 (136)	914 (513)	2762 (1110)	3386 (1413)	3992 (1943)	NA	NA	NA	NA	NA

remained in the 454-data of plastid DNA enriched *Lindenbergia* irrespective of the assembler and assembly strategy (Supplemental material: Table SIII-A). However, in both CAP3- and MIRA assemblies, we observe that gaps are introduced by assembling large read pools of both empirical and simulated datasets. This may be due to the higher risk of chimeric contigs by low-copy nuclear or mitochondrial plastid like-fragments. All datasets exhibit a local minimal gap number at a specific read pool size. The effect of a local minimum of gaps at larger sequence pools is even more dramatic in our experimental datasets (Fig. III-11; Supplemental material: Fig. SIII-1, Table SIII-F and SIII-G). In all datasets, the number of uncovered regions ranges between 50 and 90 until the plastid chromosome is entirely covered. A minimum number of gaps occur between 100,000 and 250,000 reads depending on dataset, assembler and assembly strategy. Except for *Phelipanche* and irrespective of the employed assembly strategy, the amount of gaps increases after this point with the amount of raw read. Such unknown regions tend to be

longer in MIRA-assemblies, although the length of unknown regions varies greatly between the different datasets. In particular, for our experimental data it seems as if not only the abundance of ptDNA in the read pool contributes to this effect. Gap number and length may also reflect their individual genomic peculiarities with respect to rearrangements, complexity, and repeat elements. For CAP3, the length of unalignable regions ranges between 200-400 bp in *Schwalbea* and *Lindenbergia*, and around 500bp in *Striga* and *Phelipanche (*Supplemental material: Fig. SIII-2, Table SIII-H). Similar to the situation for *in silico* datasets, adding more reads to the *Lindenbergia* and *Schwalbea* read pool barely influences gap sizes in MIRA assemblies *(*Supplemental material: Fig. SIII-2, Table SIII-I). Length of uncovered region seems to be influenced by the assembly strategy in that we observe a weak correlation of larger gaps when read sorting precedes the MIRA-assembly. The effect is more pronounced in *Striga* and *Phelipanche*. In *Lindenbergia*, *Striga* and *Schwalbea* the amount and length largely correlates with the number and length of low-complexity (mostly AT-rich) plastid regions (compare chapter IV). The very long gaps of up to 3,000 bp in *Phelipanche* supposedly have another reason. As mentioned earlier, one may be misassembled plastid-like contigs that align only partially with the reference genome. Another explanation may be the very complex general structure of the *Phelipanche* plastid genome that, although highly reduced compared to "normal" green plant plastomes, it contains a high number of repeated elements as well as numerous reorganization and several untypical sequences that are unique to the *Phelipanche* plastome (see chapter IV for details). The latter elements cannot be detected during read filtering using standard plastid read archives, and thus require manual refinement.

3.2.7. Read depth strictly depends on the abundance of the desired genomic region.

Coverage of up to approx. 200x can be obtained from 454-datasets with an ptDNA abundance of 9% when the largest possible read pool size is used for CAP3-based reconstruction of the plastid chromosome. In contrast, rarely more than 20x are reached with ptDNA-amounts around 1%. Coverage of the plastid genome increases with read pool size in CAP3 assemblies (Fig. III-12, Table III-L). In contrast, no such clear relation between coverage and either the ptDNA-abundance or the assembled read pool size exists in MIRA assemblies of the simulated data. Low-ptDNA datasets show a steady, nearly linear increase of coverage up to a dataset-specific read pool size. However, coverage of MIRA-contigs drops by adding more reads to the assembly pool. In ptDNA-rich datasets (>5%), we observe decreasing coverages with MIRA reaching a local minimum between 400,000 and 600,000 reads. More reads lead to a nearly linear coverage increase. The turning point ranges around 20-30x coverage (Table III-L). Around the same read pool sizes, we also observed the largest contig sizes as well as least and shortest gaps (see sections 3.6, 3.7, 3.9). The complex relation of coverage and pool size may partly underlie

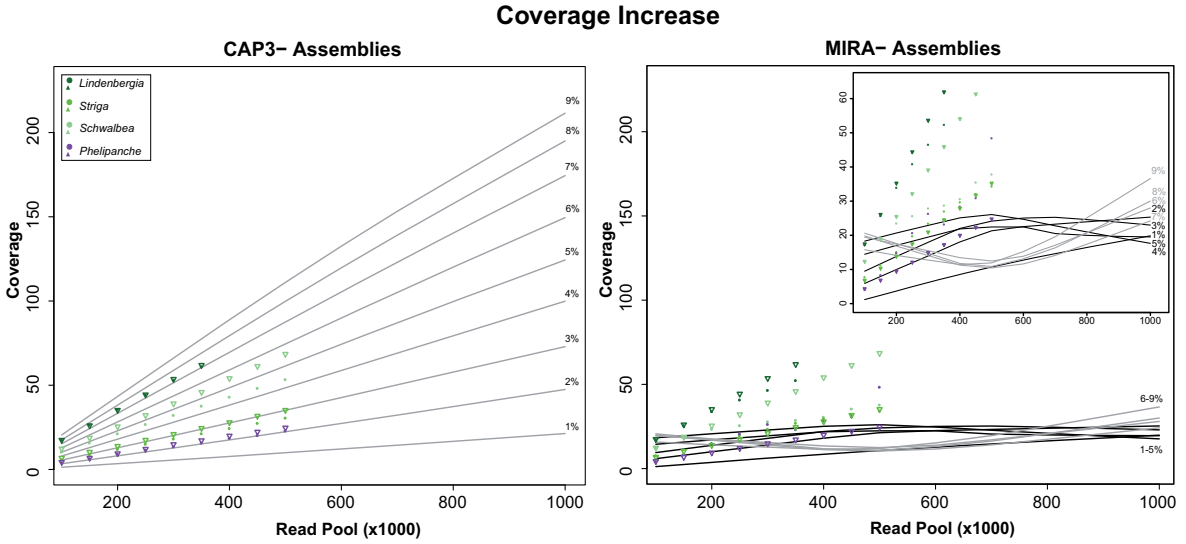


Fig. III-12 **Relationship of coverage and read pool size**. Increase of coverage with expanded read pools follows a strict linear trend in CAP3 assemblies of both simulated data (gray lines) and empirical datasets (colored dots). Data obtained for empirical datasets (triangles) assembled without read presorting align well to the pattern of the pre-clustering method. Coverage does not develop uniformly in MIRA assemblies, and deviates strongly between experimental and simulated datasets. For better visualization, *in-silico* dataset with ptDNA up to 5% are drawn as black lines; ptDNA datasets above 5% are illustrated in gray. The inset displays a close-up view of the *Arabidopsis* results.

the aforementioned results on decreasing contig lengths, extreme underestimation of ptDNA abundance as well as contig number development at different read pools. On the other hand, assembly at moderate coverages (ca. 20 to 30X) appears to reduce the risk of potential chimeric contigs. The difference of coverage between the two assemblers is not that severely pronounced in our experimental datasets (Fig. III-12, Supplemental Material: Fig. SIII-3). Above that, the latter exhibit a much stronger distortion in assemblies without read presorting. Up to 20-25x, plastid chromosome coverage increases linearly with read number. Thereafter, the coverage increment slightly descends beyond ~20x-equivalent read pools. Assemblies obtained from the *Phelipanche* dataset severely differ from the remainder. High abundances of ambiguous repeat sequences and divergent plastid-like nuclear/mitochondrial reads may cause extremes in samples for read pools larger than 250,000 sequences (Supplemental Material: Fig. SIII-3). Clustering, however, supposedly filters the majority of contaminating elements prior to the assembly, which substantially weakens the effect. The strong dependence of assembly quality on coverage entails that a reliable estimate of the number of reads required to achieve a given coverage would also allow approximating the corresponding quality or confidence level of the final assembly. Kendall and Spearman tests confirmed a general correlation throughout all *in silico*- and experimental datasets (Table III-M). Given the strong correlation of coverage and sequence pool in CAP3 assemblies, we inferred a model to estimate the required number of reads *a priori* from the abundance of a particular region in the sequence pool and the desired coverage. In order to examine the extent to which ptDNA ratios influences the coverage

Table III-L **Observed coverage after CAP3 and MIRA assemblies in simulated and experimental dataset.** Contig sizes obtained from CAP3 and MIRA assemblies using previous read sorting are averaged over 50 samples per read pool for all in-silico and experiment datasets. Standard deviation is given in brackets. [Abbr.: k - ×1000, M – ×10⁶, NA – not sampled].

	Read pool size									
Dataset	**100k**	**200k**	**300k**	**400k**	**500k**	**600k**	**700k**	**800k**	**900k**	**1 M**
CAP3 with previous read sorting										
Arath – 1%	1.4 (0.1)	3.4 (0.1)	5.5 (0.1)	7.7 (0.1)	9.8 (0.2)	12.2 (0.2)	14.3 (0.2)	16.6 (0.2)	19.0 (0.2)	21.4 (0.2)
Arath – 2%	3.5 (0.1)	8.0 (0.2)	12.7 (0.2)	17.4 (0.2)	22.3 (0.3)	27.3 (0.3)	32.3 (0.3)	37.2 (0.4)	42.3 (0.3)	47.6 (0.3)
Arath – 3%	5.7 (0.1)	12.7 (0.2)	20.0 (0.3)	27.5 (0.3)	35.0 (0.4)	42.7 (0.5)	50.4 (0.6)	57.9 (0.5)	65.3 (0.5)	73.0 (0.6)
Arath – 4%	8.2 (0.2)	17.8 (0.3)	27.9 (0.4)	38.1 (0.4)	48.5 (0.5)	58.7 (0.7)	69.0 (0.6)	79.8 (0.8)	89.8 (0.9)	99.6 (0.9)
Arath – 5%	10.5 (0.2)	22.7 (0.3)	35.5 (0.4)	48.4 (0.5)	61.5 (0.7)	74.7 (0.9)	87.4 (1.1)	99.9 (1.2)	112.4 (1.2)	124.1 (1.0)
Arath – 6%	13.0 (0.2)	28.0 (0.4)	43.5 (0.5)	59.1 (0.7)	74.6 (0.8)	90.2 (1.2)	105.8 (1.0)	120.4 (1.5)	135.0 (1.5)	149.0 (1.2)
Arath – 7%	15.4 (0.3)	33.0 (0.5)	51.5 (0.6)	69.7 (1.0)	87.6 (1.0)	105.9 (1.2)	123.5 (1.1)	140.8 (1.7)	157.6 (1.5)	174.1 (1.9)
Arath – 8%	17.7 (0.3)	38.0 (0.4)	58.7 (0.6)	79.6 (0.8)	100.4 (1.3)	120.5 (1.4)	139.8 (1.7)	158.5 (1.6)	177.0 (1.8)	194.0 (2.0)
Arath – 9%	20.3 (0.3)	43.3 (0.5)	66.6 (0.7)	89.4 (1.0)	111.5 (1.3)	133.3 (1.4)	153.1 (10.1)	174.8 (2.0)	193.2 (1.7)	210.0 (2.3)
Lindenbergia	17.4 (0.2)	35.7 (0.3)	53.1 (0.3)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	10.4 (0.2)	21.2 (0.2)	32.2 (0.3)	43.0 (0.3)	NA#	NA	NA	NA	NA	NA
Striga	5.9 (0.1)	12.0 (0.2)	18.2 (0.3)	24.5 (0.3)	30.6 (0.3)	NA	NA	NA	NA	NA
Phelipanche	4.3 (0.3)	9.1 (0.5)	13.7 (0.9)	18.7 (1.2)	23.1 (1.2)	NA	NA	NA	NA	NA
CAP3 without previous read sorting										
Lindenbergia	17.5 (0.2)	35.3 (0.3)	53.9 (0.5)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	12.4 (0.2)	25.5 (0.3)	39.2 (0.4)	53.0 (0.4)	66.9 (0.4)	NA	NA	NA	NA	NA
Striga	6.8 (0.2)	13.9 (0.3)	21.0 (0.4)	28.1 (0.5)	35.3 (0.5)	NA	NA	NA	NA	NA
Phelipanche	4.4 (0.3)	9.6 (0.7)	15.0 (0.7)	20.0 (0.9)	25.1 (1.3)	NA	NA	NA	NA	NA
MIRA with previous read sorting										
Arath – 1%	1.1 (0.1)	3.7 (0.2)	6.3 (0.3)	9.2 (0.8)	11.4 (1.0)	13.1 (0.9)	15.3 (0.9)	17.3 (1.2)	19.0 (1.4)	20.0 (1.1)
Arath – 2%	4.7 (0.2)	11.0 (0.5)	17.6 (1.5)	24.3 (2.6)	21.7 (0.9)	23.1 (0.8)	23.0 (0.5)	23.8 (0.6)	24.8 (0.5)	25.3 (0.4)
Arath – 3%	7.1 (0.3)	15.1 (1.7)	20.6 (0.7)	23.5 (0.6)	24.3 (0.7)	25.3 (0.4)	26.1 (0.5)	25.4 (0.6)	24.0 (0.5)	22.4 (0.5)
Arath – 4%	10.2 (0.7)	20.4 (1.3)	28.5 (1.1)	26.1 (0.7)	27.3 (0.6)	25.6 (0.7)	22.4 (0.6)	20.3 (0.4)	18.8 (0.4)	18.4 (0.3)
Arath – 5%	14.0 (1.8)	24.1 (4.4)	20.5 (0.8)	17.3 (6.2)	26.8 (3.2)	22.3 (0.6)	19.5 (0.5)	18.7 (0.5)	19.2 (0.4)	20.7 (0.5)
Arath – 6%	17.9 (2.7)	20.0 (0.7)	15.6 (0.5)	11.1 (0.4)	12.0 (0.4)	11.9 (0.4)	15.3 (2.3)	23.6 (2.3)	24.6 (0.5)	27.3 (0.5)
Arath – 7%	20.1 (2.0)	18.4 (0.7)	12.2 (0.5)	9.3 (0.3)	9.5 (0.3)	10.8 (0.3)	13.5 (0.7)	17.2 (0.5)	20.8 (0.4)	24.5 (0.5)
Arath – 8%	22.6 (2.3)	17.4 (0.6)	10.5 (0.4)	9.2 (0.3)	10.3 (0.3)	12.6 (0.4)	15.7 (0.5)	19.7 (0.5)	24.8 (0.5)	31.1 (0.4)
Arath – 9%	25.8 (1.8)	15.3 (0.6)	9.1 (0.4)	9.4 (0.4)	11.1 (0.4)	14.6 (0.6)	18.9 (0.5)	24.3 (0.5)	30.4 (0.6)	37.5 (0.7)
Lindenbergia	17.1 (0.2)	33.9 (0.2)	33.9 (0.3)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	12.4 (0.4)	23.5 (1.2)	27.9 (0.7)	30.6 (0.6)	37.8 (0.2)	NA	NA	NA	NA	NA
Striga	7.7 (0.7)	15.1 (0.6)	23.4 (1.0)	29.7 (1.0)	34.3 (1.2)	NA	NA	NA	NA	NA
Phelipanche	4.1 (0.5)	14.8 (3.6)	26.3 (7.8)	28.4 (4.3)	48.4 (22.5)	NA	NA	NA	NA	NA
MIRA without previous read sorting										
Lindenbergia	17.0 (0.0)	27.8 (0.0)	28.6 (0.1)	NA	NA	NA	NA	NA	NA	NA
Schwalbea	12.1 (0.0)	19.2 (0.0)	27.4 (0.1)	30.5 (0.1)	31.5 (0.0)	NA	NA	NA	NA	NA
Striga	7.5 (0.0)	15.5 (0.1)	22.3 (0.1)	27.3 (0.2)	30.7 (0.2)	NA	NA	NA	NA	NA
Phelipanche	4.2 (0.1)	13.2 (0.4)	35.4 (1.1)	71.0 (2.3)	113.8 (3.4)	NA	NA	NA	NA	NA

slope with higher read numbers, we at first optimized generalized linear models to the CAP3 data (Table III-M). The estimated functions describe in how far the size of the required read pool is determined by the desired coverage and the estimated ptDNA ratio in the total genomic DNA extract. Using the inferred model functions, we determined the amount of reads necessary to reach 20x coverage. As little as 30 million sequenced basepairs are necessary for the confident assembly of the plastid chromosome from a

Table III-M **PtDNA ratio and optimal coverage.** Generalized linear models have been computed to evaluate the linear correlations between plastome coverage and read number for simulated datasets assemblies and all empirical datasets. Coefficients of each model function are provided including their respective standard error (SE) and p-values. General correlations were tested using the Kendall and the Spearman correlation test. Based upon the inferred models, the sequencing effort has been calculated for a 20x coverage.

	Results for generalized linear models				Kendall		Spearman		Reads	Mbp.
Dataset	**Intercept (SE)**	**p-value**	**Slope (SE)**	**p-value**	**Tau**	**p_τ -value**	**Rho**	**p_ρ -value**	**(20 x)**	**(20 x)**
CAP3 with previous read sorting										
Arath – 1%	50,195.3 (846.5)	<0.001***	44,900.0 (65.9)	<0.001***	0.95	<0.001***	0.995	<0.001***	948,194	341.4
Arath – 2%	39,312.3 (735.5)	<0.001***	20,381.7 (25.6)	<0.001***	0.95	<0.001***	0.995	<0.001***	446,945	160.9
Arath – 3%	30,439.5 (591.1)	<0.001***	13,312.4 (13.3)	<0.001***	0.95	<0.001***	0.995	<0.001***	296,688	106.8
Arath – 4%	25,649.1 (614.0)	<0.001***	9,759.2 (10.0)	<0.001***	0.95	<0.001***	0.995	<0.001***	220,832	79.5
Arath – 5%	18,802.8 (697.1)	<0.001***	7,844.7 (9.1)	<0.001***	0.95	<0.001***	0.995	<0.001***	175,697	63.3
Arath – 6%	13,268.5 (738.7)	<0.001***	6,554.2 (8.1)	<0.001***	0.949	<0.001***	0.995	<0.001***	144,353	52
Arath – 7%	9,756.6 (768.4)	<0.001***	5,632.4 (7.1)	<0.001***	0.95	<0.001***	0.995	<0.001***	122,405	44.1
Arath – 8%	1,977.1 (1,047.1)	0.06	5,054.3 (8.6)	<0.001***	0.95	<0.001***	0.995	<0.001***	103,064	37.1
Arath – 9%	-9,068.6 (2,013.3)	<0.001***	4,677.2 (15.0)	<0.001***	0.947	<0.001***	0.994	<0.001***	84,476	30.4
Lindenbergia	362.4 (209.8)	0.085	5,659.8 (5.7)	<0.001***	0.937	<0.001***	0.992	<0.001***	113,559	28.5
Schwalbea	2,891.3 (177.3)	<0.001***	9,258.0 (5.6)	<0.001***	0.954	<0.001***	0.996	<0.001***	188,051	48.3
Striga	4,290.3 (313.5)	<0.001***	16,211.2 (17.4)	<0.001***	0.954	<0.001***	0.996	<0.001***	328,515	103.8
Phelipanche	9,570.6 (1,316.6)	<0.001***	21,106.5 (96.7)	<0.001***	0.945	<0.001***	0.993	<0.001***	431,700	148.5
CAP3 without previous read sorting										
Lindenbergia	2,651.2 (230.8)	<0.001***	5,564.5 (6.2)	<0.001***	0.937	<0.001***	0.992	<0.001***	113,941	28.6
Schwalbea	10,832.6 (517.2)	<0.001***	7,232.6 (13.1)	<0.001***	0.954	<0.001***	0.996	<0.001***	155,484	39.9
Striga	4,295.3 (411.8)	<0.001***	14,042.4 (19.8)	<0.001***	0.954	<0.001***	0.996	<0.001***	285,143	90.1
Phelipanche	12,915.0 (1,072.9)	<0.001***	19,372.2 (79.3)	<0.001***	0.947	<0.001***	0.994	<0.001***	400,359	137.7
MIRA with previous read sorting										
Lindenbergia	3,123.3 (221.2)	<0.001***	5,753.4 (10.6)	<0.001***	0.896	<0.001***	0.98	<0.001***	118,191	29.7
Schwalbea	4,949.5 (1,148.2)	<0.001***	7,884.3 (76.7)	<0.001***	0.884	<0.001***	0.973	<0.001***	162,636	41.8
Striga	11,272.3 (1,282.2)	<0.001***	12,331.8 (74.1)	<0.001***	0.926	<0.001***	0.989	<0.001***	257,909	81.5
Phelipanche	67,220.5 (4,994.5)	<0.001***	10,416.3 (240.3)	<0.001***	0.794	<0.001***	0.915	<0.001***	275,546	94.8
MIRA without previous read sorting										
Lindenbergia	1,925.7 (466.3)	<0.001***	6,035.5 (30.6)	<0.001***	0.868	<0.001***	0.968	<0.001***	122,636	30.8
Schwalbea	3,821.7 (560.1)	<0.001***	8,421.1 (52.2)	<0.001***	0.868	<0.001***	0.968	<0.001***	172,244	44.3
Striga	7,195.7 (975.0)	<0.001***	12,790.8 (66.3)	<0.001***	0.914	<0.001***	0.986	<0.001***	263,011	83.1
Phelipanche	104,403.8 (3,325.6)	<0.001***	4,793.6 (121.1)	<0.001***	0.91	<0.001***	0.983	<0.001***	200,276	68.9

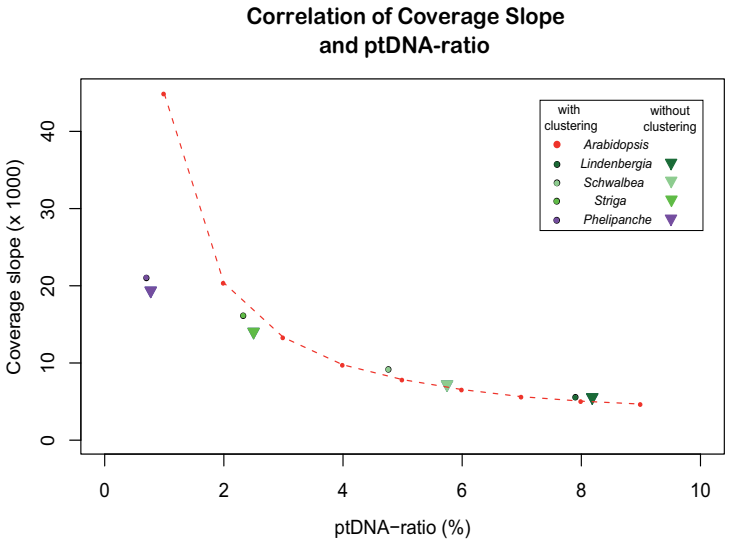


Fig. III-13 Correlation of coverage and ptDNA abundance. Coverage increments derived from linear regression of coverage data was plotted against the respective ptDNA-ratio.

dataset with an estimated ptDNA abundance of 9 %. In contrast, low ptDNA abundances of 1% require 454 sequencing of nearly one full picotiter plate in order to obtain the necessary amount of 340 Mbp (Table M). Thus, the regression model slope inclines substantially in ptDNA-poor datasets.

*3.2.8. The exponential decay function y(x) ≅ a*e(-bx) allows an a priori estimation of the optimal read pool size for the assembly of a specific genomic region.*

In order to analyze the extent to which ptDNA ratio influences the slope, we plotted the slopes of each *Arabidopsis* regressions against the respective ptDNA ratio (Fig. III-13). The trend curve evokes an exponential decay function describing the relationship of plastid DNA ratio and increase of coverage. Results of our experimental datasets fit very well to the *Arabidopsis* data suggesting that this trend may be used for an *a priori* estimate of the optimal assembly read pool size. We fitted a nonlinear regression model through our data assuming that coverage slope depends on ptDNA-ratio (Table III-N). The function parameters A and B were estimated with high confidence ($p << 0.001$***). The inferred correlation coefficient R ($R = 0.9642$) as well as the F test ($p << 0.001$***) strongly support our inferred model (Fig. III-14). The results are suitable to roughly predict the required sequence amount for any given ptDNA ratio. Hence, we may approximate the read number required for an optimal assembly *a priori* based on a given assumed ptDNA ratio (within the given confidence intervals) as:

$$\mathbf{n \approx 70{,}000e^{-0.5XC}}$$

where n is read number, C is the desired coverage, and X is the ptDNA ratio.

Table III-N **Nonlinear regression model estimation describing the dependence of coverage slope from ptDNA ratio.** The nonlinear model function $y(x) \cong a \times e^{(-bx)}$ describes the correlation of coverage slope and ptDNA-ratio throughout all CAP3-results from the read presorting assembly. The coefficient *a* as well as the scale parameter *b* of the inferred model are summarized below including its quality statistics, and its upper and lower confidence interval.

Nonlinear model estimates for function $y(x) \cong a \times e^{(-bx)}$						Goodness of fit		
Parameter	**Estimate**	**t-value (df = 7)**	**p-level**	**Lower Conf. limit**	**Upper Conf. limit**	**Correlation Coefficient R**	**F-test**	**p_F-level**
A	68027 (9121)	7.458	<0.001***	46,459.13	89,594.57	0.9642	102.536	<0.001***
B	0.49 (0.07)	6.839	<0.001***	0.32	0.66			

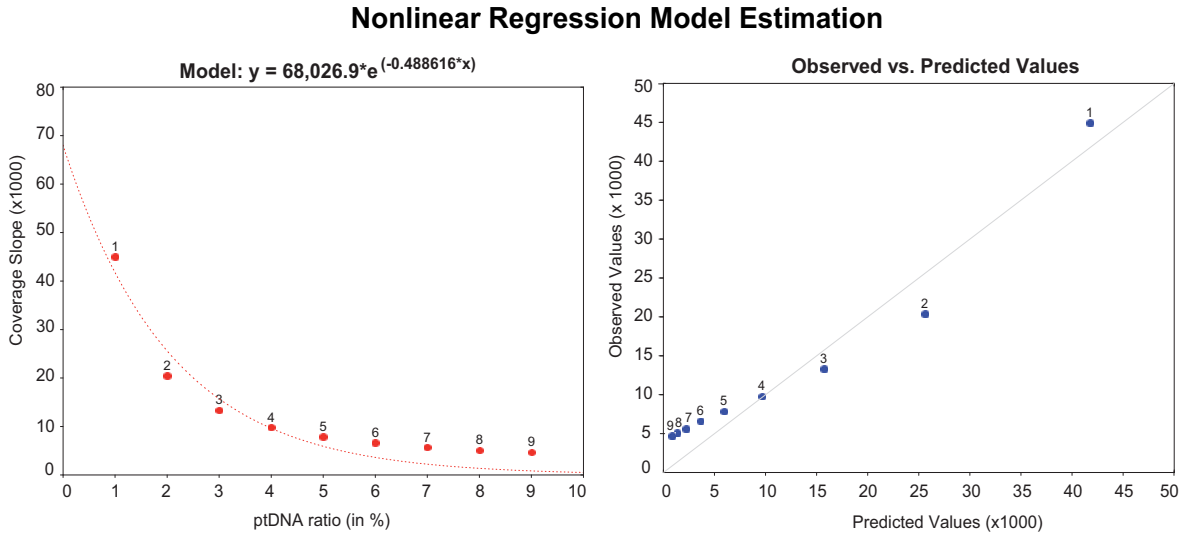


Fig. III-14 **Results from nonlinear regression model estimation.** The left hand side of the figure illustrates the curve of the inferred exponential decay function with the corresponding *Arabidopsis* data points. The right hand side illustrates the insignificant deviation of *Arabidopsis* values from those predicted by the model. Data point annotations refer to the amount of ptDNA.

Note that this estimate is merely meant as a rule of thumb that may help to set the boundaries for the required sequencing or assembly effort. There are certainly other factors that could not be examined here but are likely to define success and costs of any shotgun sequencing project. Also note that certain parameters of sample preparation procedures may not have been considered in the process of 454 simulation of the *Arabidopsis* data during *in-silico* pyrosequencing.

4. CONCLUSIONS AND OUTLOOK

Using a total genomic shotgun-pyrosequencing approach and subsequent *in silico* extraction of a desired genomic region provides an ideal tool for studying molecular evolution in organisms that cannot be examined by standard procedures. Using a resampling scheme and the plastid chromosome as reference genomic region, this study demonstrates for the first time the need for an *a priori* estimation of optimal read pool sizes for the assembly of genomic subsequences. Counter-intuitively perhaps, our results imply that use of more than the necessary amount of reads does not necessarily improve overall assembly quality. We were able to show that the average contig length reaches a local optimum depending on the abundance of the desired genomic fragment in the read pool. Even more, within the range of 20-30x coverage, both simulated datasets as well as experimentally generated datasets from four very different species exhibited least or shortest gaps, and a low risk for putative chimeric contigs. Larger read pools rather negatively affect the reconstruction of a specific desired genomic region from shotgun 454 sequencing projects. In particular, assembly of suboptimal sequence pools bears the risk to produce suboptimal contig length of the desired genome region. We could further show that in case of experimental data where many genomic peculiarities are unknown, both an adequate read pool size as well the employed assembly method severely influences the success of reconstruction. A read presorting strategy, for instance, saves immense computational effort and time during locus specific assembly of large sequence pools. Smaller read pools require of course fewer resources (computation time, RAM etc.). Even if comparatively large read pools are required because of an expected low plastid DNA content, clustering reads prior to the assembly process reduces the computational effort up to 80 %. Besides such savings, clustering reads prior to the assembly turned out extremely useful for difficult plant genomes by reducing the amount of erroneous assembly by incorporating plastid-like sequences from either the nuclear and/or mitochondrial genome into the plastome-supercontig.

One of the major focuses of this study was to elaborate an approximation of the optimal amount of sequences for locus specific assembly of genomic regions from 454-data. Often, such estimations do require many unknown parameters. Based upon CAP3-results, we employed the simple correlation of read depth development and abundance of the desired genomic region in the 454-sequence dataset to approximate the optimal read pool (n) for assembly by an exponential decay model *a priori* as $n \approx (70{,}000\ e^{(-0.5X)})C$, with X being the assumed abundance of plastid DNA; *C* represents the desired coverage. Abundance of a desired genomic region is a parameter that can well be experimentally determined using, for instance PCR and/or blotting approaches. This could contribute to

the ease and success of a planned project and, in return, help reducing costs. Clearly, we must consider that more factors than those investigated here may affect the assembly success in a positive or negative way. Our experimental data suggests robustness up to a certain degree of complexity and base composition of the genomic data. Nevertheless, it is highly recommended to confine the read pool *in silico* rather than the amount of overall sequenced base pairs. The analyses of the holoparasitic species *Phelipanche* that is highly distinctive in terms of molecular evolution suggest that target regions particularly rich in (large) repeated elements essentially require further assembly optimization procedures.

Further studies focusing on NGS technologies others than pyrosequencing are required. Amount and average length of read data varies greatly between the different platforms. This poses the need for technology specific "guidelines" in order to assure a resilient data foundation of biological inferences. Paired-end *Illumina* sequencing is another highly interesting option for sequencing of un-enriched genomic DNA-extracts. Unpaired *Illumina*-runs have already been used in conjunction with reference-based (mapping) assemblies (Zhang et al. 2011) for the reconstruction of the entire plastid chromosome sequences. However, their applicability to de novo reconstruction of genomic segments has yet to be addressed. Studies trying to do so would greatly contribute as "guidance" not only for experimentalists. What is at least as important, results from such studies may provide the much-needed feedback for programmers developing software for NGS data assembly.

5. ACKNOWLEDGEMENTS

We would like to thank Norman J. Wickett (PennState University) for invaluable discussion on bioinformatics. We are especially greatful to Michael Krug (University of Bonn) and Ben Stöver (University of Muenster) for excellent technical help regarding hardware maintenance, and Holger Angenent (University of Muenster) for assistance with the ZIVSMP computing cluster at the University of Muenster.

This project received funding from the Austrian Science Fund (FWF grant 19404 to G.M.S), and the US National Science Foundation (N.S.F. grants DEB-0120709 and DBI-0701748 to C.W.D.).

6. AUTHORS' CONTRIBUTIONS

S.W. conceived of the study, designed the computational steps and contributed to Perl-programming, performed all steps for experimental data generation and plastome reconstruction, analyzed the data and drafted the manuscript. K.F.M. contributed to the conceptual layout of this study, programmed the Perl-pipeline, critically discussed data analysis and revised the manuscript. G.M.S., D.Q. and C.W.D. contributed to the conceptual layout and critically revised the manuscript. M.L. performed plastid enrichment and analyzed the respective data; M.L., P.S.S. and D.E.S critically revised the manuscript.

This chapter will be published in a modified version as research article in a peer-reviewed journal, tentatively as:

Wicke S, Latvis M, Schneeweiss GM, Quandt D, dePamphilis CW, Soltis PS, Soltis DE, and Müller KF. Assembly and reconstruction of specific genomic segments using whole genome shotgun pyrosequencing.

7. REFERENCES

Cattolico R et al. 2008. Chloroplast genome sequencing analysis of *Heterosigma akashiwo* CCMP452 (West Atlantic) and NIES293 (West Pacific) strains. BMC Genomics. 9:211.

Chevreux B. 2011. Sequence assembly with MIRA3 - The definitive guide. http://mira-assembler.sourceforge.net/docs/DefinitiveGuideToMIRA.pdf.

Chevreux B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics. 99:45-56.

Dean FB, Nelson JR, Giesler TL, Lasken RS. 2001. Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed rolling circle amplification. Genome Res. 11:1095-1099.

Dolezel J, Bartos J, Voglmayr H, Greilhuber J. 2003. Nuclear DNA content and genome size of trout and human. Cytometry. 51:127-8; author reply 129.

Downie SR, Palmer JD. 1992. Restriction site mapping of the chloroplast DNA inverted repeat - a molecular phylogeny of the Asteridae. Ann. Mo. Bot. Gard. 79:266-283.

Forrest LL, Wickett NJ, Cox CJ, Goffinet B. 2011. Deep sequencing of *Ptilidium* (Ptilidiaceae) suggests evolutionary stasis in liverwort plastid genome structure. Plant Ecol. Evol. 144:29-43.

Gao L, Su Y-J, Wang T. 2010. Plastid genome sequencing, comparative genomics, and phylogenomics: Current status and prospects. J. Syst. Evol. 48:77-93.

Guisinger MM, Kuehl Jennifer V., Boore Jeffrey L., Jansen RK. 2011. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Mol. Biol. Evol. 28:583-600.

Hawkins TL, Detter JC, Richardson PM. 2002. Whole genome amplification -- applications and advances. Curr. Opin. Biotechnol. 13:65-67.

Huang X, Madan A. 1999. CAP3: A DNA Sequence assembly program. Genome Res. 9:868-877.

Jansen RK et al. 2005. Methods for obtaining and analyzing whole chloroplast genome sequences. Methods Enzymol. 395:348 - 384.

Leaver CJ, Gray MW. 1982. Mitochondrial genome organization and expression in higher plants. Ann. Rev. Plant Physiol. 33:373-402.

Maniatis T, Fritsch E, Sambrook J. 1982. Molecular Cloning. A Laboratory Manual.

McNeal JR et al. 2006. Using partial genomic fosmid libraries for sequencing complete organellar genomes. Biotechniques. 41:69 - 73.

Meader S, Hillier LW, Locke D, Ponting CP, Lunter G. 2010. Genome assembly quality: Assessment and improvement using the neutral indel model. Genome Res. 20:675-684.

Moore M et al. 2006. Rapid and accurate pyrosequencing of angiosperm plastid genomes. BMC Plant Biol. 6:17.

Narzisi G, Mishra B. 2011. Comparing de novo genome assembly: The long and short of it. PLoS ONE. 6:e19175.

Oliver M et al. 2010. Chloroplast genome sequence of the moss *Tortula ruralis*: gene content, polymorphism, and structural arrangement relative to other green plant chloroplast genomes. BMC Genomics. 11:143.

Pascoe MJ, Ingle J. 1978. Distinction between nuclear satellite DNAs and chloroplast DNA in higher plants. Plant Physiol. 62:975-977.

Raubeson LA et al. 2007. Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. BMC Genomics. 8:174.

Richter DC, Ott F, Auch AF, Schmid R, Huson DH. 2008. MetaSim—A sequencing simulator for genomics and metagenomics. PLoS ONE. 3:e3373.

Sandberg J, Stahl PL, Ahmadian A, Bjursell MK, Lundeberg J. 2009. Flow cytometry for enrichment and titration in massively parallel DNA sequencing. Nucl. Acids Res. 37:e63.

Schwartz DC, Waterman MS. 2010. New generations: Sequencing machines and their computational challenges. J. Comp. Science Tech. 25:3-9.

Truernit E, Hibberd JM. 2007. Immunogenic tagging of chloroplasts allows their isolation from defined cell types. Plant J. 50:926-932.

Wakasugi T et al. 1997. Complete nucleotide sequence of the chloroplast genome from the green alga *Chlorella vulgaris*: the existence of genes possibly involved in chloroplast division. Proc. Natl. Acad. Sci. USA. 94:5967 - 5972.

Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D. 2011. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol. Biol. 76:273-297.

Wolf PG et al. 2005. The first complete chloroplast genome sequence of a lycophyte, *Huperzia lucidula* (Lycopodiaceae). Gene. 350:117 - 128.

Zhang Y-J, Ma P-F, Li D-Z. 2011. High-throughput sequencing of six bamboo chloroplast genomes: Phylogenetic implications for temperate woody bamboos (Poaceae: Bambusoideae). PLoS ONE. 6:e20596.

8. SUPPLEMENTAL MATERIAL

8.1. Figures

8.2. Tables

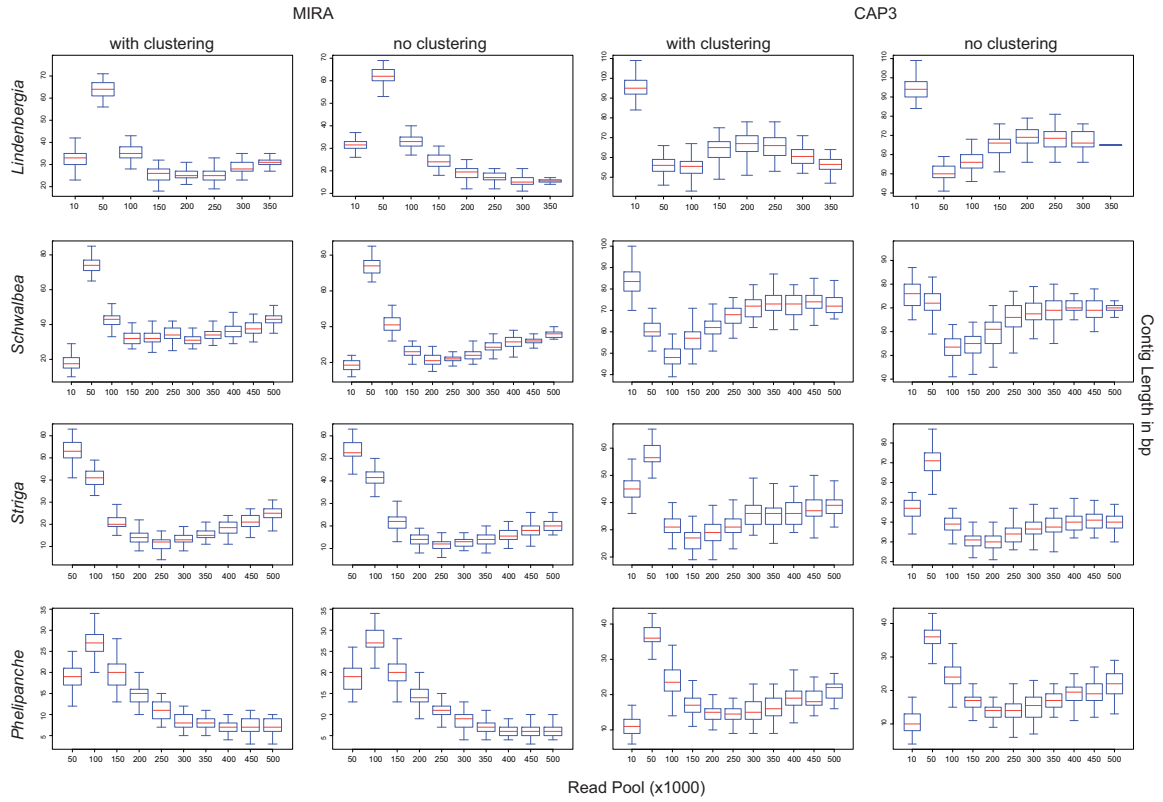


Figure SIII-1 **Number of gaps in empirical datasets.** Box-Whisker plots summarize the number of gaps and/or unknown regions in the plastome occurring per CAP3-or MIRA-assembly with and without previous clustering for *Lindenbergia, Schwalbea, Striga* and *Phelipanche.*

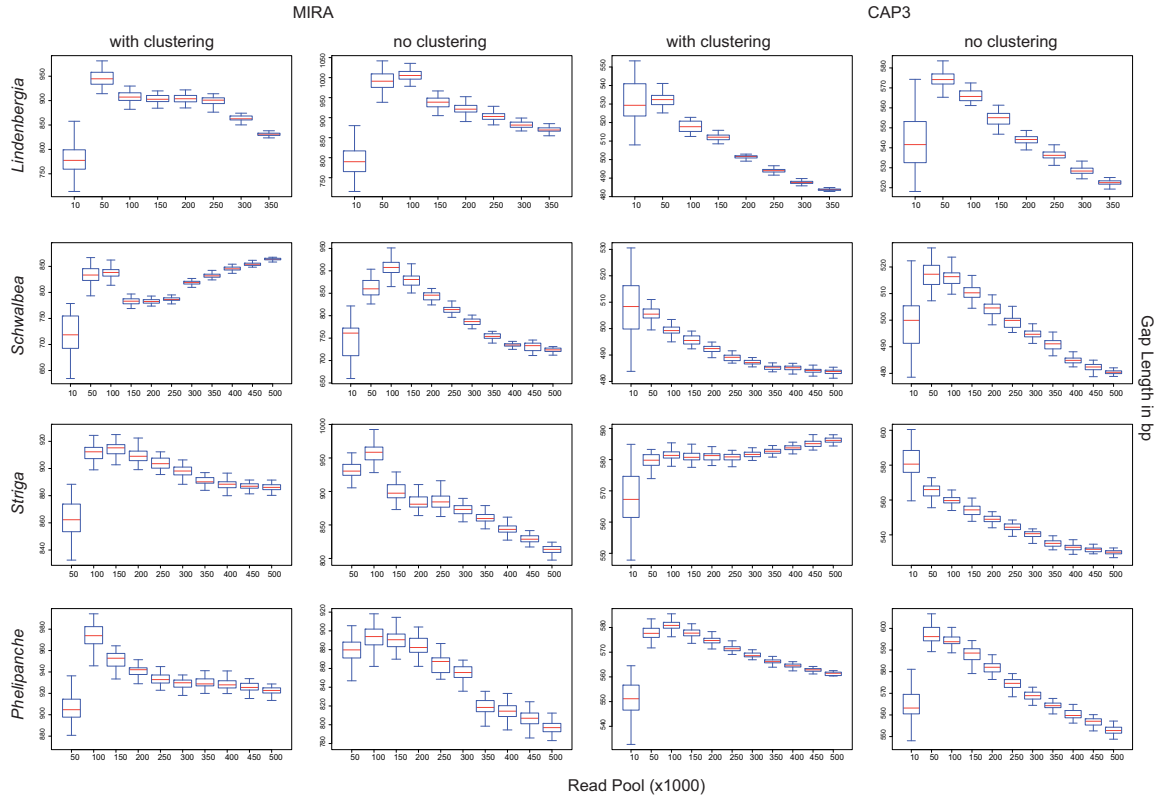


Figure SIII-2 **Average length of gaps/unknown regions.** Box-Whisker plots illustrate the lengths of gaps and/or unknown regions in the plastome occurring on average per CAP3-or MIRA-assembly with and without previous clustering for *Lindenbergia, Schwalbea, Striga* and *Phelipanche.*

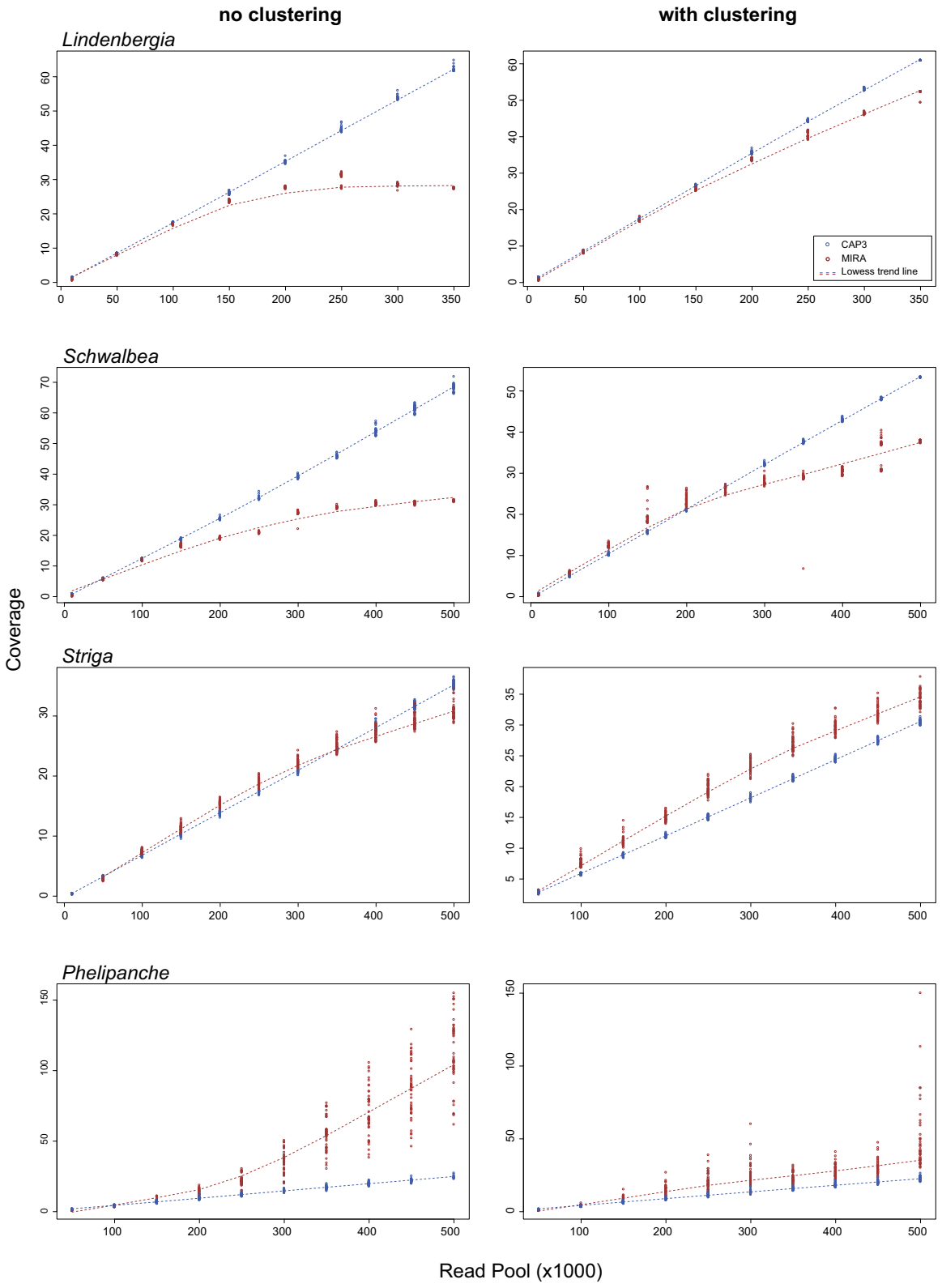


Fig. SIII-3 **Coverage in empirical datasets.** Plastid genome coverage increases steadily with read number in empirical datasets. Linear dependence of coverage and read number in CAP3-assemblies (blue) is shown by a least squares line (LSL) through the iteration median of each read number per taxon. Brown dots/lines illustrate results from MIRA-assemblies.

Table SIII-A **Summary of assembly results of plastid-enriched *Lindenbergia***. Results regarding general assembly statistics as well as plastid specific statistics have been averaged over 50 samples for a read pool size of 10,000 sequences of plastid enriched *Lindenbergia*. Results are sorted according to the employed assembler and assembly strategy. Standard deviation (SE) is provided in brackets.

	No. contigs	No. singlets	Contig length (bp)	No. ptDNA-contigs	ptDNA-contig length (bp)	Base quality	No. gaps	Gap length (bp)
CAP3 with previous read sorting								
Mean	219	821	1,154	163	1,526	60.29	54	660
(SE)	(2)	(4)	(7)	(2)	(19)	(0.24)	(1)	(23)
CAP3 without previous read sorting								
Mean	231	854	1,115	170	1,470	60.11	52	612
(SE)	(4)	(6)	(13)	(2)	(27)	(0.45)	(2)	(55)
MIRA with previous read sorting								
Mean	23	856	6,539	23	14,631	21.35	17	2085
(SE)	(2)	(6)	(496)	(2)	(2883)	(0.07)	(2)	(270)
MIRA without previous read sorting								
Mean	23	903	6527	17	10,142	21.36	35	2447
(SE)	(2)	(8)	(702)	(2)	(1823)	(0.09)	(4)	(150)

Table SIII-B **Contig length differences for CAP3-assemblies with and without previous clustering.** The medians of CAP3-contig lengths obtained without ("w/o", first row per read number) and with previous read clustering ("CPC", second row per read number) were compared with each other. Length differences (in bp) are summarized below. For both assembly strategies, Mann-Whitney-U tests were performed to evaluate statistically significant differences in length between the various read numbers per taxon. Asterisks indicate significance levels (<0.05*, <0.01**).
[Abbreviations: Lind. – *Lindenbergia*; Schw. – *Schwalbea*; Pheli. – *Phelipanche*; k – ×1000; NA – no data]

		Read Pool Size																			
		50k		**100k**		**150k**		**200k**		**250k**		**300k**		**350k**		**400k**		**450k**		**500k**	
	Taxon	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**
10k	***Lind.***	3	33**	-12**	24**	-17**	13**	-28**	3**	-35**	-5**	-42**	-13**	-46**	-19**	NA	NA	NA	NA	NA	NA
	Schw.	-3	17**	-9**	16**	-13**	10**	-16**	5**	-19**	0	-21**	-5*	-23**	-9**	-23**	-15**	-24**	-18**	-25**	-20**
	Striga	13**	-15**	14**	-21**	13**	-26**	14**	-32**	14**	-36**	14**	-40**	15**	-45**	16**	-48**	18**	-49**	19**	-50**
	Pheli.	26**	33**	30**	31**	27**	25**	24**	19**	20**	11**	17**	6**	15**	1**	13**	-3**	12**	-6**	10**	-10**
50k	***Lind.***			-15**	-8**	-20**	-19**	-31**	-30**	-38**	-38**	-45**	-46**	-49**	-52**	NA	NA	NA	NA	NA	NA
	Schw.			-6**	-1	-10**	-7**	-13**	-13**	-16**	-17**	-18**	-23**	-20**	-26**	-20**	-32**	-21**	-35**	-22**	-37**
	Striga			2**	-6**	1*	-12**	1**	-17**	1*	-22**	2**	-25**	3**	-31**	4**	-33**	5**	-34**	6**	-36**
	Pheli.			3**	-2**	0	-8**	-3**	-14**	-6**	-22**	-9**	-27**	-11**	-32**	-13**	-36**	-15**	-39**	-16**	-43**
100k	***Lind.***					-6**	-11**	-16**	-22**	-24**	-29**	-30**	-37**	-34**	-43**	NA	NA	NA	NA	NA	NA
	Schw.					-4**	-6**	-7**	-12**	-10**	-16**	-12**	-22**	-14**	-25**	-14**	-31**	-15**	-34**	-16**	-36**
	Striga					-1	-5**	0	-11**	0	-15**	0	-19**	1**	-25**	2**	-27**	4**	-28**	5**	-30**
	Pheli.					-3**	-5**	-6**	-12**	-9**	-19**	-12**	-25**	-15**	-30**	-16**	-34**	-18**	-37**	-19**	-41**
150k	***Lind.***							-10**	-11**	-18**	-19**	-25**	-27**	-28**	-33**	NA	NA	NA	NA	NA	NA
	Schw.							-3**	-6**	-6**	-10**	-8**	-16**	-10**	-19**	-10**	-25**	-11**	-28**	-12**	-30**
	Striga							1	-5**	0	-10**	1*	-14**	2**	-19**	3**	-21**	4**	-23**	5**	-24**
	Pheli.							-3**	-7**	-6**	-14**	-9**	-20**	-12**	-24**	-13**	-29**	-15**	-32**	-16**	-36**
200k	***Lind.***									-8**	-8**	-14**	-16**	-18**	-22**	NA	NA	NA	NA	NA	NA
	Schw.									-3**	-5**	-5**	-10**	-7**	-13**	-7**	-20**	-8**	-22**	-9**	-24**
	Striga									0	-5**	0	-8**	1*	-14**	2**	-16**	4**	-17**	5**	-19**
	Pheli.									-3**	-7**	-6**	-13**	-9**	-18**	-10**	-22**	-12**	-25**	-13**	-29**
250k	***Lind.***											-6**	-8**	-10**	-14**	NA	NA	NA	NA	NA	NA
	Schw.											-2**	-5**	-4**	-9**	-4**	-15**	-5**	-18**	-5**	-19**
	Striga											1	-4**	2*	-9**	3**	-11**	4**	-13**	5**	-14**
	Pheli.											-3**	-6**	-5**	-10**	-7**	-15**	-9**	-18**	-10**	-22**
300k	***Lind.***													-4**	-6**	NA	NA	NA	NA	NA	NA
	Schw.													-2**	-4**	-2**	-10**	-3**	-12**	-3**	-14**
	Striga													1**	-6**	2**	-8**	3**	-9**	4**	-11**
	Pheli.													-2**	-5**	-4**	-9**	-6**	-12**	-7**	-16**
350k	***Lind.***															NA	NA	NA	NA	NA	NA
	Schw.															0	-6**	-1**	-9**	-2**	-11**
	Striga															1**	-2**	2**	-3**	3**	-5**
	Pheli.															-2**	-5**	-3**	-7**	-5**	-11**
400k	***Lind.***																	NA	NA	NA	NA
	Schw.																	-1**	-2**	-1**	-4**
	Striga																	1	-1**	2**	-3**
	Pheli.																	-2**	-3**	-3**	-7**
450k	***Lind.***																			NA	NA
	Schw.																			0*	-2**
	Striga																			1**	-2**
	Pheli.																			-1**	-4**

Table SIII-C **Contig length differences for MIRA-assemblies with and without previous clustering.** The medians of MIRA-contig lengths obtained without clustering ("w/o", first row per read number) and with previous read clustering ("CPC", second row per read number) were compared with each other. Length differences (in bp) are summarized below. Asterisks indicate significance levels from Mann Whitney-U tests for differences between two read pool sizes per dataset (<0.05*, <0.01**). Unlike CAP3, MIRA-assemblies from 10.000 reads did not yield any plastid contigs for *Striga* and *Phelipanche*, and thus those read pool size was omitted here. [Abbreviations: Lind. – *Lindenbergia*; Schw. – *Schwalbea*; Pheli. – *Phelipanche*; k – ×1000; NA – no data]

		Read Pool Size																			
		50k		**100k**		**150k**		**200k**		**250k**		**300k**		**350k**		**400k**		**450k**		**500k**	
	Taxon	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**
10k	***Lind.***	167**	201**	130**	215**	125**	149**	126**	131**	123**	113**	85**	91**	54**	79**	NA	NA	NA	NA	NA	NA
	Schw.	115**	99**	120**	146**	65**	120**	64**	85**	68**	52**	100**	26**	113**	-7**	127**	26*	136**	-28**	145**	-37**
	Striga	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	Pheli.	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
50k	***Lind.***			-37**	14**	-42**	-52**	-41**	-70**	-44**	-88**	-82**	-110**	-113**	-122**	NA	NA	NA	NA	NA	NA
	Schw.			5	48**	-50**	21**	-51**	-14**	-47**	-46**	-14**	-73**	-2	-106**	12**	-73**	21**	-127**	30**	-136**
	Striga			50**	28**	53**	-33**	47**	-49**	41**	-46**	36**	-57**	28**	-71**	26**	-57**	25**	-102**	24**	-117**
	Pheli.			69**	14**	48**	11**	37**	3	28**	-12**	25**	-24**	24**	-61**	23**	-24**	21**	-73**	18**	-83**
100k	***Lind.***					-4	-67**	-3	-84**	-6**	-102**	-45**	-124**	-76**	-136**	NA	NA	NA	NA	NA	NA
	Schw.					-55**	-27**	-56**	-62**	-52**	-94**	-19**	-121**	-7**	-154**	7**	-121**	16**	-174**	25**	-184**
	Striga					3*	-61**	-3*	-78**	-9**	-74**	-14**	-85**	-22**	-99**	-24**	-85**	-25**	-130**	-26**	-145**
	Pheli.					-21**	-3	-32**	-12**	-41**	-27**	-44**	-38**	-45**	-76**	-46**	-38**	-48**	-87**	-51**	-97**
150k	***Lind.***							1	-18**	-2**	-36**	-40**	-57**	-71**	-70**	NA	NA	NA	NA	NA	NA
	Schw.							-1	-35**	3**	-67**	36**	-94**	49**	-127**	62**	-94**	71**	-147**	81**	-157**
	Striga							-6**	-16**	-12**	-13**	-17**	-24**	-25**	-38**	-27**	-24**	-28**	-69**	-29**	-84**
	Pheli.							-11**	-8**	-20**	-23**	-23**	-35**	-24**	-72**	-25**	-35**	-27**	-84**	-30**	-94**
200k	***Lind.***									-3**	-18**	-41**	-40**	-73**	-52**	NA	NA	NA	NA	NA	NA
	Schw.									4**	-32**	36**	-59**	49**	-92**	63**	-59**	72**	-112**	81**	-122**
	Striga									-5**	3	-11**	-8**	-19**	-21**	-21**	-8**	-22**	-52**	-23**	-67**
	Pheli.									-9**	-15**	-12**	-27**	-13**	-64**	-14**	-27**	-17**	-75**	-19**	-85**
250k	***Lind.***											-38**	-21**	-70**	-34**	NA	NA	NA	NA	NA	NA
	Schw.											32**	-27**	45**	-60**	59**	-27**	68**	-80**	77**	-90**
	Striga											-5**	-11**	-13**	-25**	-15**	-11**	-17**	-56**	-17**	-71**
	Pheli.											-3**	-12**	-4**	-49**	-5**	-12**	-7**	-60**	-10**	-70**
300k	***Lind.***													-31**	-12**	NA	NA	NA	NA	NA	NA
	Schw.													13**	-33**	26**	28**	35**	-54**	45**	-63**
	Striga													-8**	-14**	-10**	0	-11**	-44**	-12**	-60**
	Pheli.													-1	-37**	-2	0	-4**	-49**	-7**	-59**
350k	***Lind.***															NA	NA	NA	NA	NA	NA
	Schw.															14**	33**	22**	-20**	32**	-30**
	Striga															-2**	14**	-3**	-31**	-4**	-46**
	Pheli.															-1	37**	-3**	-11**	-6**	-21**
400k	***Lind.***																	NA	NA	NA	NA
	Schw.																	9**	-54*	18**	-63**
	Striga																	-2**	-44**	-2**	-60**
	Pheli.																	-2**	-49**	-5**	-59**
450k	***Lind.***																			NA	NA
	Schw.																			10**	-9**
	Striga																			-1	-15**
	Pheli.																			-3**	-10**

Table SIII-D **Plastid contig length differences for CAP3-assemblies with and without previous clustering.** The medians of true ptDNA -contig lengths of CAP3 assemblies obtained without clustering ("w/o", first row per read number) and with previous read clustering ("CPC", second row per read number) were compared with each other. Length differences (in bp) are summarized below.

Asterisks indicate significance levels from Mann Whitney-U tests for differences between two read pool sizes per dataset (<0.05*, <0.01**).

[Abbreviations: Lind. – *Lindenbergia*; Schw. – *Schwalbea*; Pheli. – *Phelipanche*; k – ×1000; NA – no data]

		Read Pool Size																			
		50k		**100k**		**150k**		**200k**		**250k**		**300k**		**350k**		**400k**		**450k**		**500k**	
	Taxon	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**
10k	***Lind.***	211**	278**	220**	219**	181**	191**	145**	164**	123**	136**	102**	116**	95**	106**	NA	NA	NA	NA	NA	NA
	Schw.	255**	208**	319**	277**	309**	275**	297**	255**	285**	241**	274**	224**	265**	211**	272**	197**	271**	187**	265**	176**
	Striga	300**	178**	457**	318**	488**	350**	462**	359**	434**	336**	417**	330**	407**	305**	398**	290**	385**	287**	380**	281**
	Pheli.	234**	199**	443**	448**	562**	591**	574**	644**	595**	693**	597**	708**	566**	688**	560**	699**	547**	664**	555**	639**
50k	***Lind.***			8	-59**	-30**	-87**	-66**	-114**	-89**	-142**	-110**	-162**	-117**	-172**	NA	NA	NA	NA	NA	NA
	Schw.			64**	70**	53**	68**	42**	47**	30**	33**	19**	17**	10	4	17**	-10*	16**	-20**	10*	-32**
	Striga			157**	140**	189**	173**	162**	181**	135**	159**	118**	153**	108**	128**	99**	113**	86**	110**	80**	104**
	Pheli.			209**	250**	328**	392**	341**	445**	361**	494**	364**	509**	332**	489**	326**	501**	313**	465**	321**	440**
100k	***Lind.***					-38**	-28**	-74**	-55**	-97**	-83**	-118**	-103**	-125**	-113**	NA	NA	NA	NA	NA	NA
	Schw.					-11*	-2	-22*	-22**	-34**	-36**	-45**	-53**	-54**	-66**	-47**	-80**	-48**	-90**	-54**	-101**
	Striga					31*	32**	5	41*	-23*	19**	-40**	13*	-50**	-12	-59**	-27**	-72**	-31**	-77**	-36**
	Pheli.					119**	143**	132**	196**	152**	244**	155**	259**	124**	240**	117**	251**	105**	216**	113**	191**
150k	***Lind.***							-36**	-27**	-59**	-55**	-80**	-75**	-87**	-85**	NA	NA	NA	NA	NA	NA
	Schw.							-11	-20**	-24**	-34**	-35**	-51**	-44**	-64**	-37**	-78**	-38**	-88**	-43**	-99**
	Striga							-27	9	-54**	-14	-71**	-20**	-81**	-45**	-90**	-60**	-103**	-63**	-108**	-69**
	Pheli.							13	53**	33	102**	36*	117**	4	97**	-2	108**	-14	73**	-7	48**
200k	***Lind.***									-23**	-28**	-44**	-48**	-51**	-58**	NA	NA	NA	NA	NA	NA
	Schw.									-12**	-14**	-23**	-31**	-32**	-44**	-25**	-58**	-26**	-68**	-32**	-79**
	Striga									-28**	-23**	-45**	-29**	-54**	-54**	-64**	-68**	-77**	-72**	-82**	-78**
	Pheli.									21	49*	23	64**	-8	44**	-15	55**	-27**	20	-19	-5
250k	***Lind.***											-21**	-20**	-28**	-31**	NA	NA	NA	NA	NA	NA
	Schw.											-11	-17**	-20**	-30**	-13**	-44**	-14**	-54**	-20**	-65**
	Striga											-17	-6*	-27**	-31**	-36**	-46**	-49**	-49**	-54**	-55**
	Pheli.											2	15	-29	-5	-35*	7	-48**	-28	-40	-54
300k	***Lind.***													-7**	-10**	NA	NA	NA	NA	NA	NA
	Schw.													-9**	-13**	-2	-27**	-3	-37**	-9	-49**
	Striga													-10	-25**	-19**	-40**	-32**	-43**	-37**	-49**
	Pheli.													-31	-20	-38*	-8	-50**	-43**	-42	-69**
350k	***Lind.***															NA	NA	NA	NA	NA	NA
	Schw.															7**	-14**	6	-24**	0	-35**
	Striga															-9*	-15**	-22**	-18**	-27**	-24**
	Pheli.															-6	11	-19	-24**	-11	-49**
400k	***Lind.***																	NA	NA	NA	NA
	Schw.																	-1	-10**	-7	-21**
	Striga																	-13	-3	-18	-9**
	Pheli.																	-12	-35*	-5	-60**
450k	***Lind.***																			NA	NA
	Schw.																			-6	-12**
	Striga																			-5	-6
	Pheli.																			8	-25

Table SIII-E **Plastid contig length differences for MIRA-assemblies with and without previous clustering.** The medians of true ptDNA-contig lengths of MIRA assemblies obtained without clustering ("w/o", first row per read number) and with previous read clustering ("CPC", second row per read number) were compared with each other. Length differences (in bp) are summarized below.
Asterisks indicate significance levels from Mann Whitney-U tests for differences between two read pool sizes per dataset (<0.05*, <0.01**).
[Abbreviations: Lind. – *Lindenbergia*; Schw. – *Schwalbea*; Pheli. – *Phelipanche*; k – ×1000; NA – no data]

		Read Pool Size																			
		50k		**100k**		**150k**		**200k**		**250k**		**300k**		**350k**		**400k**		**450k**		**500k**	
	Taxon	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**
10k	***Lind.***	615**	606**	601**	1,246**	545**	819**	464**	851**	418**	816**	315**	774**	232**	781**	NA	NA	NA	NA	NA	NA
	Schw.	440**	443**	924**	1,183**	670**	2,108**	726**	2,280**	678**	1,809**	528**	1,472**	492**	1,061**	460**	844**	402**	929**	377**	758**
	Striga	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	Pheli.	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
50k	***Lind.***			-14	640**	-70**	213**	-151**	245**	-197**	210**	-300**	168**	-383**	175**	NA	NA	NA	NA	NA	NA
	Schw.			484**	740**	230**	1,664**	286**	1,837**	237**	1,366**	88**	1,029**	51**	618**	19	401**	-38**	486**	-64**	315**
	Striga			596**	1,077**	926**	762**	941**	905**	930**	1,273**	834**	1,092**	724**	920**	673**	806**	619**	693**	537**	641**
	Pheli.			506**	477**	1,085**	1,426**	1,907**	2,995**	3,147**	4,097**	4,830**	5,203**	7,474**	3,973**	9,018**	5,027**	9,951**	4,703**	12,381**	6,372**
100k	***Lind.***					-56**	-427**	-137**	-395**	-183**	-431**	-286**	-472**	-369**	-465**	NA	NA	NA	NA	NA	NA
	Schw.					-254**	924**	-198**	1,097**	-246**	626**	-396**	288**	-433**	-122**	-464**	-339**	-522**	-254**	-547**	-425**
	Striga					329**	-315**	345**	-172**	334**	196**	238**	15	127**	-157**	77**	-271**	22**	-385**	-59**	-436**
	Pheli.					578**	950**	1,401**	2,518**	2,640**	3,620**	4,323**	4,727**	6,968**	3,496**	8,511**	4,551**	9,445**	4,227**	11,875**	5,896**
150k	***Lind.***							-81**	32*	-127**	-4	-230**	-45**	-313**	-38**	NA	NA	NA	NA	NA	NA
	Schw.							56**	173	7	-299**	-143**	-636**	-179**	-1,046**	-211**	-1,263**	-268**	-1,178**	-294**	-1,349**
	Striga							16	143**	4	511**	-91**	330**	-202**	159**	-253**	44	-307**	-69**	-388**	-121**
	Pheli.							822**	1,569**	2,062**	2,671**	3,745**	3,777**	6,389**	2,546**	7,933**	3,601**	8,866**	3,277**	11,297**	4,946**
200k	***Lind.***									-46**	-35*	-149**	-77**	-232**	-70**	NA	NA	NA	NA	NA	NA
	Schw.									-49**	-471**	-199**	-808**	-235**	-1,219**	-267**	-1,436**	-324**	-1,351**	-350**	-1,522**
	Striga									-11	368**	-107**	187**	-218**	15	-268**	-99**	-323**	-212**	-404**	-264**
	Pheli.									1,240**	1,102**	2,923**	2,208**	5,567**	978**	7,111**	2,032**	8,044**	1,708**	10,475**	3,377**
250k	***Lind.***											-103**	-42**	-186**	-35**	NA	NA	NA	NA	NA	NA
	Schw.											-150**	-337**	-186**	-748**	-218**	-965**	-275**	-880**	-301**	-1,051**
	Striga											-96**	-181**	-207**	-353**	-257**	-467**	-311**	-581**	-393**	-632**
	Pheli.											1,683**	1,106**	4,327**	-124	5,871**	930*	6,804**	606*	9,235**	2,275**
300k	***Lind.***													-83**	7	NA	NA	NA	NA	NA	NA
	Schw.													-36**	-410**	-68**	-628**	-126**	-542**	-151**	-714**
	Striga													-111**	-172**	-161**	-286**	-216**	-399**	-297**	-451**
	Pheli.													2,644**	-1,231**	4,188**	-176	5,121**	-500	7,552**	1,169**
350k	***Lind.***															NA	NA	NA	NA	NA	NA
	Schw.															-32**	-217**	-89**	-132**	-115**	-303**
	Striga															-50	-115**	-105**	-228**	-186**	-279**
	Pheli.															1,544*	1,054	2,477**	730**	4,908**	2,399**
400k	***Lind.***																	NA	NA	NA	NA
	Schw.																	-57**	85*	-83**	-86**
	Striga																	-54**	-113**	-136**	-165**
	Pheli.																	933	-324	3,364**	1,345**
450k	***Lind.***																			NA	NA
	Schw.																			-25**	-171**
	Striga																			-81**	-52**
	Pheli.																			2,430**	1,669**

Table SIII-F **Differences of gaps from CAP3-contig/reference alignments.** The differences between the medians of the amount of alignment gaps of CAP3-contigs with a reference sequence have been compared between different read pools of empirical datasets. Results are summarized according to the applied assembly strategy, i.e. without clustering ("w/o", first row per read number) and with previous read clustering ("CPC", second row per read number).
Asterisks indicate significance levels from Mann Whitney-U tests for differences between two read pool sizes per dataset (<0.05*, <0.01**).
[Abbreviations: Lind. – *Lindenbergia*; Schw. – *Schwalbea*; Pheli. – *Phelipanche*; k – ×1000; NA – no data]

		Read Pool Size																			
		50k		**100k**		**150k**		**200k**		**250k**		**300k**		**350k**		**400k**		**450k**		**500k**	
	Taxon	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**
10k	***Lind.***	-39**	-44**	-40**	-38**	-30**	-28**	-28**	-25**	-29**	-26**	-35**	-28**	-39**	-29**	NA	NA	NA	NA	NA	NA
	Schw.	-24**	-4**	-36**	-23**	-27**	-21**	-22**	-15**	-16**	-10**	-12**	-9**	-11**	-7**	-11**	-6**	-10**	-7**	-12**	-6**
	Striga	12**	24**	-14**	-8**	-18**	-16**	-16**	-17**	-14**	-13**	-9**	-11**	-9**	-10**	-9**	-7**	-8**	-6**	-6**	-7**
	Pheli.	25**	26**	13**	14**	6**	7**	4**	4**	4**	4**	4**	6**	5**	7**	8**	10**	7**	9**	11**	12**
50k	***Lind.***			-1	6**	9**	16**	11**	19**	10**	19**	5**	16**	1	15**	NA	NA	NA	NA	NA	NA
	Schw.			-12**	-19**	-3**	-17**	2	-11**	8**	-6**	12**	-5**	13**	-3**	13**	-2	14**	-3*	12**	-2*
	Striga			-26**	-32**	-30**	-40**	-28**	-41**	-26**	-37**	-21**	-35**	-21**	-34**	-21**	-31**	-20**	-30**	-18**	-31**
	Pheli.			-13**	-12**	-19**	-19**	-21**	-22**	-22**	-22**	-21**	-21**	-20**	-19**	-17**	-17**	-18**	-17**	-14**	-14**
100k	***Lind.***					10**	10**	12**	13**	11**	13**	5**	10**	1	9**	NA	NA	NA	NA	NA	NA
	Schw.					9**	2	14**	8	20**	13**	24**	14**	25**	16**	25**	17**	26**	16**	24**	17**
	Striga					-4**	-8**	-2*	-9**	0	-5**	5**	-3*	5**	-2	5**	1	6**	2*	8**	1
	Pheli.					-7**	-7**	-9**	-10**	-9**	-10**	-9**	-9**	-8**	-7**	-5**	-5**	-6**	-5**	-2	-2**
150k	***Lind.***							2	3**	1	3*	-5**	0	-9**	-1	NA	NA	NA	NA	NA	NA
	Schw.							5**	6**	11**	11**	15**	13**	16**	14**	16**	15**	17**	14**	15**	15**
	Striga							2**	-1	4**	3**	9**	6**	9**	7**	9**	9**	10**	10**	12**	9**
	Pheli.							-2**	-3**	-3**	-3**	-2**	-2*	-1*	0	2*	3**	1*	2**	5**	5**
200k	***Lind.***									-1	-1*	-7**	-3	-11**	-4**	NA	NA	NA	NA	NA	NA
	Schw.									6**	5**	10**	7**	11**	8**	11**	9**	12**	8**	10**	9**
	Striga									2*	4**	7**	7**	7**	8**	7**	10**	8**	11**	10**	10**
	Pheli.									-1	0	0	2**	1*	3**	4**	6**	3**	5**	7**	8**
250k	***Lind.***											-6**	-3	-10**	-4**	NA	NA	NA	NA	NA	NA
	Schw.											4**	2	5**	3	5**	4**	6**	3**	4**	4**
	Striga											5**	3*	5**	4**	5**	6**	6**	7**	8**	6**
	Pheli.											1	2	2*	3**	5**	6**	4**	5**	8**	8**
300k	***Lind.***													-4**	-1*	NA	NA	NA	NA	NA	NA
	Schw.													1	2	1	3**	2*	2	0	3**
	Striga													0	1	0	4**	1	5**	3**	4**
	Pheli.													1	2*	4**	4**	3**	4**	7**	7**
350k	***Lind.***															NA	NA	NA	NA	NA	NA
	Schw.															0	1	1	0	-1	1
	Striga															0	3	1*	4*	3**	3*
	Pheli.															3**	3**	2**	2**	6**	5**
400k	***Lind.***																	NA	NA	NA	NA
	Schw.																	1	-1	-1	0
	Striga																	1	1	3**	0
	Pheli.																	-1	-1	3*	3**
450k	***Lind.***																			NA	NA
	Schw.																			-2	1
	Striga																			2	-1
	Pheli.																			4*	3**

Table SIII-G **Differences of gaps from MIRA-contig/reference alignments.** The differences between the medians of the amount of alignment gaps of MIRA-contigs with a reference sequence have been compared between different read pools of empirical datasets. Results are summarized according to the applied assembly strategy, i.e. without clustering ("w/o", first row per read number) and with previous read clustering ("CPC", second row per read number). Asterisks indicate significance levels from Mann Whitney-U tests for differences between two read pool sizes per dataset (<0.05*, <0.01**). Unlike CAP3, MIRA-assemblies from 10,000 reads did not yield any plastid contigs for *Striga* and *Phelipanche*.
[Abbreviations: Lind. – *Lindenbergia*; Schw. – *Schwalbea*; Pheli. – *Phelipanche*; k – ×1000; NA – no data]

		Read Pool Size																			
		50k		**100k**		**150k**		**200k**		**250k**		**300k**		**350k**		**400k**		**450k**		**500k**	
	Taxon	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**
10k	***Lind.***	31**	31**	2**	2**	-7**	-8**	-8**	-12**	-8**	-15**	-5**	-17**	-2**	-16**	NA	NA	NA	NA	NA	NA
	Schw.	57**	56**	26**	23**	15**	8**	15**	3**	17**	4**	14**	6**	17**	10**	19**	13**	20**	14**	26**	18**
	Striga	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	Pheli.	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
50k	***Lind.***			-29**	-29**	-38**	-38**	-39**	-43**	-39**	-45**	-36**	-47**	-33**	-47**	NA	NA	NA	NA	NA	NA
	Schw.			-31**	-33**	-42**	-48**	-42**	-53**	-40**	-52**	-43**	-50**	-40**	-46**	-38**	-43**	-37**	-42**	-31**	-38**
	Striga			-12**	-11**	-33**	-31**	-39**	-39**	-41**	-41**	-40**	-40**	-38**	-39**	-35**	-37**	-32**	-35**	-28**	-33**
	Pheli.			8**	8**	1	1*	-4**	-5**	-8**	-8**	-11**	-10**	-11**	-12**	-12**	-13**	-12**	-13**	-12**	-13**
100k	***Lind.***					-9**	-9**	-10**	-14**	-10**	-16**	-7**	-18**	-4**	-18**	NA	NA	NA	NA	NA	NA
	Schw.					-11**	-15**	-11**	-20**	-9**	-19**	-12**	-17**	-9**	-13**	-7**	-10**	-6**	-9**	0	-5**
	Striga					-21**	-20**	-27**	-28**	-29**	-30**	-28**	-29**	-26**	-28**	-23**	-26**	-20**	-24**	-16**	-22**
	Pheli.					-7**	-7**	-12**	-13**	-16**	-16**	-19**	-18**	-19**	-20**	-20**	-21**	-20**	-21**	-20**	-21**
150k	***Lind.***							-1	-5**	-1	-7**	2**	-9**	5**	-9**	NA	NA	NA	NA	NA	NA
	Schw.							0	-5**	2*	-4**	-1	-2**	2*	3**	4**	6**	6**	6**	11**	10**
	Striga							-6**	-8**	-8**	-10**	-7**	-9**	-5**	-8**	-2**	-7**	1	-4**	5**	-2**
	Pheli.							-5	-6**	-9**	-9**	-12**	-11**	-12**	-13**	-13**	-14**	-13**	-14**	-13**	-14**
200k	***Lind.***									0	-3**	3**	-5**	6**	-4**	NA	NA	NA	NA	NA	NA
	Schw.									2**	1	-1	3**	2**	8**	4**	11**	6**	11**	11**	15**
	Striga									-2**	-2**	-1	-1	1*	0	5**	2**	7**	4**	11**	6**
	Pheli.									-4**	-3**	-7**	-5**	-7**	-7**	-8**	-8**	-8**	-8**	-8**	-8**
250k	***Lind.***											3**	-2**	6**	-2**	NA	NA	NA	NA	NA	NA
	Schw.											-3**	2**	0	7**	2*	10**	4**	10**	9**	14**
	Striga											1	1*	3**	2**	7**	4**	9**	6**	13**	8**
	Pheli.											-3	-2**	-3**	-4**	-4**	-5**	-4**	-5**	-4**	-5**
300k	***Lind.***													3**	1	NA	NA	NA	NA	NA	NA
	Schw.													3**	5**	5**	8**	7**	8**	12**	12**
	Striga													2**	1*	6**	3**	8**	5**	12**	7**
	Pheli.													0	-2**	-1**	-3**	-1*	-3**	-1**	-3**
350k	***Lind.***															NA	NA	NA	NA	NA	NA
	Schw.															2**	3**	4**	4**	9**	8**
	Striga															4**	2**	6**	4**	10**	6**
	Pheli.															-1*	-1*	-1*	-1*	-1*	-1*
400k	***Lind.***																	NA	NA	NA	NA
	Schw.																	2	1	7**	5**
	Striga																	3**	3**	7**	5**
	Pheli.																	0	0	0	0
450k	***Lind.***																			NA	NA
	Schw.																			6**	4**
	Striga																			4**	2**
	Pheli.																			0	0

Table SIII-H **Differences of gap lengths from CAP3-contig/reference alignments.** The differences between the medians of the length of alignment gaps of CAP3-contigs with a reference sequence have been compared between different read pools of empirical datasets. Results are summarized according to the applied assembly strategy, i.e. without clustering ("w/o", first row per read number) and with previous read clustering ("CPC", second row per read number).
Asterisks indicate significance levels from Mann Whitney-U tests for differences between two read pool sizes per dataset (<0.05*, <0.01**).
[Abbreviations: Lind. – *Lindenbergia*; Schw. – *Schwalbea*; Pheli. – *Phelipanche*; k – ×1000; NA – no data]

		Read Pool Size																			
		50k		**100k**		**150k**		**200k**		**250k**		**300k**		**350k**		**400k**		**450k**		**500k**	
	Taxon	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**
10k	***Lind.***	-302**	-232**	-209**	-219**	-279**	-249**	-331**	-288**	-349**	-313**	-378**	-331**	-380**	-335**	NA	NA	NA	NA	NA	NA
	Schw.	-529**	-779**	-348**	-691**	-365**	-636**	-393**	-664**	-443**	-696**	-464**	-708**	-500**	-739**	-506**	-764**	-527**	-766**	-528**	-776**
	Striga	-1,557**	-1,496**	-1,522**	-1,542**	-1,394**	-1,449**	-1,358**	-1,411**	-1,422**	-1,443**	-1,454**	-1,416**	-1,453**	-1,467**	-1,491**	-1,490**	-1,479**	-1,496**	-1,509**	-1,493**
	Pheli.	-3,856**	-4,159**	-4,078**	-4,357**	-3,989**	-4,255**	-3,792**	-4,129**	-3,663**	-3,938**	-3,635**	-3,903**	-3,634**	-3,921**	-3,667**	-3,990**	-3,663**	-3,991**	-3,729**	-4,063**
50k	***Lind.***			93**	14	23	-16	-29*	-56**	-47**	-80**	-76**	-99**	-78**	-103**	NA	NA	NA	NA	NA	NA
	Schw.			181**	88**	164**	144**	136**	115**	86**	84**	65**	71**	29**	40**	22**	15**	2	14**	1	3
	Striga			36*	-46**	163**	47**	200**	85**	135**	53**	104**	80**	104**	29**	67**	6	78**	0	49**	3
	Pheli.			-222**	-198**	-134**	-96**	64	30	193**	221**	221**	256**	222**	238**	189**	169**	193**	168**	127**	96**
100k	***Lind.***					-70**	-30**	-122**	-70**	-140**	-94**	-169**	-113**	-171**	-117**	NA	NA	NA	NA	NA	NA
	Schw.					-17	55**	-45**	27	-95**	-5	-116**	-17**	-152**	-48**	-159**	-73**	-179**	-75**	-180**	-85**
	Striga					128**	93**	164**	131**	100**	99**	68**	126**	69**	75**	31*	52**	42*	46**	13	49**
	Pheli.					88**	102**	286**	228**	414**	418**	442**	454**	444**	436**	411**	366**	415**	366**	349**	294**
150k	***Lind.***							-52**	-40**	-70**	-64**	-99**	-83**	-101**	-87**	NA	NA	NA	NA	NA	NA
	Schw.							-27**	-28**	-78**	-60**	-99**	-73**	-135**	-104**	-141**	-129**	-161**	-130**	-163**	-141**
	Striga							37	38**	-28	6*	-60**	33**	-59**	-18	-97**	-41	-85**	-47**	-115**	-44**
	Pheli.							198**	126**	326**	317**	354**	352**	356**	334**	323**	265**	327**	264**	261**	192**
200k	***Lind.***									-18**	-24**	-47**	-43**	-49**	-47**	NA	NA	NA	NA	NA	NA
	Schw.									-50**	-32**	-71**	-44**	-107**	-75**	-114**	-100**	-134**	-101**	-136**	-112**
	Striga									-65**	-32	-96**	-5	-96**	-56**	-133**	-78**	-122**	-85**	-151**	-81**
	Pheli.									128**	190**	156**	225**	158**	208**	125**	138**	129**	138**	63	66*
250k	***Lind.***											-29**	-19*	-31**	-23**	NA	NA	NA	NA	NA	NA
	Schw.											-21**	-12**	-57**	-44**	-63**	-69**	-84**	-70**	-85**	-81**
	Striga											-32*	27	-31*	-24**	-69**	-46**	-57**	-53**	-87**	-50**
	Pheli.											28	35	30	18	-3	-52	0	-52	-65	-124**
300k	***Lind.***													-2	-4*	NA	NA	NA	NA	NA	NA
	Schw.													-36**	-31**	-42**	-56**	-63**	-57**	-64**	-68**
	Striga													1	-51**	-37*	-74**	-26*	-80**	-55**	-77**
	Pheli.													2	-18	-31	-87**	-28	-88**	-93	-160**
350k	***Lind.***															NA	NA	NA	NA	NA	NA
	Schw.															-6	-25**	-27**	-26**	-28**	-37**
	Striga															-38	-23	-26	-29**	-56**	-26**
	Pheli.															-33	-70	-29	-70	-95	-142**
400k	***Lind.***																	NA	NA	NA	NA
	Schw.																	-20**	-1	-22**	-12**
	Striga																	11	-6	-18*	-3
	Pheli.																	4	0	-62	-72
450k	***Lind.***																			NA	NA
	Schw.																			-2	-11**
	Striga																			-29*	3
	Pheli.																			-66	-72

Table SIII-I **Differences of gap lengths from MIRA-contig/reference alignments.** The differences between the medians of the length of alignment gaps of CAP3-contigs with a reference sequence have been compared between different read pools of empirical datasets. Results are summarized according to the applied assembly strategy, i.e. without clustering ("w/o", first row per read number) and with previous read clustering ("CPC", second row per read number). Asterisks indicate significance levels from Mann Whitney-U tests for differences between two read pool sizes per dataset (<0.05*, <0.01**). Unlike CAP3, MIRA-assemblies from 10.000 reads did not yield any plastid contigs for *Striga* and *Phelipanche*.
[Abbreviations: Lind. – *Lindenbergia*; Schw. – *Schwalbea*; Pheli. – *Phelipanche*; k – ×1000; NA – no data]

		Read Pool Size																			
		50k		**100k**		**150k**		**200k**		**250k**		**300k**		**350k**		**400k**		**450k**		**500k**	
	Taxon	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**	**w/o**	**CPC**
10k	***Lind.***	-2,664**	-2,813**	-2,753**	-2,932**	-2,713**	-2,843**	-2,733**	-2,821**	-2,682**	-2,713**	-2,630**	-2,631**	-2,631**	-2,544**	NA	NA	NA	NA	NA	NA
	Schw.	-5,074**	-4,833**	-5,197**	-4,944**	-5,011**	-4,218**	-4,792**	-4,090**	-4,696**	-3,645**	-4,909**	-3,526**	-4,914**	-3,838**	-4,927**	-4,037**	-4,993**	-3,832**	-4,980**	-3,998**
	Striga	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	Pheli.	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
50k	***Lind.***			-89**	-119**	-48**	-30*	-69**	-8	-18	100**	34**	182**	34**	269**	NA	NA	NA	NA	NA	NA
	Schw.			-124**	-112**	62**	614**	282**	743**	378**	1,188**	164**	1,307**	160**	995**	147**	796**	81**	1,000**	93**	835**
	Striga			-582**	-520**	-556**	-685**	-316**	-373**	-53	41	-35	207**	-19	-16	-134**	-144**	-81**	-141**	-167**	-181**
	Pheli.			-1,402**	-1,420**	-1,536**	-1,554**	-1,410**	-1,306**	-882**	-736**	886**	441*	2,559**	538*	3,443**	1,040**	3,994**	1,045**	3,661**	1,472**
100k	***Lind.***					41*	89**	20*	110**	71**	219**	123**	301**	122**	388**	NA	NA	NA	NA	NA	NA
	Schw.					186**	726**	406**	854**	501**	1,300**	288**	1,418**	284**	1,106**	270**	907**	204**	1,112**	217**	946**
	Striga					25	-166**	265**	147**	528**	561**	547**	727**	562**	503**	448**	376**	501**	378**	414**	339**
	Pheli.					-133**	-134**	-8	113	520**	684**	2,288**	1,861**	3,961**	1,958**	4,846**	2,460**	5,397**	2,465**	5,064**	2,892**
150k	***Lind.***							-20	21	31*	130**	83**	212**	82**	299**	NA	NA	NA	NA	NA	NA
	Schw.							220**	128*	315**	574**	102**	692**	98**	380**	84**	181**	18	386**	31	221**
	Striga							240**	313**	503**	726**	521**	893**	537**	669**	422**	542**	475**	544**	389**	504**
	Pheli.							125**	247**	653**	818**	2,421**	1,995**	4,094**	2,092**	4,979**	2,594**	5,530**	2,599**	5,197**	3,026**
200k	***Lind.***									51**	108**	103**	191**	102**	277**	NA	NA	NA	NA	NA	NA
	Schw.									96**	446**	-118**	564**	-122**	252**	-135**	53	-201**	258**	-189**	92*
	Striga									263**	414**	281**	580**	297**	356**	182**	229**	235**	231**	149*	192**
	Pheli.									528**	570**	2,296**	1,747**	3,969**	1,845**	4,854**	2,347**	5,405**	2,352**	5,072**	2,779**
250k	***Lind.***											52**	82*	51**	169**	NA	NA	NA	NA	NA	NA
	Schw.											-213**	119	-218**	-194*	-231**	-392**	-297**	-188**	-284**	-353**
	Striga											18	166	34	-57	-80*	-185*	-27	-183*	-114*	-222**
	Pheli.											1,768**	1,177**	3,441**	1,274**	4,326**	1,776**	4,877**	1,782**	4,544**	2,209**
300k	***Lind.***													-1	87**	NA	NA	NA	NA	NA	NA
	Schw.													-4	-312**	-17	-511**	-84**	-306**	-71**	-472**
	Striga													16	-224*	-99*	-351**	-46	-349**	-132**	-388**
	Pheli.													1,673**	97	2,558**	599*	3,109**	604**	2,776**	1,031**
350k	***Lind.***															NA	NA	NA	NA	NA	NA
	Schw.															-13	-199**	-79**	6	-67**	-160**
	Striga															-115**	-127*	-62*	-125*	-148**	-165**
	Pheli.															885**	502	1,436**	507*	1,103**	934**
400k	***Lind.***																	NA	NA	NA	NA
	Schw.																	-66**	205**	-53**	39
	Striga																	53	2	-33	-37
	Pheli.																	551	5	218	432
450k	***Lind.***																			NA	NA
	Schw.																			13	-165**
	Striga																			-86	-39
	Pheli.																			-333	427

– CHAPTER IV –

BROOMRAPE PLASTID GENOMES REVEAL DISTINCT PATTERNS OF FUNCTIONAL AND PHYSICAL GENE DELETION UNDER RELAXED SELECTIVE CONSTRAINTS

ABSTRACT. Non-photosynthetic plants possess highly reduced plastid chromosomes compared to their autotrophic relatives. Caused by the loss of photosynthesis, deletions of photosynthesis as well as housekeeping protein-coding genes display prominent convergences in several unrelated lineages. However, little is known about *how* functional and structural genome reduction takes place. In the present study, we trace the complex history of genome reduction in a group of closely related parasites of the broomrape family (Orobanchaceae). This group represents a wide array of intermediates in the process of plastome reduction that allows assessing elementary patterns of the deletion of dispensable DNA-fragments. To this end, we sequenced the plastid genomes from several photosynthetic and non-photosynthetic parasitic broomrape species using different sequencing strategies. We thoroughly analyzed the structural evolution with respect to co-linearity, gene content, and functionality of genes. By reconstruction of plastome rearrangement history and ancestral gene contents, we assessed molecular evolutionary patterns of gene loss and the deletions of DNA-segments under relaxed selective constraints. We provide convincing evidence that functional plastome reduction already occurs in early stages of heterotrophy suggesting that the establishment of obligate parasitism can be viewed as the major prior relaxing selective constraints. Increasing amounts of plastid repetitive DNA in parasites eventually entail increased rates of improper and/or illegitimate recombination leading to the eventual deletion of plastid-chromosomal fragments. Amongst others, our analyses reveal that the functional and physical plastome reduction coincides with a measurable increase in A/T-content in both coding and non-coding plastid DNA-fractions. Furthermore, we demonstrate that vicinity to functionally relevant groups mainly determines longevity and retention of dispensable genes and gene regions.

KEYWORDS. Reductive genome evolution, plastid genome reconfiguration, gene loss, pseudogenization, nucleotide compositional bias, plastid repetitive DNA, relaxed selective pressure, parasitic plants.

CONTENTS.

This chapter contains approx. 16,000 words, 11 figures, 6 tables, plus 20 pages of supplemental information.

1. INTRODUCTION

The plastid chromosome encodes numerous subunits involved in photosynthesis reactions, but also several proteins for the genetic apparatus including structural RNAs. In most plants, the plastid chromosomes exhibits a quadripartite structure characterized by two single copy regions (LSC - large single copy region; small single copy region - SSC), which are separated by two virtually identical large inverted repeat regions (IR). Due to a constantly high selective pressure on the functionality of photosynthesis and plastid housekeeping genes, the plastid chromosome in green plants evolves highly conservatively in terms of structure and nucleotide substitution rates. Apart from these, plastomes exhibit a rather small number of small dispersed (SDR) and simple sequence repeats (SSR). Moreover, repeats larger than 50 bp appear to be commonly suppressed in conservatively evolving plastid chromosomes (e.g. Raubeson et al. 2007; Wicke et al. 2011). Only few autotrophic lineages are known for their non-canonical plastid chromosome structure (e.g. Cosner et al. 2004; Jansen et al. 2007; Cai et al. 2008; Guisinger et al. 2011; reviewed in Wicke et al. 2011). An important role in stabilizing the plant plastid genome may be the mostly yet not exclusive uniparental inheritance (e.g. Bock 2007; Zhang and Sodmergen 2010) as well as effective recombination and DNA-repair processes (reviewed in Maréchal and Brisson 2010). These processes are probably associated with the presence of the two large inverted chromosomal segments (IRs). Prominent changes in plastid chromosome structure have been shown to be mostly correlated with an increase of both SDRs and SSRs (Chumley et al. 2006; Haberle et al. 2008; Cai et al. 2008; Guisinger et al. 2011). Moreover, breakpoints of inversions and other rearrangements are often flanked by tRNA genes that may provide anchors for illegitimate recombination and/or improper repair (Hiratsuka et al. 1989; Maréchal and Brisson 2010; Guisinger et al. 2011). Structural changes in the plastid genome due to gene loss, including gene loss after functional gene transfer to the nucleus, occur rarely among photosynthetic land plants. Nevertheless, few gene deletions were observed in highly rearranged plastomes (Jansen et al. 2007; Magee et al. 2010; reviewed in Wicke et al. 2011).

A group of plants whose plastid chromosome evolution is mainly driven by gene loss is non-photosynthetic plants. Non-photosynthetic representatives, i.e. holoparasitic plants, have completed the transition from a semi-autotrophic to a fully heterotrophic way of life, and thus completely rely on a host for nutritional and water supply. Thus, having abandoned photosynthesis, plastids of non-photosynthetic plants are deprived of their major function. In contrast, hemiparasites retain the ability to carry out photosynthesis to a greater or lesser extent. Holoparasites are well known to exhibit unusual modes of plastid genome evolution due to extensive functional reduction eventually leading to the deletion of large plastome regions (reviewed in Wicke et al. 2011; Krause 2011). Dramatic physical and functional diminution has been reported for several (holo-)heterotrophic plants (Wolfe

et al. 1992a; Funk et al. 2007; McNeal et al. 2007; Delannoy et al. 2011; Logacheva et al. 2011). Gene losses are primarily observed in photosynthesis relevant genes (elements for photosystems, electron transport etc.), but also some genes of the genetic apparatus have been reported lost or are pseudogenized (reviewed in Wicke et al. 2011). The extent of plastome reduction among different groups of parasitic plants, even among close relatives, appears to proceed at different tempos and may attest highly lineage specific evolution. Some lineages of the broomrape family (Orobanchaceae), for instance preserve and even express the *rbc*L, whereas others only retain a pseudogenized copy, or lost it completely (Delavault et al. 1995; Lusson et al. 1998; Wolfe and dePamphilis 1997; Randle and Wolfe 2005; Young and dePamphilis 2005; Wicke et al. 2011). Similarly, the set of plastid-encoded tRNAs differs substantially within and among various lineages of parasitic plants (Wolfe et al. 1992a, b; Funk et al. 2007; McNeal et al. 2007; Wickett et al. 2008; Delannoy et al. 2011; Logacheva et al. 2011). Apart from distortion of gene order due to gene deletions, there are several differences in the structural evolution of the plastid chromosomes in parasites, most prominently affecting the IR-segments with contraction or expansion up to complete deletion of one segment (Downie and Palmer 1992; Delannoy et al. 2011; Funk et al. 2007; McNeal et al. 2007). Besides this, smaller inversions have been reported in the single copy regions (Funk et al. 2007; McNeal et al. 2007). It is tempting to assume that these independent structural changes associate to relaxed evolutionary pressures normally selecting for structurally stable and rather slowly evolving plastomes in order to ascertain unconfined genome functionality. Hence, we would hypothesize that parasitic plant plastomes feature numerous non-canonical elements. Those elements could potentially provide evidence how physical reduction takes place, and what determines genomic changes under relaxed selective constraints. Insights from closely related hemi-and holoparasitic plants are widely lacking, and currently known aspects regarding plastome reduction do not allow concluding patterns of plastid genome reduction under relaxed evolutionary constraints caused by different degrees of heterotrophy. Are gene functions lost "randomly" after the release of selective constraints on the plastid chromosome? What happens to DNA segments that have newly become dispensable due to the loss of their coding function? Are they lost independently and randomly, or under certain constraints? How rapidly does genome reduction take place?

So far, aspects of molecular evolutionary processes of reductive genome evolution under relaxed or lacking selection have not yet been addressed. Therefore, we selected a phylogenetically well-understood group of parasitic angiosperms in order to test specifically hypotheses on the events and series of functional and physical genome reduction. Orobanchaceae are an ideal group to assess evolutionary patterns of plastid genome reduction, gene loss, and pseudogenization among very closely related species. Orobanchaceae are root parasites, i.e. they are capable of connecting to the roots of a host plant via a specialized organ, the haustorium. The group exhibits a diversity of divergent

modes with respect to host dependence. Above that, Orobanchaceae include a fully autotrophic lineage, *Lindenbergia*, which is the sister group to all other species. Some Orobanchaceae are loosely in need of a host plant. Those facultative hemiparasites are capable to autotrophically fulfill their lifecycle, whereas obligate parasites completely rely on a host for germination and reproduction even if they are able to carry out photosynthesis (see e.g. Westwood et al. 2010 for a recent review of parasitism in angiosperms). Complemented by holoparasitic members, Orobanchaceae provide a unique model to broadly and comparatively investigate (molecular) evolutionary questions. In particular, the broomrape-clade itself qualifies as a model group for studying processes of plastid genome reduction (clade III sensu Bennett and Mathews 2006; Schneeweiss et al. 2004; Park et al. 2008), as it is entirely holoparasitic, and thus bears high potential of reflecting various (intermediate) stages of plastome non-functionalization and physical deletions.

In the current study, we employ complete plastid chromosomes to reconstruct patterns of functional and physical genome reduction in an unprecedentedly wide comparative approach. Although reductive evolution is expected to occur in all non-photosynthetic lineages of Orobanchaceae, previous studies suggest that several different levels of plastome reduction occur among very closely related species (e.g. Downie and Palmer 1992; Delavault et al. 1996; Wolfe and dePamphilis 1997; dePamphilis et al. 1997). We will test different hypotheses relating to aspects of functional genome reduction, gene loss and the deletion of newly dispensable genome segments. Here, we: i) examine the structural and functional plastome evolution under relaxed selective constraints as well as the role of repetitive DNA in the process of reductive genome evolution DNA, ii) investigate the series of gene losses and test whether the loss of a plastid DNA-fragment occurs unrestrained after the loss of selective pressures, and iii) investigate the role of transcription units and conserved essential elements during physical genome reduction.

2. RESULTS

1.1 Broomrape plastid chromosomes exhibit a great diversity of both architecture and functional properties.

We sequenced the plastid chromosomes of a representative set of parasitic Orobanchaceae species (photosynthetic as well as non-photosynthetic), plus the autotrophic and first branching lineage *Lindenbergia*. The architecture and properties of broomrape plastomes exhibit an extraordinary size spectrum and coding capacity. Among them, we find non-photosynthetic members with hardly any genome reduction.

Nevertheless, the same group also harbors the smallest plastid chromosome found in land plants so far. The plastomes of the photosynthetic Orobanchaceae representatives *Lindenbergia* and *Schwalbea* are ~156 kb and 161 kb in size, respectively, and thus do not deviate from the majority of other angiosperms. Besides small shifts (<100 bp) regarding the junctions of the IR-region into the large and small single copy region, the *Lindenbergia* plastome is highly similar and co-linear to that of *tobacco*. The hemiparasitic, i.e. photosynthetically active *Schwalbea* exhibits few gene losses as well as localized structural changes: The *accD-rbcL*-region in the plastid LSC is inverted relative to that of tobacco and *Lindenbergia*. Furthermore, the IR-region expands leading to a duplication of the *ycf1*-gene and some subunits of the *ndh*-complex. Several of the *ndh*-genes normally encoded in the SSC region are pseudogenized: *ndhF* is split into three truncated fragments, which are scattered in the IR and SSC; *ndhA* is truncated lacking one exon and its typical intron. Genes for NdhD and NdhG contain several indels and multiple premature stop codons; a start codon is lacking in *ndhG*. Besides this, the *accD*-gene is very likely non-functional as its coding region is 5'-truncated. The remainder putative genic region of *accD* displays extreme sequence divergence including large indels and several premature stop codons. Its status as a pseudogene needs experimental verification (expression analysis), since we cannot rule out the possibility of an alternative start codon as well as RNA-editing of the premature stop codons. Unpredictable differences in size exist in the non-photosynthetic holoparasites. Sizes of holoparasitic plastomes range from 121 kb in *Myzorrhiza* to only 45 kb in *Conopholis* (summarized in Table IV-A). Reflecting this size diversity, the coding capacity, i.e. number of functionally retained genes, of the plastid chromosome varies greatly among (holoparasitic) Orobanchaceae (Table IV-A, Supplemental material: Table SIV-A).

Table IV-A **Overview of physical properties of plastid chromosomes in non-parasitic and parasitic Orobanchaceae.** Besides plastome size, the table presents a summary of the gene content for each species, the proportions of coding to non-coding regions as well as their respective G/C-content.

Taxon	Size (bp)	Gene content (prot.[1]/ tRNA[1]/ rRNA[2])	% prot.-coding regions	% non-coding regions	% struc. RNA	% G/C [3]	% G/C[4] protein coding	% G/C[4] non-coding	% G/C[4] struct. RNA
Lindenbergia	155,088	79/30/4	51.068	41.291	7.641	37.799	38.158	33.422	54.567
Schwalbea	160,944	74/30/4	50.822	41.814	7.364	38.080	38.817	33.732	54.313
Epifagus	70,072	20/16/4	43.071	41.579	15.350	36.007	34.724	29.838	50.201
Conopholis	45,776	20/18/4[5]	49.294	37.863	12.843	33.922	35.662	28.881	52.445
Cistanche	94,387	25/25/4	33.403	54.355	12.242	36.566	36.213	32.658	53.441
Boulardia	80,360	27/21/4	42.118	44.146	13.736	35.755	35.723	29.835	53.819
O. crenata	87,529	31/26/4	39.875	36.916	13.209	35.185	35.374	29.838	53.394
O. gracilis	65,634	28/25/4[5]	49.418	36.085	14.497	34.512	34.758	28.244	52.949
Myzorrhiza	120,996	41/30[6]/4[6]	36.846	53.440	9.714	36.692	37.188	32.174	53.651
P. purpurea	62,897	28/22/4[5,7]	50.145	39.742	10.113	31.087	34.619	24.949	51.659
P. ramosa	61,709	28/22/4[5,7]	43.504	46.428	10.068	32.039	34.911	26.772	52.117

1 – refers to the number of unigenes; 2 – unless mentioned otherwise, rDNA is present in 2 sets; 3 – %G/C of the entire plastid chromosome; 4 - %G/C of unique sequences (IR-region removed where present); 5 – IR loss/loss of one rDNA operon; 6 – one rRNA operon may be non-functional; 7 – partial duplication of rrn16.

Housekeeping genes (*rps-*, *rpl*-genes, structural RNAs) are widely conserved, and only few reading frames are pseudogenized or have been lost in broomrapes. In several cases (*rps7, 8, 12, 18*), unambiguous identification of gene start and stop will require experimental verification; since the remainder sequence implies functionality of the gene, we do not classify those as pseudogenes here (Supplemental Material: Table IV-A). Genes related to photosynthesis are most commonly non-functional or have already been lost from holoparasite plastomes. Pseudogenization and/or eventual gene loss seems to be lineage specific. However, we observe a recurrent pattern regarding the series of losses. Some of the photosynthesis-related genes seem to be less prone to be lost, while others are frequently absent. Frequent loss concerns subunits for the electron transport (*ndh-*, *pet*-genes), and the plastid-encoded polymerase (*rpo*-genes). Several genes for photosystem subunits (*psaA, B, C; psbA, C, K, Z*) are repeatedly found as pseudogenes while others have already been lost from the plastomes (e.g. *psaI, J, K; psbB, D, E, F, I, J, L, M, T*). Retained photosystem-pseudogenes are frequently localized in close vicinity to elements of the genetic apparatus (see sections 1.5. and 1.6. for detailed analyses). Most interestingly, subunits for the ATP-Synthase complex (*atp*-genes) are present in several cases with apparently functional reading frames.

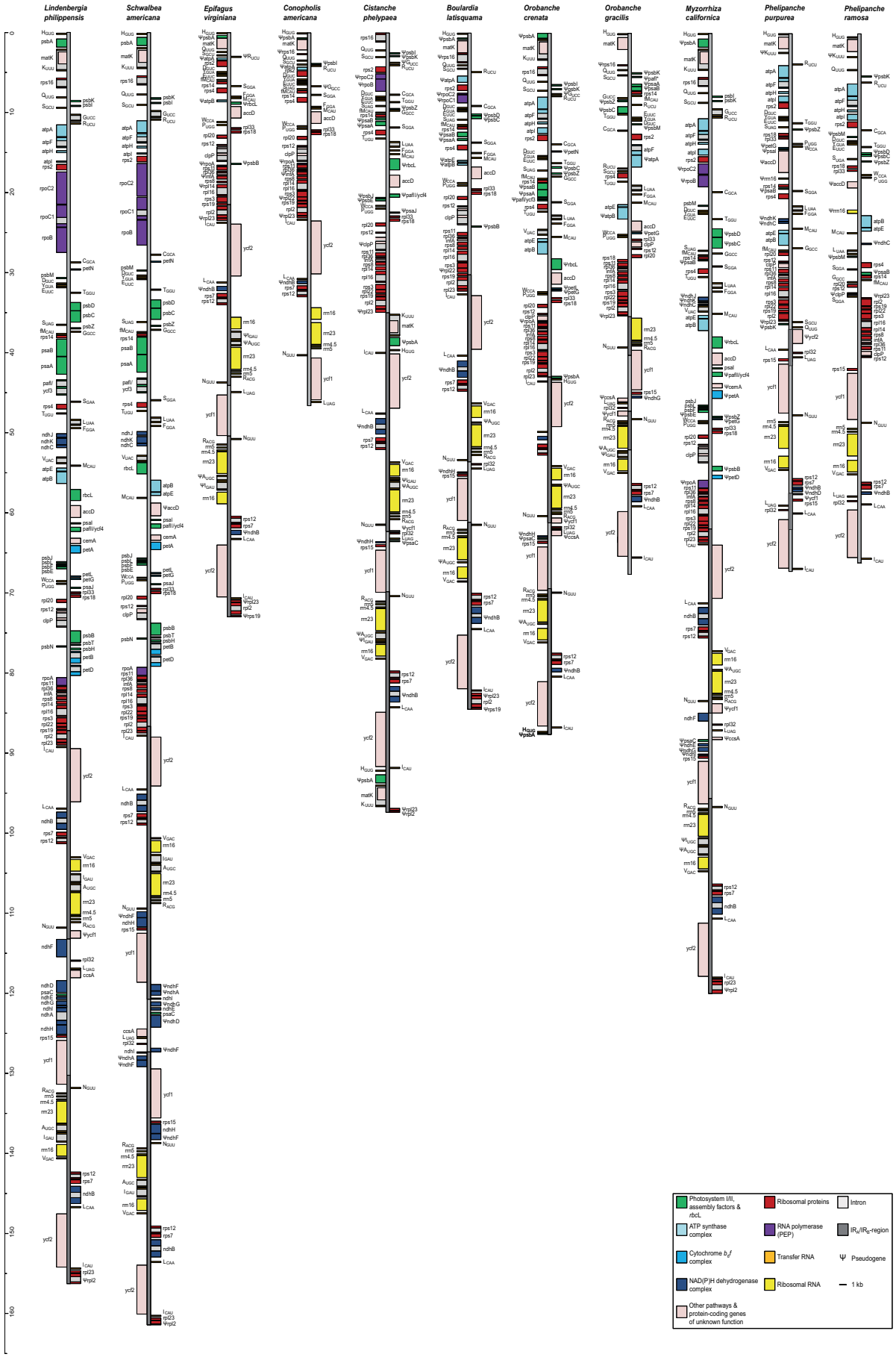
The functionality of the genes *ycf*1 and *ycf*2 in Orobanchaceae remains unclear and requires experimental verification by transcription analysis. While we could unambiguously identify the reading frame for the *ycf*2-gene by PCR-based Sanger re-sequencing, there is evidence for pseudogenization of the *ycf*1-gene in *Phelipanche*. We detected only short fragments with some similarity to *ycf*1-genes. We were not able to determine the gene start unambiguously, and the identifiable reading frame seems to be truncated. Unlike other plastid gene regions, we cannot entirely rule out sequencing errors in the *ycf1*-region by Sanger re-sequencing due to frequent repeats and extreme homopolymeric stretches in the region. Therefore, *ycf1* was excluded from subsequent analyses (Supplemental material: Table SIV-A).

1.2 Gene rearrangements and large-scale structural reconfiguration accompany the process of functional plastid genome reduction in hemi- and holoparasites.

Gene order and gene content are virtually identical among most eudicot plastid chromosomes, including *Lindenbergia*. In parasites, including the hemiparasitic *Schwalbea*, gene deletions cause multiple local gene disruptions (Fig. IV-1). Besides, variations concerning the large IR-segments are another major cause for both plastome size variation and structural differences (e.g. *Conopholis* vs. *Epifagus*). With the exception of *Lindenbergia, Epifagus, Boulardia,* and *Myzorrhiza*, we observe notable extensions and reductions of the large IR-region. Complete loss of one IR-segment occurs in *Conopholis* and *Phelipanche*

ramosa. In many cases, we also observe localized rearrangements as well as inversions in the LSC-region (Figure IV-1); those rearrangements appear to be unrelated to actual events of gene loss as inversion breakpoints do not entail deletions in the respective regions compared to closest relatives. LSC-inversions mostly coincide with modifications of the IR-regions (e.g. *Schwalbea, Cistanche, Orobanche,* and *Phelipanche*). IR expansions occur in several independent lineages. In *Cistanche*, the translocation of the Ψ*psb*A-*trn*K-*mat*K region leads to an IR expansion, and disrupts the *rpl*23 reading frame. In *Schwalbea*, IRs expand relative to *Lindenbergia* due to the repetition of gene fragments of truncated Ψ*ndh*-genes and of *ycf1*. Despite multiple gene losses in the single copy regions, mainly in the LSC, the *Orobanche crenata* plastid chromosome is largely co-linear to *Lindenbergia*. In contrast, *O. gracilis* shows a smaller inversion in the LSC and its IR-segments only include the largest part of the rDNA operon relative to *Lindenbergia*. Gene order is most extensively re-configured in *Phelipanche* (Fig: IV-1). Besides multiple gene losses, the *rpl32-trnL*$_{UAG}$ region (usually located in the SSC) has been duplicated by relocation into the *ycf2-rps7* region in *P. purpurea*. The IR-regions are reduced and do no longer include the rDNA-operon. In all *Phelipanche* species, a huge duplicated fragment of plastid *rrn16* replaces the *rbcL*-gene between *atpB* and *accD*. Long stretches of non-coding DNA (~2.5kb) border the truncated *rrn16*-like fragments. The fragment does not exhibit any significant similarity to known plastid (spacer) DNA-regions. Apart from one group I intron (*trnL*$_{UUA}$), *Phelipanche* plastomes retain only two group I-introns (*rpl16, atpF,*) that do not coincide with the loss of photosynthesis–related genes. We notice that particular group IIA introns are absent from their common host genes (3'-*rps12, clpP, rpl2, trnK*$_{UUU}$/*matK*). With the exception of *clpP*, this type of introns associates with *MatK*, which is likely to play an essential role during the splicing of these plastid introns (Zoschke et al. 2010). In contrast to other Orobanchaceae, the *matK* reading frame in *Phelipanche* is highly divergent, and does contain at least two internal stop codons in *P. ramosa*. Given its unusual evolution with generally higher nucleotide substitution rates and frequent indels, we hesitate, however, to classify it as pseudogene solely based on its primary sequence data. The "concerted loss" of several IIA introns might indicate an impaired splicing mechanism due to either a reduced activity or malfunction of *MatK* or of any of the nuclear-encoded factors. According to a gamma correlation test there is an slight trend towards more localized plastid genome inversions associated to the parasitic lifestyle ($p = 0.183$). The relation of frequent structural changes after the loss of photosynthesis is, however, clearly rejected ($p_{inversions} = 0.525$, $p_{IR} = 0.774$).

Fig. IV-1 **Physical maps of the plastid chromosomes of photosynthetic and non-photosynthetic members of Orobanchaceae.** All genes are colored according to functional complexes. Pseudogenes are demarcated by a "Ψ"-symbol in the gene ID. Genes are colored according to their functionality.


Lindenbergia philippensis
Schwalbea americana
Epifagus virginiana
Conopholis americana
Cistanche phelypaea
Boulardia latisquama
Orobanche crenata
Orobanche gracilis
Myzorrhiza californica
Phelipanche purpurea
Phelipanche ramosa
Photosystem I/II, assembly factors & rbcL
ATP synthase complex
Cytochrome b6f complex
NAD(P)H dehydrogenase complex
Other pathways & protein-coding genes of unknown function
Ribosomal proteins
RNA polymerase (PEP)
Transfer RNA
Ribosomal RNA
Intron
IRA/IRB-region
Ψ Pseudogene
1 kb

1.3 The amount of plastid repetitive DNA increases significantly under relaxed selective constraints and destabilizes the plastid chromosome of parasite.

The amount of repetitive DNA segments is generally low in plastid genomes with the majority of cases being forward or palindromic repeats. Number and size of plastid DNA repeats are considered directly related to structural stability of plastomes (reviewed in Wicke et al. 2011). Conservatively evolving plastomes, i.e. those that show little variation in overall structure across different lineages, show similar patterns of repeat distributions and repeat sizes. Non-canonically evolving plastomes (e.g. in Fabaceae, Geraniaceae) have been shown to be significantly richer in the overall number of repeated elements and contain a considerably higher amount of large repeats (>100 bp). Similarly, structural modifications of the plastome due to ongoing functional reduction/gene loss may likely coincide with an increase of repeated DNA elements. In particular, we may also assume that severe structural changes as observed in *Phelipanche*, *O. gracilis* and *Conopholis* are reflected in facets of plastid repetitive DNA that are different from those of structurally more conservatively evolving broomrape plastomes. In fact, we find that number and overall length of repeated sequences in parasitic plant plastomes are apparently different from close autotrophic relatives. The total number and ratio of repeats does not notably differ between *Nicotiana* and *Lindenbergia*. Including the IR, repetitive DNA accounts for a ~19 % of the total plastid chromosome in each species. Except for the *Epifagus*/*Conopholis* clade and *Myzorrhiza*, we observe a notable increase in the total repeat number in parasites (including the hemiparasitic *Schwalbea*) relative to tobacco and *Lindenbergia* (Fig IV-2). In *Epifagus* and *Conopholis*, the absolute repeat numbers are similar to the autotrophs. Nevertheless, given the smaller plastome size, we notice that the ratio is strongly skewed in these parasites as well. If repeats were distributed uniformly in the plastome,

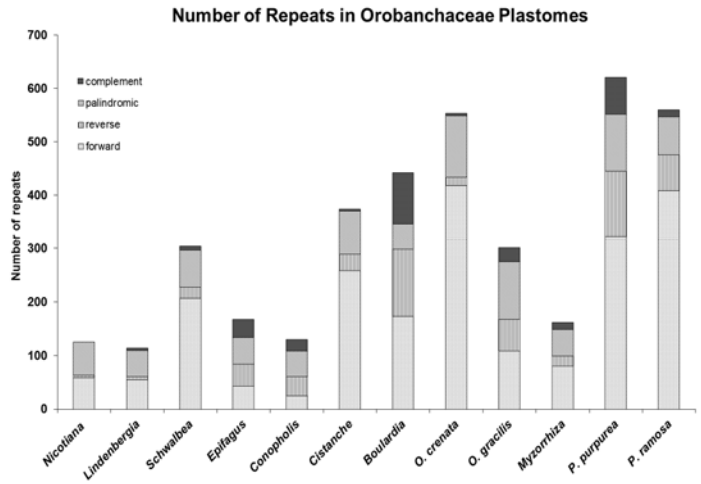


Fig. IV-2 **Number of repetitive DNA elements in Orobanchaceae.** The ratio of different repeat classes to each other is severely skewed in hemi- and holoparasites compared to *Nicotiana* and *Lindenbergia*. Reverse and complement repeats occur more frequently in parasite plastomes than in autotrophs.

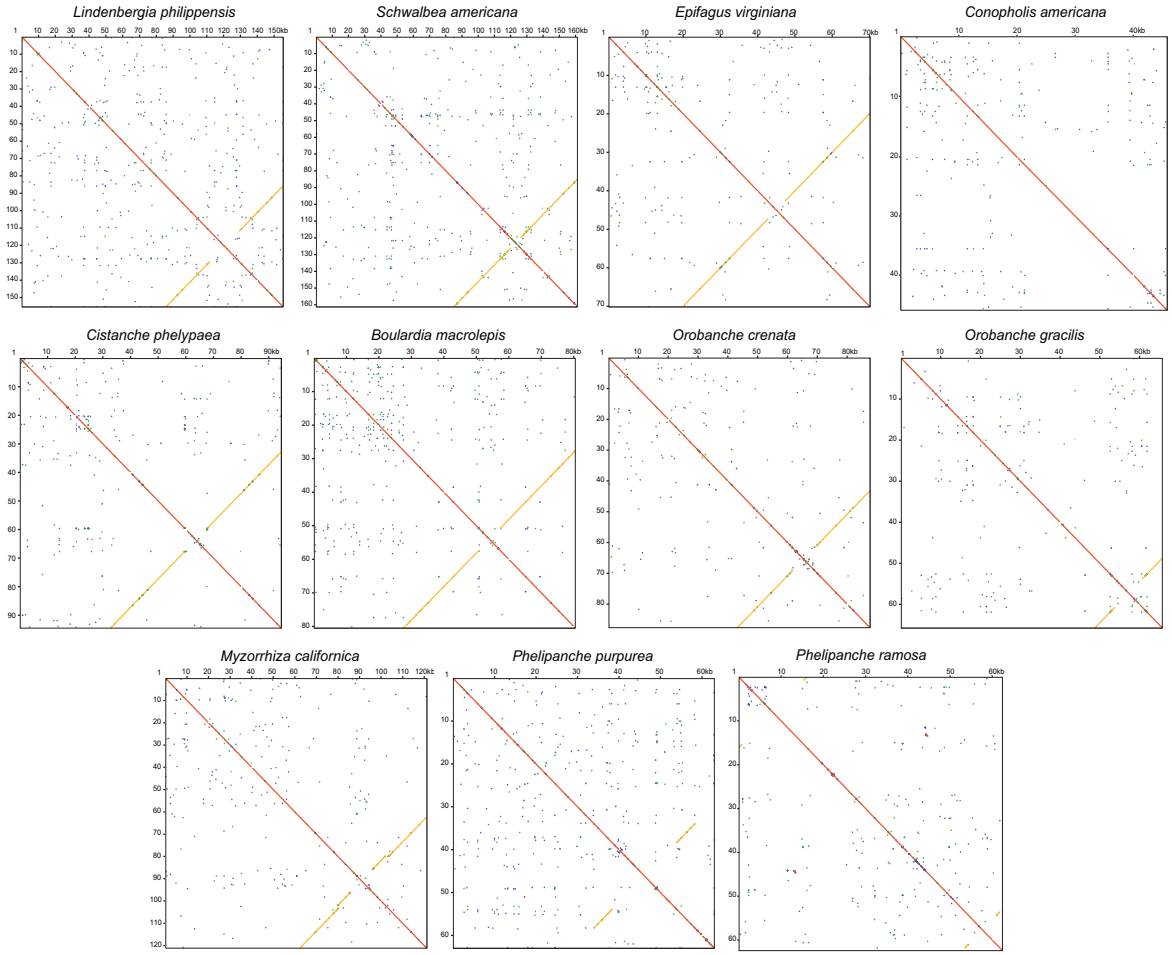


Fig. IV-3 **Self-dotplots of plastid chromosomes from non-parasitic and parasitic Orobanchaceae.** Red/dots lines mark identical matches. Orange dots indicate identical palindromic repeats, with an uninterrupted orange line displaying the large inverted repeat region. Blue and green dots indicate direct and reverse repeats. Axes are scaled in kb.

we would encounter a repeated element every ~1.3kb in *tobacco/Lindenbergia*. In *Epifagus/Conopholis*, however, repeats would appear every ~0.5 kb. In the remainder parasites the frequency would be even higher with up to one repeat per every ~ 150 bp. Taken into account the total length covered by repeats, the *Conopholis* plastome contains only ~ 8 % of its genome as duplicated sequences. In contrast, repeats account for ~35% in *Phelipanche* plastomes and nearly 42% in *O. gracilis*. These amounts do not differ from those found in other parasitic plants, where repeated elements account for 30 to 45% (in Myzorrhiza only 17%) of the plastome. We analyzed the distribution of repeats in broomrape plastomes with the help of self-dotplots. According to the plots (Fig. IV-3), longer repeats are dispersed nearly uniformly in *Lindenbergia*. In contrast, repeated elements are less conservatively distributed across the plastome of parasites. In *Schwalbea*, we see slight accumulations around the IR-SSC junction as well as an accumulation near the middle of the LSC. A similar pattern occurs in *Boulardia* and *Myzorrhiza* as well as in

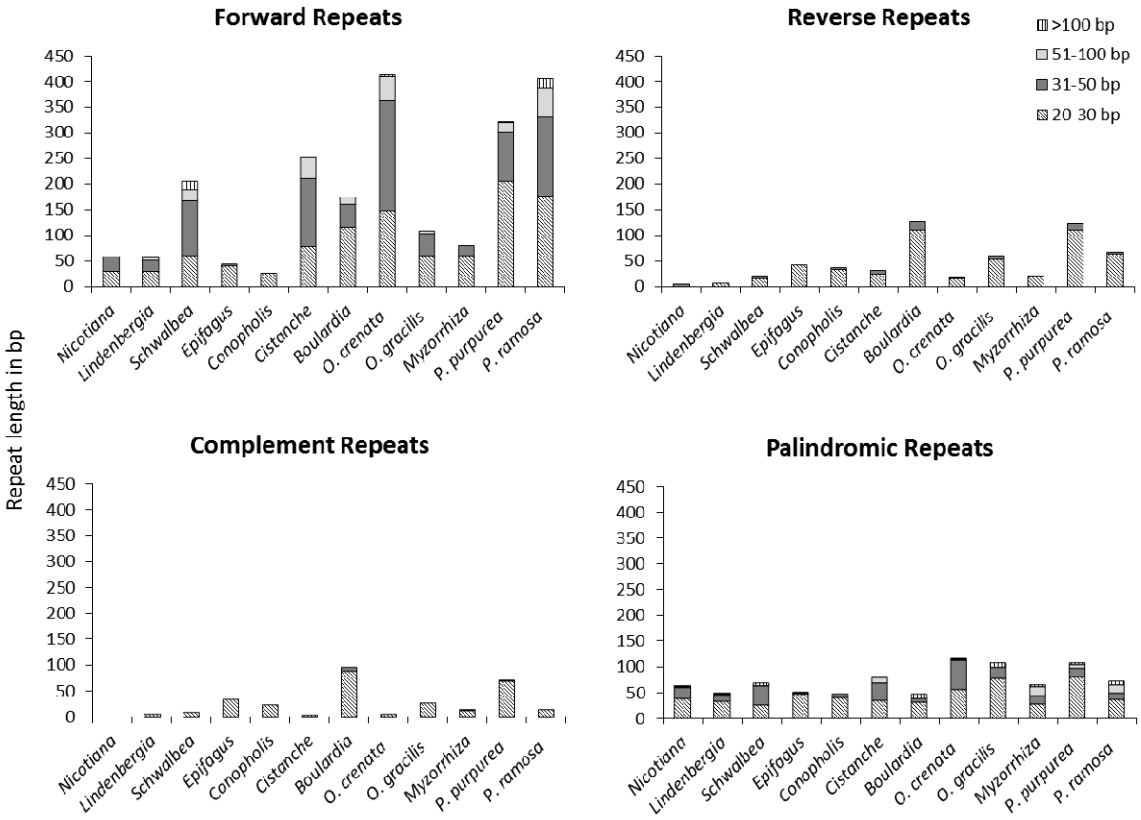


Fig. IV-4 **Length of plastid repeats in photosynthetic and non-photosynthetic Orobanchaceae and close relatives.** Repeats have been subdivided in to four different size classes, i.e. repeats of l20-30, 31-50, 51-100 bp and longer than 100 bp. The ratios of these size classes to one another are sorted according to repeat types (forward, reverse, complement, palindromic).

Orobanche, although less pronounced in the latter. Those local increases around the IR-SSC-boundary may co-locate with a putative origin of replication, which maps to this region in *Nicotiana*. Consequently, the LSC-repeats may accumulate around putative sites of termination from Cairns-type replication, which is known to terminate approximately 180° opposite of the initiation site. Alternatively, those local accumulations might originate from high similarities of genic regions to one another (e.g. tRNA regions). The largest repeat-poor regions correspond to the large *rpl*/*rps*-operon at the LSC-IR-boundary. Interestingly, parasite IR-segments (where present) are poorer in repeated elements than their counterparts in *Lindenbergia* and *Schwalbea*.

Statistical analysis supports a relation between parasitism in Orobanchaceae and an increasing number of plastid repeats ($p = 0.013^*$; Table IV-B), and a significantly increasing repeat density ($p = 0.012^*$). However, the proportion of plastid repetitive DNA in the plastome is not significantly altered with the general transition to heterotrophy ($p = 0.618$). The degree of functional genome reduction does not associate with variation in the

Table IV-B **Correlation analysis of structural plastome changes with parasitism and/or DNA repeats.** All test statistics incl. the p-value resulting from a gamma correlation test are provided for each tested hypothesis. Asterisks mark the significance level (<0.05*, <0.01**). IR=inverted repeat;

Hypothesis	N	Gamma	Statistic Z	p-value
Total number of repeats increases in parasitic Orobanchaceae-plastomes	12	1.00000	2.49136	0.013*
Parasitism correlates with an increase of proportional repeat length	12	0.20000	-0.49827	0.618
Parasitism correlates with an increased repeat density	12	1.00000	-2.49136	0.012*
Functional genome reduction correlates with repeat number	12	0.09091	0.41144	0.680
Parasitism correlates with structural changes (inversions)	12	1.00000	1.33097	0.183
Parasitism correlates with structural changes (IR)	12	1.00000	2.06661	0.039*
Loss of photosynthesis correlates w/structural changes (inversions)	12	-0.14286	-0.28638	0.774
Loss of photosynthesis correlates w/structural changes (IR)	12	-0.26316	-0.63523	0.525
IR-changes relate to repeat number	12	0.48936	1.86896	0.061
Repeat density correlates with structural changes (inversions)	12	-0.62162	-2.10644	0.035*
Repeat density correlates with structural (IR) changes	12	-0.40426	-1.54393	0.122
Repeat density correlates with low total G/C-content	12	0.666667	3.01719	0.003**
Repeat density correlates with low non-coding G/C-content	12	0.630769	2.83302	0.005**

amount of plastid repetitive DNA (p = 0.680). On average, plastid repeats are rarely longer than 100 bp (Fig. IV-4). Duplicated elements of this size class (either in forward orientation or as palindromes) mainly account for the dramatic increase in overall repeat number in broomrape plastomes. However, there seems to be an accumulation in the *Phelipanche*-clade. The majority of such large repeated elements in *Phelipanche* originates from genic regions (e.g. from *rrn16*, Ψ*clpP*). A similar pattern concerning the proliferation of larger plastid repeats (including duplication of gene fragments) has previously been described for some photosynthetic angiosperms that show an aberrant behavior of the large inverted repeat region (reviewed in Wicke et al. 2011). Many broomrape plastome have experienced structural reconfiguration around the IR-region. In broomrape plastomes, these two observations also notably interrelate with each other (p = 0.039*). Moreover, our results show that the evolution of plastid repetitive DNA in parasitic Orobanchaceae strongly associates to the presence/absence of changes in gene synteny caused by inversion events (p = 0.035*). Moreover, repeat density significantly increases in G/C-poor plastomes ($p_{totalGC}$ = 0.003**, $p_{non\text{-}codingGC}$ = 0.005**) suggesting the proliferation of particularly A/T-rich elements or homopolymeric stretches. To some extent, proliferation of plastid repeats also associates with structural changes around the IR-region (expansion, constriction, loss), but the correlation is at best only marginally significant ($p_{Rep.\text{-}No.}$ = 0.061, $p_{Rep.\text{-}density}$ = 0.122).

1.4 Evolutionary patterns of pseudogenization and gene deletion reveal different tempos of functional reduction after the loss of photosynthesis in the broomrape ancestor.

We reconstructed the putative ancestral content of protein and tRNA-genes using maximum likelihood and an unconstrained model allowing for different rates of state

changes. Based upon our data matrix, the transition from a functional plastid gene to a pseudogene was estimated to occur nearly eight times as frequent than the immediate loss of a previously functional gene (estimated state change parameters: $q_{gene \rightarrow \Psi} = 3.422$; $q_{gene \rightarrow loss} = 0.443$, $q_{\Psi \rightarrow loss} = 17.446$). The parameter for back-changes, i.e. re-functionalization or reversing of a pseudogenization or gene loss event, was estimated as $q = 0$. The series of major pseudogenization and gene loss events has been graphically summarized for each node in Fig. IV-5. Detailed results for every plastid protein-coding gene and all tRNAs are presented in the supplemented material in Fig. SIV-1 and SIV-2, respectively.

Based upon our study and a set of initially 112 unique genes (excluding conserved but cryptic ORFs such as *ycf15*, *ycf62*), we identified a minimum common gene set of 20 protein genes that seem to be essential in the broomrape clade. Of those, 17 encode ribosome subunits (13 ribosomal protein genes). Of the non-ribosomal open reading frames only the *matK*-gene, *ycf1* and *ycf2* are retained in all examined taxa; *ycf1* and *2* are likely to be involved in other non-ribosomal processes as well. However, as of this writing, their roles in plastid metabolism are unknown. The genes *accD* and *clpP* may be considered in an expanded set of potentially essential parasite plastid genes. Both subunits are involved in non-photosynthetic pathways and are found frequently retained. However, both genes have been deleted from some of the herein investigated taxa suggesting that, in those cases, their function may be substituted by nuclear/cytosolic subunits. Besides protein-coding genes, the minimum common set of retained structural RNA genes encompasses 14 unique tRNAs plus 4 different rRNA species in broomrape plastomes. Our results imply that several subunits for the -complex appear to have been functionally lost during the transition to hemiparasitism in Orobanchaceae. The "long" retention of *ndhB* compared to other *ndh*-genes likely correlates with its localization in the inverted repeats. The majority of genes, in particular those encoding subunits of the photosynthetic apparatus have turned non-functional during or shortly after the transition to holoparasitism. The number of putative pseudogenes of photosystem genes and subunits for the electron transport imply that their last common ancestor had probably already lost the ability of photosynthesis. According to our results, the loss of photosynthesis did not lead to a physical deletion of those elements at this point. Only five genes (*ndhA, petB, psbT/N/H*) have been recovered as candidates for early gene losses. Analyses with a denser taxon sampling might however revise their reconstructed state at this node to pseudogenized rather than early loss. Along with photosynthesis-genes, the function of the plastid-encoded polymerase (normally responsible for transcription of most photosynthesis-genes) is highly likely to have become obsolete along the way to holoparasitism as well. Relaxation of selective pressure to maintain all other elements from the genetic apparatus apparently took place later. Repeatedly but independently in the different broomrape lineages, several tRNAs as well as some ribosomal protein genes are

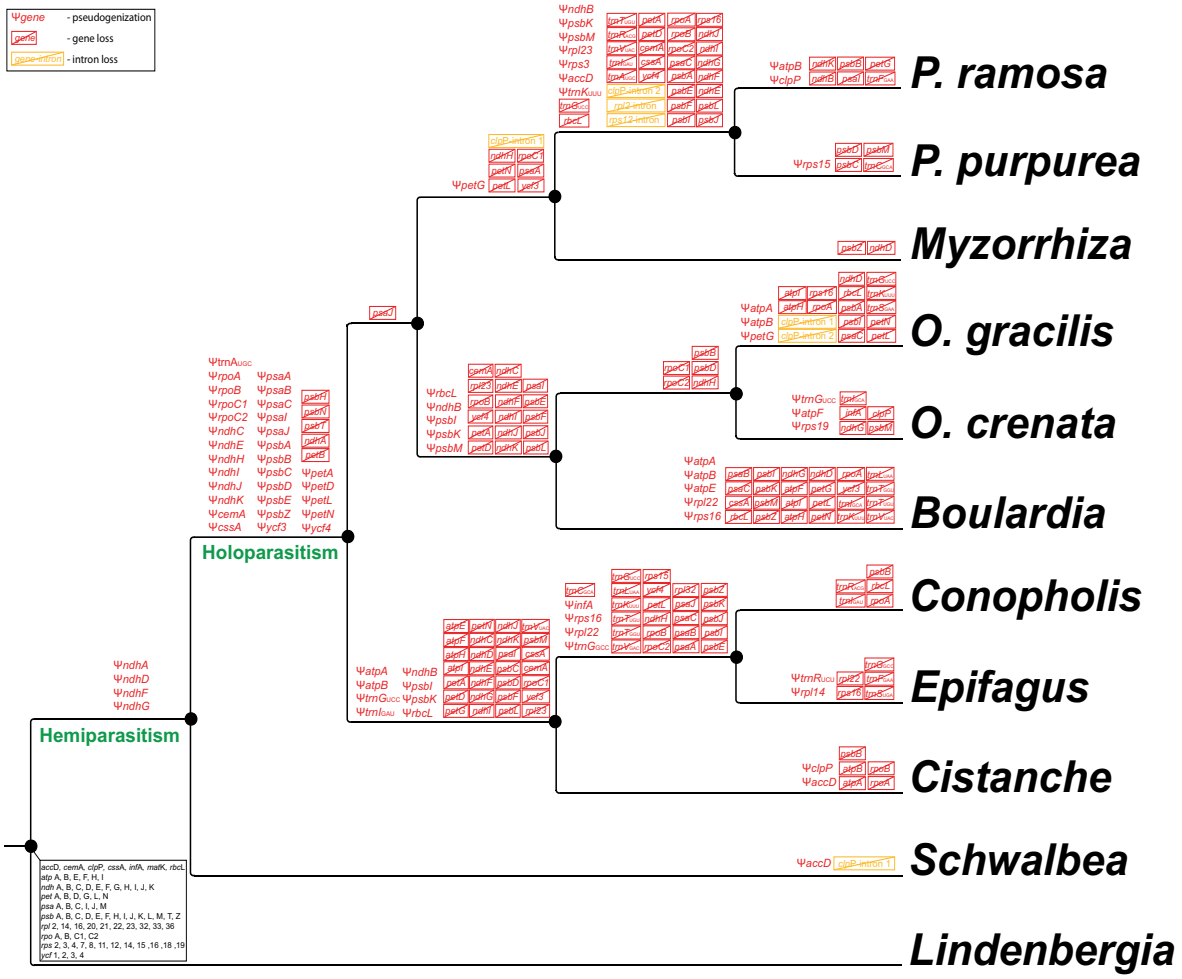


Fig. IV-5 **Graphical summary of gene losses and pseudogenization events based upon the reconstruction of plastid gene contents at ancestral nodes in Orobanchaceae.** The complete set of functional protein-coding plastid genes is provided at the root node. Pseudogenization of genes is demarcated by a "Ψ"-symbol along the branch. Boxed gene names indicate its deletion from the plastome. Intron losses are illustrated in orange.

lost. Among them, *rps16*, *rpl22* and *rpl23* as well as the tRNAs Ala (GCA), Ile (AUC), Gly (GGA), and Val (GUA) turn non-functional or are lost early during the evolution of broomrapes. However, most non-functionalization events occur localized at terminal branches. Those pseudogenizations/losses of genes for the genetic apparatus includes *rpl14*, *32* and *rps3*, *15*, *19*, *infA*, and *clpP*.

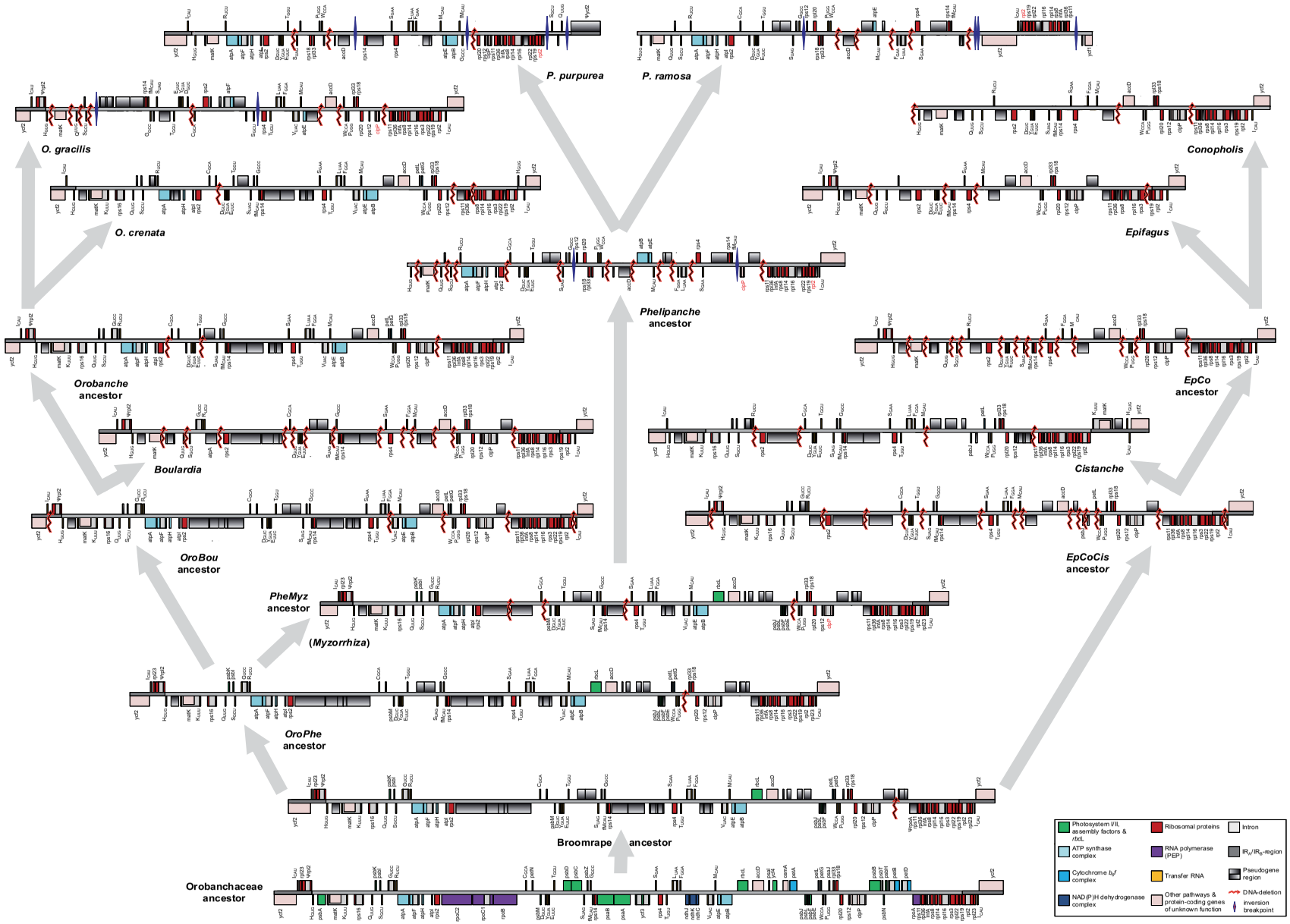
Most interestingly, not all genes for photosynthesis have been lost evenly rapidly from the plastome. In contrast, some subunits of photosynthesis-related elements have been maintained as functional reading frames after the transition to holoparasitism. This includes all genes for the ATP-Synthase complex, *rbcL*, as well as some subunits for photosystem II (*psbF, I, J, K, L, M, T*). This is most surprising, as we would generally

assume that gene regions and plastome fractions evolve similarly freely when molecular evolutionary selection sets out by the loss of photosynthesis. Consequently, this would imply that each subunit of every sub-complex would have the same chance to be lost. Our results indicate however, that pseudogenization and deletion of genes seems to occur not free of constraints.

1.5 Reconstruction of ancestral plastid genomes reveal the series of functional losses and suggest a major protecting role of conserved genetic elements.

We employed a Bayesian approach to reconstruct the rearrangement history of broomrape plastid genomes in that we subdivided the plastid chromosome of Orobanchaceae into locally co-linear blocks. With this approach, we are able to show that - with one exception only – structural plastome reconfiguration occurs locally on terminal branches, and may thus be interpreted as species-specific changes. However, the severe plastome reorganizations with large- scale inversions in the *Phelipanche* lineage are likely to have occurred in their common ancestor, and raised further independent inversion events in *P. ramosa* and *P. purpurea* (Fig IV-6). Questions concerning the IR-expansion/contraction history can only be answered on an argumentative basis, since currently available algorithms cannot unambiguously treat and process repeat regions. Considering the phylogenetic relationships of the taxa investigated and the frequency with which we observe changes around IR boundaries in the different clades, we have reason to believe that many expansion/contraction events have been present at earlier nodes. In particular, the probability of an ancestral IR-reduction at the *Phelipanche* node appears as highly likely for several reasons: In contrast to *Myzorrhiza* (sister taxon to *Phelipanche*), both *P. purpurea* and *P. ramosa* show an extremely aberrant IR-behavior. *Phelipanche purpurea* retains at least a partial IR- region, and is sister to the remaining species of this section (Schneeweiss et al. 2004; Park et al. 2008). Although currently unconfirmed by PCR-based verification, a third species, *P. lavandulacea*, also exhibits a substantial IR-reduction (own data, unpublished), which is highly similar to that of *P. purpurea* indicating that IR-reduction might display the ancestral state.

Fig. IV-6 **Evolution of the plastid LSC and adjacent regions from an autotrophic ancestor towards non-photosynthetic taxa.** Based on likelihood reconstructions of both node-specific ancestral gene contents and rearrangement history, the illustration summarizess the continuous functional and physical reduction from a putative autotrophic ancestor's plastome towards holoparasitic broomrape plastomes. Blackish-shaded gene boxes indicate events of pseudogenization. Jagged arrows through the chromosome bar mark a gene or segmental deletion; a blue diamond indicates the breakpoints of an inversion. We set aside an illustration of due to its high similarity to the inferred putative ancestor of the lineage. [Abbr.: Bou – *Boulardia*, Cis – *Cistanche*, Co – *Conopholis*, Ep – *Epifagus*, Myz – *Myzorrhiza* Oro – *Orobanche*, Phe – *Phelipanche*.]


P. purpurea
P. ramosa
O. gracilis
Conopholis
O. crenata
Epifagus
Phelipanche ancestor
Orobanche ancestor
EpCo ancestor
Boulardia
Cistanche
OroBou ancestor
EpCoCis ancestor
PheMyz ancestor
(Myzorrhiza)
OroPhe ancestor
Broomrape ancestor
Orobanchaceae ancestor
Photosystem I/II, assembly factors & rbcL
ATP synthase complex
Cytochrome b6f complex
NAD(P)H dehydrogenase complex
Ribosomal proteins
RNA polymerase (PEP)
Transfer RNA
Other pathways & protein-coding genes of unknown function
Intron
IRA/IRB-region
Pseudogene region
DNA-deletion
Inversion breakpoint

1.6 Neighboring-genes and the plastid operon-structure significantly determine the tempo of genome segment evolution, but leave the retention of ATP-Synthase genes unexplained.

In order to identify recurrent patterns in the series of gene losses, we reconstructed the plastid chromosome structure at major ancestral nodes of the Orobanchaceae tree by combining the results from our analyses of plastome rearrangement history with those of the state reconstruction analysis regarding gene functionality (ASR, Fig. IV-9). ASR-results strongly suggest that a scenario where genes and plastome regions are deleted randomly may only apply to a delimited extent (Fig. IV-12). Although a fraction of the plastid genome appears to contain a substantial proportion of dispensable regions in the ancestors, the amount of true losses vs. retention of pseudogene regions appears to be unevenly distributed and apparently proceeds with a locus-specific tempo. We have reason to assume that after the relaxation of selective pressures, the deletion of dispensable DNA-fractions depends upon the proximity to conserved genes (hereafter termed "neighboring-gene effect"). Deletion of plastomic regions seems to affect more frequently dispensable genes that are more distantly localized to conserved housekeeping genes. This may explain the "early" deletion of *rpoC1* whose distance is farther away from upstream and downstream subunits of the adjacent *rpo*-subunits C2 and B. Similarly, we observe that *psaA* is deleted more frequently than *psaB*, the latter of which is localized shortly downstream of *rps14*. The distance to other, i.e. more essential genes appears to protect the region from deletion. Another reason for the different tempo of deletion may be the organization of genes in operons or operon-like transcription units. In this respect, multi-functional operons (e.g. *rpoA-rpl*/*rps*-operon, *rps14-psaB*/*A*-operon) apparently contribute to a longer retention of dispensable segments, whereas loss of the entire operon containing genes of a similar function seems to occur more frequently (e.g. *psbI*/*K*-operon, *petB*/*D-psbH*/*N*/*T*/*B*-operon). Thus, retention of dispensable plastomic regions may be severely influenced by the operon-structure in that multi-functional operons contribute to a longer survival of similar or equal functional classes ("operon-effect"). Genes of different complexes are not uniformly encoded in operons of similar or equal functions; the "operon-effect" might also provide a possible answer to the question if different functional classes are lost at different tempos after the loss of selective constraints. We grouped our genes according to their survival rate, i.e. the longevity of a gene over the tree considering the node depth at which a dispensable gene was functionally lost (Fig. IV). For each of these genes we approximated the distance per node to its next conserved genes. We therefore defined "conserved" genes as those that were found universally present in Orobanchaceae (Supplemental material: Table SIV-A and SIV-D). In order to evaluate the "operon effect", we encoded each gene in a binary matrix according to whether it is found encoded in transcription units in photosynthetic

Table IV-C **Results from a gamma correlation test evaluating the "operon-effect" and "neighboring-gene-effect" upon plastid gene loss.** Test statistics are summarized for the different hypotheses covering aspects of the "neighboring-gene-effect" or the "operon-effect". Asterisks mark the significance level (<0.05*, <0.01**, <0.001***).

Hypothesis	N	Gamma	Statistic Z	p-value
"Operon-effect"				
Longevity of dispensable genes relates to operon-structure	77	-0.35081	-2.7859	0.005**
Deletion of a dispensable gene relates to operon structure	77	0.22852	1.7442	0.081
Deletion of dispensable genes relates to operon-type (multi vs. equal)	57	0.16741	1.1974	0.231
Deletion of dispensable genes relates to operon-type (multi vs. similar)	57	-0.54161	-4.0742	<0.001***
Longevity of dispensable genes relate to operon-type (multi-equal)	57	-0.24654	-1.8261	0.068
Longevity of dispensable genes relate to operon-type (multi-similar)	57	0.44670	3.4871	<0.001***
Genes of the same operon share the same fate of being lost	57	-0.01634	-0.1654	0.868
Genes of the same operon share similar longevity	*57*	*-0.03534*	*0.3706*	*0.711*
"Neighboring-gene-effect"				
Deletion of a dispensable region relates to its distance to conserved genes	77	0.30402	3.6987	<0.001***
Survival as pseudogenes relates to its distance to conserved genes	77	-0.29851	-3.6374	<0.001***
Longevity of dispensable genes relates to its distance to conserved genes	77	-0.33925	-4.2846	<0.001***

angiosperms (Supplemental material: Table SIV-E). We further grouped genes according to their localization in operons of multi-function (e.g. *rps2-atpA/F/H/I,* i.e. housekeeping and photosynthesis-function), operons of similar-function (e.g. *petB/D-psbH/ N/T/B*-operon, i.e. all related to photosynthesis), or equal-function (e.g. *ndhC/K/J*, -complex). The results of our hypothesis tests evaluating aspects of gene loss and pseudogenization in parasitic broomrapes (Table IV-C) support our assumptions derived from structural analyses that operons severely influence the tempo of gene loss. The deletion of a gene marginally relates to its localization in an operon (p = 0.0811). However, taking into account the node depth at which genes are likely to have been lost (or how long they survive), we are able to show that the "operon-effect" does play a crucial role in deletion of dispensable plastid DNA (p = 0.0053**). In this process, the type of the operon is an important factor. Deletion of unessential genes encoded in multifunctional operons are seen less frequently than those encoded in transcription units of similar function (p < 0.001***); no such effect is evident between multi-functional operons and those encoding subunits of equal functional complexes (p = 0.2312), supposedly due to the relatively small amount of equal-function operons in the plastid and the small amount of observations. Above that, our analysis reveals that the longevity of a gene clearly relates to the operon nature ($p_{SimFun.}$ < 0.001***, $p_{EquFun.}$ = 0.0678). Based upon these results, the question arises whether genes of the same operon are likely to share the same fate. If so, we would assume that all unessential genes are lost or retained according to the residing operon. This hypothesis cannot be supported. Genes loss and longevity of genes happen independent of the operon type (p_{loss} = 0.8686, ($p_{longevity}$ = 0.7109). Thus, other factors are likely to influence the continuous deletions of dispensable DNA from the plastomes of non-photosynthetic plants. One such factor may be the physical distance to a conserved gene. That is, the closer a dispensable region is located to an essential gene the longer the dispensable region will survive after relaxation of selective constraints. In this case, we disregard the case of co-localization of both the

essential and nonessential gene within the same transcription unit. Indeed, we find that its vicinity to an essential gene significantly influences the presence of an unessential gene (e.g. photosynthesis-related elements) as a pseudogene ($p < 0.001$***). Likewise, we find significantly more genes deleted the farther they are located from essential elements ($p = 0.001$***). Essential genes thus seem to exhibit a protecting factor shaping the tempo with which dispensable genes are deleted from the plastome ($p < 0.001$***).

1.7 Strong nucleotide compositional bias leads to notable relaxations of codon usage in the plastid chromosome of parasitic plants.

Functional genome reduction affects the G/C-content of plastid chromosomes in parasitic plants (Table IV-A). Compared to photosynthetic relatives, holoparasitic species are on average 3-4 % richer in A/T-content. Only marginal deviation exists in the G/C-content of structural RNAs, whereas major differences exist in both coding and non-coding plastid regions. IR-lacking holoparasites exhibit a notably higher A/T-content in non-coding plastid DNA regions with only ~25 % G/C in *Phelipanche* and 28 % in *Conopholis* compared to more than 30 % in the remainder parasites and non-parasites. A/T-richness constitutes a recombinogenic factor increasing particularly illegitimate recombination, and may thus contribute a possible explanation for the number of species-specific structural changes in Orobanchaceae. Higher rates of recombination may lead to local disruptions in gene synteny. We conducted a gamma correlation test in order to evaluate a possible correlation between structural evolution of broomrape genomes and their G/C-contents (Table IV-D). The test reconfirms that G/C-content drops in parasites ($p = 0.046$*), and, particularly decreases in holoparasites ($p = 0.003$**). However, changes around the IR-segments appear to be uncorrelated to plastid G/C-content ($p_{non\text{-}coding} = 0.743$, $p_{total} = 0.465$,). In addition, our analyses show that there is a prominent trend towards more inversions in A/T-richer plastomes. The dropping G/C-value in non-coding region contributes to alteration in gene synteny with marginal significance ($p_{total} = 0.082$, $p_{non\text{-}coding} = 0.052$). Compositional bias towards A/T is about the same order of magnitude in plastid coding regions as in non-coding segments. Figure IV-7 and IV-8 suggest that changes at the third codon position alone may be insufficient to account for the nucleotide bias. In autotrophs and photosynthetic *Schwalbea*, the median G/C-content differs significantly between positions 1, 2 and 3 with position 1 having the highest rate of G/C bases. In parasites, this distinct variation diminishes remarkably (Fig. IV-8). The observed bias towards A/T evidently affects the first and second codon position to the point that the G/C-content of the first codon position converges to that of the second one (Fig. IV-7, IV-8). This is noteworthy as the nucleotide substitutions of the first and second codon position are typically very low in genes under purifying selection. In contrast to the third position, substitutions in position 1 and 2 lead with a higher probability to a non-synonymous

Table IV-D **Results from statistical tests evaluating the relation of G/C-content to structural features.** All test statistics incl. the p-value resulting from gamma correlation tests are provided for each tested hypothesis. The significance levels are marked by asterisks (<0.05*, <0.01**).

Hypothesis	N	Gamma	Statistic Z	p-value
GC-content drops in parasites	12	-0.800000	-1.99309	0.046*
GC-content drops in holoparasites	12	1.000000	2.89470	0.003**
Total GC-contents relates to the number of observed inversions	12	-0.513514	-1.74010	0.082
Total GC-content relates to changes around the IR-segments	12	-0.191489	-0.73133	0.465
Non-coding GC-contents relates to the number of inversions	12	-0.567568	-1.93801	0.052
Non-coding GC-content relates to changes around the IR-segments	12	-0.086957	-0.32753	0.743

amino acid change potentially affecting protein functionality. Thus, if the retained genes in parasite plastomes evolve under selective pressure, significant deviations from this pattern of base pair composition and distribution are not expected. This compositional bias indicates relaxed patterns regarding the evolution of codon-positions in parasite plastid coding regions. In order to investigate whether the functional class of those genes retained in the plastomes of parasites causes the compositional bias, we analyzed the differences in base pair distribution between matching gene pairs of parasites and close photosynthetic relatives. To this end, we conducted a series of Wilcoxon tests between three autotrophic plants (*Nicotiana*, *Mimulus*, and *Lindenbergia*) and all parasitic species. Firstly, we evaluated whether the observed differences in the overall G/C-content truly exist between photosynthetic and non-photosynthetic plastid coding regions. With the exception of *Myzorrhiza*, we can clearly reconfirm that all holoparasites show a significantly smaller G/C-usage in plastid coding regions than photosynthetic plants (Supplemental material: Table SIV-B). In the majority of cases, observed differences were highly significant (p-value < 0.001***). In contrast, the base composition among photosynthetic representatives (*Nicotiana*, *Mimulus*, *Lindenbergia*, and hemiparasitic *Schwalbea*) does not significantly differ from each other, although *Mimulus* is marginally richer in G/C than other autotrophs. Similar to most of the holoparasites, the hemi-parasitic *Schwalbea* does also

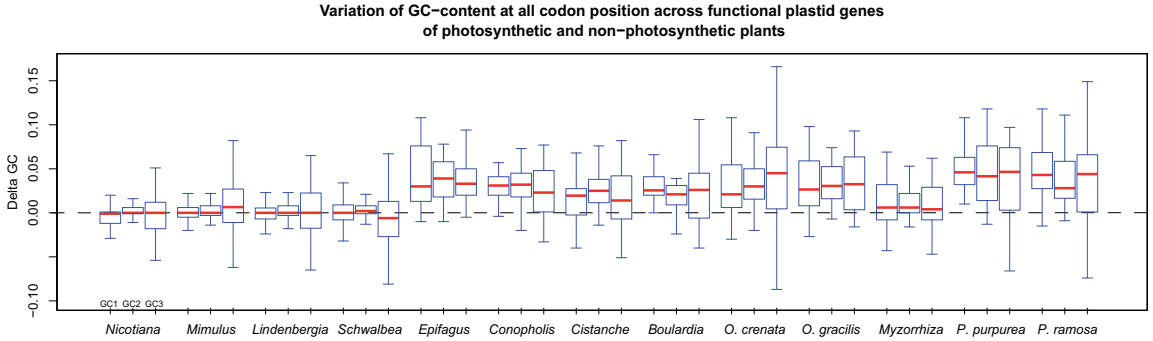


Fig. IV-7 **Median difference of GC content at the first (GC1), second (GC2), and third (GC3) codon of Orobanchaceae to an autotrophic reference.** Red bars mark the codon-position specific median difference in GC of Orobanchaceae to the photosynthetic reference (*Aucuba japonica*), which was determined in pairwise comparisons over all functional genes per species. A thin dashed line marks the reference (i.e. zero).

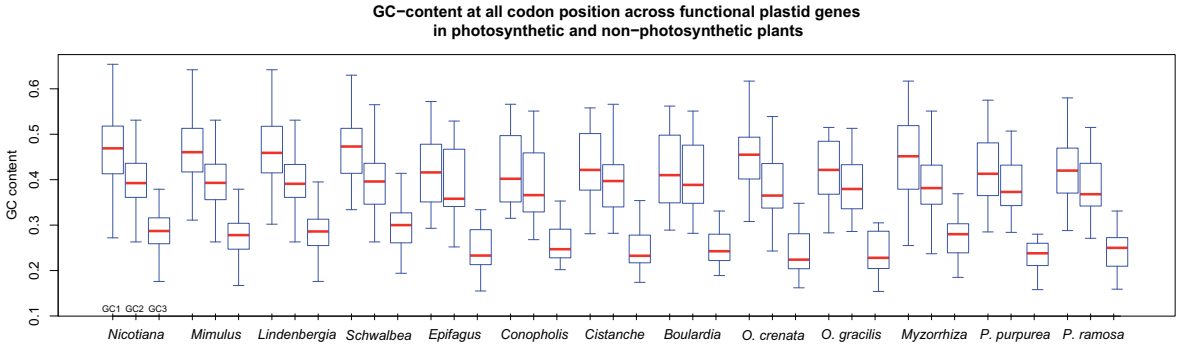


Fig. IV-8 **Codon-position specific G/C-content in non-parasitic and parasitic taxa.** G/C-content at the first (GC1), second (GC2), and third (GC3) codon position is illustrated for two representative autotrophs and Orobanchaceae. Red bars mark the median G/C-content over all functional genes per species.

exhibit a higher A/T-content than non-parasites. On a nucleotide specific level, least variation exists in the relative amount of T's at any codon position. Boxplots in Fig. IV-7 and IV-8 imply that the compositional bias varies considerably in the retained coding regions of parasites compared to autotrophic lineages. In order to evaluate if such changes uniformly affect all genic regions, we conducted gene- specific Wilcoxon tests between parasites and autotrophs. Therefore, we compared the A/T-usage for the individual codon positions between a set of photosynthetic lamiid angiosperms and parasitic Orobanchaceae. We performed analyses for 31 plastid protein genes. This corresponds to the smallest set of common genes and includes the *atp*-genes, which are found potentially functional in some of the holoparasites. We excluded *ycf1* because of its susceptibility to sequencing/assembly error accumulation due to homopolymeric stretches. Our test results strongly support our hypothesis that variation in the G/C-content at the third codon position does not account for the major shift towards an overall A/T-richness in the parasite plastid coding regions (Table IV-E). Moreover, our results show that all codon positions are equally affected exhibiting a general shift towards A and T. Differences in the third codon position exist between parasite and non-parasites in about half of the tested gene set. No differences in the G/C-content at the first two codon positions exist between parasites and non-parasites in only four genes (*rps14, rpl22, 23, 32*). In 14 out of 31 plastid genes, however, we show that the amount of A/T-bases at both the first and second site is significantly higher in parasites. Thirteen genes more exhibit significant changes in either the first or the second codon position. Interestingly, among those we find three *atp*-genes (*atpE, H, I*), *accD* and *clpP*. Thus, the compositional drift towards A/T at positions 1 and 2 of the majority of parasite plastid genes is of similar magnitude as that normally observed only for the third position. Interestingly, three of the genes identified to contain similar G/C-contents in parasites and non-parasites are lost in holoparasites *and* are known to be pseudogenized or deleted from the plastomes of several autotrophic lineages (*rpl22, 23,*

Table IV-E **Results from unpaired Wilcoxon test evaluating whether variation in G/C-contents among autotrophs is similar between autotrophic and parasitic plants.** The p-values for all codon positions are provided for 29 plastid genes. Asterisks denote the significance level (* <0.05, ** < 0.01). The taxon subsets included in the genewise tests are given for parasitic and non-parasitic species.

Gene-ID	Subset of non-parasitic taxa	Subset of parasitic taxa	GC1	GC2	GC3
			p-value	p-value	p-value
accD	Mg Lp Am Oe Ab Nt Sl Ca No	Ev Co Bm Oc Og Mc	0.316	0.003**	0.018*
atpA	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Oc Mc Pp Pr	0.003**	0.016*	0.032*
atpB	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Oc Mc Pp Pr	0.003**	0.005**	0.075
atpE	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Oc Og Mc Pp Pr	0.082	0.021*	0.003**
atpF	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Oc Og Mc Pp Pr	0.022*	0.016*	0.173
atpH	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Oc Mc Pp Pr	0.007**	0.792	0.023*
atpI	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Oc Mc Pp Pr	0.113	0.027*	0.500
clpP	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Ev Co Bm Oc Og Mc Pp	0.011*	0.393	0.285
infA	Mg Lp Am Jn Oe Ca No	Sa Cp Bm Oc Mc Pp Pr	0.006**	0.006**	0.032*
matK	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Ev Co Cp Bm Oc Og Mc Pp Pr	0.001**	0.001**	0.040*
rpl33	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Co Cp Bm Oc Og Mc Pp Pr	0.007**	0.002**	0.060
rpl16	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Ev Co Cp Bm Oc Og Mc Pp Pr	0.003**	0.000***	0.520
rpl2	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Ev Co Cp Bm Oc Og Mc Pp Pr	0.001**	0.001**	0.289
rpl20	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Ev Co Cp Bm Oc Og Mc Pp Pr	0.006**	0.001**	0.075
rpl22	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Cp Oc Og Mc Pp Pr	0.525	0.807	0.007**
rpl23	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Cp Oc Og Mc Pp Pr	0.090	0.419	0.297
rpl32	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Cp Bm Oc Og Mc Pp Pr	0.894	0.068	0.020*
rpl33	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Ev Co Cp Bm Oc Og Mc Pp Pr	0.016*	0.233	0.049*
rpl36	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Ev Co Cp Bm Oc Og Mc Pp Pr	0.065	0.025*	0.876
rps2	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Ev Co Cp Bm Oc Og Mc Pp Pr	0.238	0.003**	0.427
rps3	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Ev Co Cp Bm Oc Og Mc	0.001**	0.005**	0.021*
rps4	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Ev Co Cp Bm Oc Og Mc Pp Pr	0.023*	0.001**	0.059
rps7	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Ev Co Cp Bm Oc Og Mc Pp Pr	0.004**	0.001**	0.025*
rps8	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Ev Co Cp Bm Oc Og Mc Pp Pr	0.002**	0.006**	0.005**
rps11	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Ev Co Cp Bm Oc Og Mc Pp Pr	0.001**	0.001**	0.210
rps12	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Ev Co Cp Bm Oc Og Mc Pp Pr	0.027*	0.267	0.405
rps14	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Ev Co Cp Bm Oc Og Mc Pp Pr	0.124	0.646	0.003**
rps15	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Cp Bm Oc Og Mc Pr	0.001**	0.107	0.883
rps16	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Cp Oc Mc	0.027*	0.777	0.009**
rps18	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Ev Co Cp Bm Oc Og Mc Pp Pr	0.001**	0.305	0.239
rps19	Mg Lp Am Jn Oe Ab Nt Sl Ca No	Sa Ev Co Cp Bm Og Mc Pp Pr	0.189	0.003**	0.024*

Taxon abbreviations: Mg – *Mimulus*, Lp – *Lindenbergia*, Am – *Antirrhinum*, Jn – *Jasminum*, Oe – *Olea*, Ab – *Atropa*, Nt – *Nicotiana*, Sl – *Solanum*, Ca – *Coffea*, No – *Nerium*, Sa – *Schwalbea*, Ev – *Epifagus*, Co – *Conopholis*, Cp – *Cistanche*, Bm – *Boulardia*, Oc – *O. crenata*, Og – *O. gracilis*, Mc – *Myzorrhiza*, Pp – *P. purpurea*, Pr – *P. ramosa*

and *rpl32*). Thus, the similarity of G/C-content between non-parasites and parasites may be indicative that a high compositional bias already exists in *rpl22*, *rpl23*, and *rpl32* in non-parasitic taxa.

The significant albeit small difference in G/C-content of plastid protein-coding regions may possibly bias codon usage (CU) in parasites. Codon biases have been reported before for plastid genomes indicating that factors other than G/C-content mainly influence the usage of codons (Raubeson et al. 2005). If this holds true, we would not expect to find variations in codon usage between plastomes of photosynthetic and non-photosynthetic plants. Indeed, holoparasites do not differ from photosynthetic plants in codon

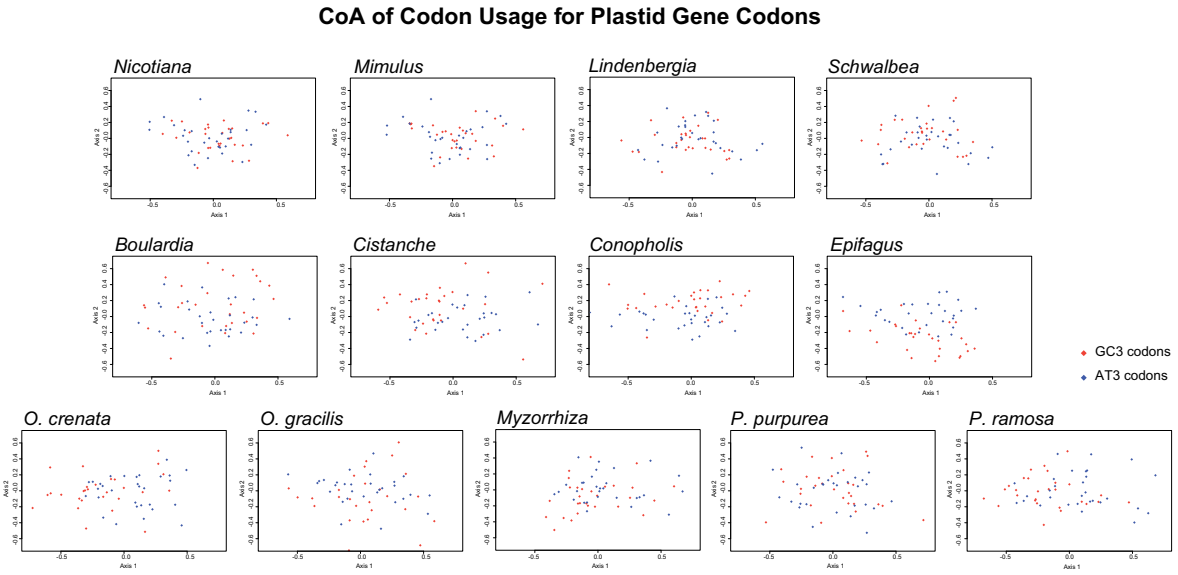


Fig. IV-9 **Correspondence analysis of codon usage (CoA-CU) for 59 degenerate codons in photosynthetic and non-photosynthetic Orobanchaceae and close relatives.** CoA-CU has been computed across all plastid genes. The most significant axes have been plotted against each other. Codons ending in G/C are highlighted in red.

preferences (Supplemental material: Table SIV-C). Although some localized underuses or overuses of selected codons do exist in *Epifagus*, *Conopholis*, *Boulardia* and *Phelipanche*, the majority of codon use is highly similar between holoparasites and autotrophs. Shifts are only observed between codons ending in A or U for isoleucine, valine, proline, and serine. Loss of a particular tRNA isoacceptor, if observed, does not generally alter codon usage. A correspondence analysis of codon usage (CoA-CU) supports the primary finding that changes in the nucleotide composition do not immediately affect codon usage. Plots of the most significant principal axes from CoA-CU for the 59 degenerated codons reveal a largely similar codon distribution between autotrophs and parasites (Fig. IV-9). Codons cluster tightly in *Lindenbergia*, *Nicotiana*, and *Mimulus*. A wider range of distribution along both axes indicates a slightly relaxed codon usage in parasitic species. Although suggested by A/T-richness of coding regions of parasites, the distribution of codons ending in G/C does not differ from A/T-ending ones. One notable exception of A/T-ending vs. G/C-ending usage appears to exist in the *Conopholis*/*Epifagus*-clade. In those taxa, we observe that G/C-codons separate more prominently from A/T-codons forming two clusters along the second axis. On gene level, CoA indicates somewhat relaxed patterns in parasites. In autotrophs, we detect a clear separation of usage of codons between genes of different functional classes (Fig. IV-10). Especially, codon usage of photosynthesis-related genes differs remarkably from that in housekeeping genes. Usage of genes functioning in pathways or genes of unknown function is interspersed. Interestingly, this clear formation

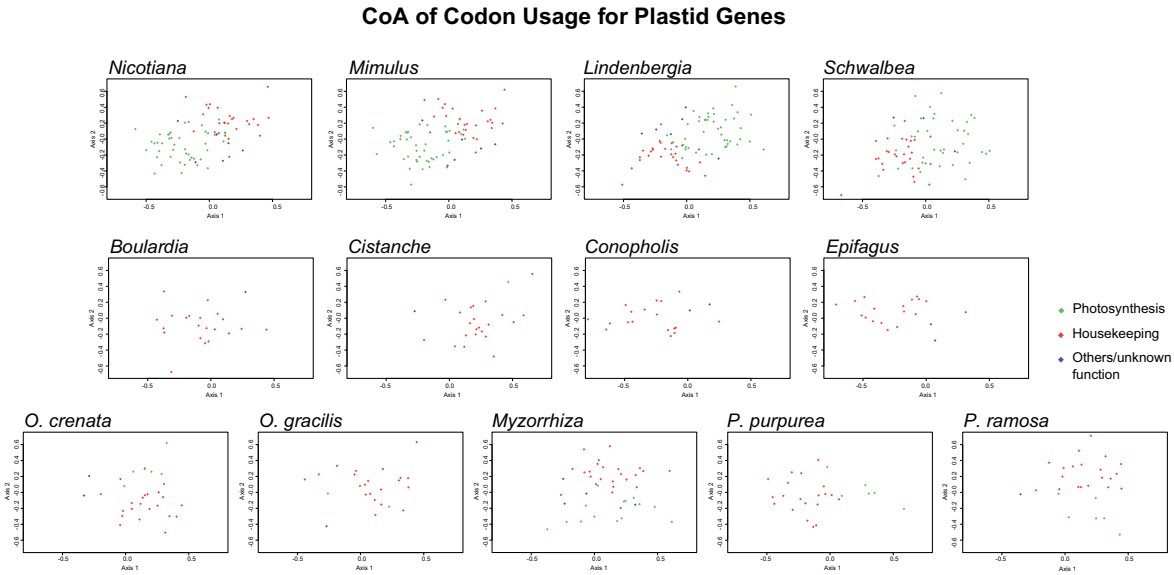


Fig. IV-10 **Correspondence analysis of codon usage (CoA-CU) for functional plastid genes in photosynthetic and non-photosynthetic Orobanchaceae and close relatives.** CoA-CU has been computed for all functional plastid genes per species. The most significant axes of the genes have been grouped into housekeeping genes (red squares), genes of the photosynthesis apparatus (green) and other or unknown function (blue).

of functional clustering is maintained in parasites harboring potentially functional photosynthesis subunits (*Myzorrhiza*, *Phelipanche*, and *Orobanche*). In *Schwalbea*, this tight clustering of photosynthesis gene codons breaks up slightly. In many yet not all holoparasites, the dense grouping of housekeeping genes in one quadrant relaxes notably. This relaxation occurs most evidently in the *Cistanche*/*Epifagus*/*Conopholis* lineage, *Boulardia* and *Orobanche gracilis* whereas function-related codon usage is distressed to a much smaller extent – if at all – in the remainder species.

3. DISCUSSION

3.1 Structural plastome evolution under relaxed selective constraints.

3.1.1 Structure of plastid chromosomes in Orobanchaceae

In the current study, we assessed evolutionary patterns of plastid genome reduction under relaxed selective pressures in a group of closely related parasitic plants. Our work revealed that plastid chromosomes in Orobanchaceae holoparasites are structurally highly diverse compared to the majority of angiosperm plastid chromosomes (Bock et al. 2007, Wicke et al. 2011). Reflecting different stages of reductive plastome evolution, the sizes of broomrape plastid chromosomes vary between 45 kb in *Conopholis* and 121 kb in *Myzorrhiza*. We demonstrated clearly that the functional relaxation also severely influences the structural evolution of the plastid chromosome. Excepting changes

of gene synteny due to gene deletion, we detected several inversions in broomrape plastomes. Frequently, the large IR-segments possess notably restructured boundaries. Our analyses of the rearrangement history revealed that most of this reconfiguration are lineage specific, and do not occur at deep nodes. We may therefore assume that the restructuring of plastid chromosomes marks a "late" phenomenon of reductive evolution. Apparently, structural maintenance relaxes with progressive plastome non-functionalization and progressive relaxation of constraints on the few preserved genes. Broomrape plastomes are rich in recombinogenic elements. G/C-content of the plastid chromosomes drops significantly after the loss of photosynthesis. We could show that A/T-richness in broomrapes relates to reconfiguration events such as partial or complete loss of one IR-segment. Stretches of A/T-microsatellites, long homo-nucleotide stretches as well as other repetitive DNA, which are abundant in broomrape plastomes, constitute target sites of, in particular, illegitimate recombination (Ogihara et al. 1988; Fejes et al. 1990; Müller et al. 1999; Gray et al. 2009; Maréchal and Brisson 2010). An elevation of repetitive DNA is irrespective of the stage of reductive evolution of the plastid chromosome. Although causalities are elusive, elevated amounts of repetitive DNA are likely to be the results of a relaxed evolutionary pressure normally selecting for plastomes with a low number of repeats of any kind in order to suppress deleterious, i.e. erroneous recombination. Structural changes thus may be the direct consequence of a dramatic increase of plastid repetitive DNA relative to autotrophs. We could clearly demonstrate that repeated elements of sizes larger than 20 bp spread out after the transition to parasitism. The effect is even more prominent in holoparasites. This size class significantly intensifies intramolecular recombination in plastid DNA (Müller et al. 1999). Even more, increase in direct (i.e. forward) and palindromic repeats provide further substrates for homologous and illegitimate recombination (Ogihara et al. 1988; Segall and Roth 1989; Sears et al. 1996; Haberle et al. 2008; reviewed in Maréchal and Brisson 2010). Improper intramolecular recombination might thus have led to the deletion of plastid DNA segments that are not under selection any more. Similar processes potentially account for certain length mutations in non-coding plastid regions of autotrophic plants (e.g. Ogihara et al. 1988; Saski et al. 2005; Daniell et al. 2006; Saski et al. 2007). Relaxation of selective pressures may lead to an increasing rate of (illegitimate) recombination or improper DNA repair that thereby creates smaller inversions and deletions of plastid DNA fragments, and eventually affects parts of the IR region as observed in *Phelipanche*, *Orobanche gracilis* and the non-photosynthetic orchid *Rhizanthella* (Delannoy et al. 2011). This scenario, i.e. increasing rate of illegitimate recombination, is probably powerful enough to enlighten the series of observed gene losses as well. It remains, however, elusive whether the IR-reduction/loss causes the severe restructuring, or vice versa. In the light of earlier results from other (autotrophic) angiosperm lineages with IR-losses, it may also be likely that the loss of one IR-segment seeds further plastome reconfiguration.

3.1.2 Functional reduction of the plastid genomes of parasitic plants

The extent of non-functionalization of photosynthesis related and housekeeping genes differs substantially among broomrape lineages. While hardly any traces of genes required for photosynthesis are detectable in the *Epifagus*/*Conopholis* clade, we do observe remnants of photosynthesis elements (intact and as pseudogenes) in *Myzorrhiza, Orobanche,* and *Phelipanche*. Genes for the electron transport and photosystems appear to be lost early after the transition to holoparasitism. We detected only few cases of putatively intact genes from those complexes (e.g. *petG, ndhB*). Genes for the ATP synthase complex display a distinctive exception. Similar to previous accounts (non-photosynthetic green alga *Prototheca wickerhamii* (Knauf and Hachtel 2002), non-photosynthetic liverwort *Aneura mirabilis* (Wickett et al. 2008), *Cuscuta*-species (Funk et al. 2007; McNeal et al. 2007)), we find all six *atp*-genes potentially functional in *Phelipanche* and *Myzorrhiza*. A reduced subset of *atp*-genes still exists in species of *Orobanche*. We identified a subset of 20 protein-coding genes as well as 18 structural RNAs as the minimum number of genes universally present in all investigated broomrape species. Housekeeping genes are repeatedly lost from broomrape plastomes; among those are *rps3, 15, 16, 19* and *rpl14, 22, 23, 32*. In a recent study, all but *rps15* have been shown to be essential in tobacco (Fleischmann et al. 2011). While no data is available for *rps19*, knockout experiments of those ribosomal protein genes resulted in hetero-transplastomic lines under selective growth conditions; homoplasmy, i.e. the elimination of the wild type plastid, could only be achieved for *rps15* knockout-plants whose ribosomes could be subsequently confirmed to function without the S15-protein (Fleischmann et al. 2011). Convergent patterns of plastid ribosomal gene losses have been reported from other nonphotosynthetic plants (reviewed in Wicke et al. 2011). In addition to losses of the above *rpl*/*rps*-genes, the highly reduced *Rhizanthella* plastome lacks a functional *rpl33*-gene. Above that, *rps12*-exons could not be attested to trans-splice correctly suggesting its pseudogenization (Delannoy et al. 2011). The plastid genomes of several other parasitic, i.e. non-photosynthetic, algae exhibits further ribosomal protein genes losses, for instance the achlorophyllous green alga *Helicosporidium* (de Koning and Keeling 2006), as well as numerous apicomplexian parasites (summarized in Fleischmann et al. 2011). In both apicomplexians as well as green algal parasites, nuclear-encoded elements for the translation apparatus are expressed and targeted to the plastid (de Koning and Keeling 2004; Jackson et al. 2011), suggesting that ribosomes are likely to be functionally assembled in parasites. In addition, plastid housekeeping genes of *Epifagus* have been shown to be expressed and processed (Wolfe, Morden, et al. 1992, 1992; Ems, Morden, et al. 1995). Since the majority of plastid encoded *rps* and *rpl*-genes have been shown to be essential for plants cell development (Ahlert et al. 2003; Rogalski et al. 2006, 2008; Fleischmann et al. 2011), their losses from the plastomes may be complemented by nuclear-encoded substitutes. Alternatively, the lack of one or more subunits likely

results in structurally aberrant ribosomes of reduced activity. It is conceivable that relaxed requirement of a plastid translation apparatus allow tolerating lower or reduced translation efficiency to some extent.

3.2 Evolutionary trends in reductive plastome evolution under relaxed selective pressure

Besides studying the series of functional diminution in parasitic angiosperms, our major intention was to identify evolutionary trends in the process of plastome reductive evolution resulting in the deletion of dispensable regions from the plastid chromosome in holoparasites. To this end, we employed reconstructed ancestral gene contents (ASR) and inferred the rearrangement history of broomrape plastomes, and combined those results with a thorough analysis of plastome architectural features. Within Orobanchaceae, ASR suggests that the thylakoid NAD(P)H-dehydrogenase complex (i.e. genes *ndhA-K*) probably displays the early-most functional loss accompanying the transition to a heterotrophic lifestyle; this conclusion is also in line with reports on *ndh*-gene losses in some photosynthetic seed plant lineages (Wakasugi et al. 1994; Wu et al. 2010; Schäferhoff 2011; Chris Blazier et al. 2011; reviewed in Wicke et al. 2011). Loss of the plastid Ndh-subcomplex may not affect the fitness of heterotrophs, since its function has been shown to be non-essential for cell survival (Peltier and Cournac 2002; Suorsa et al. 2009). The non-functionalization of genes for photosystems (*psa, psb* genes) and photosynthetic electron transport (*pet, atp* genes) occurs at a lineage-specific tempo, and is notably influenced by the localization of the respective dispensable gene within multifunctional operons and, more significantly, its distance to essential neighboring elements. While pseudogenization of several subunits appear to have occurred along or shortly after the transition to holoparasitism in the broomrape clade, deletion of photosynthesis-related genic regions from the plastome is only likely for very few subunits. The small degree of plastome reduction in *Myzorrhiza* compared to other lineages of broomrapes may also indicate the retention of the ability of photosynthesis for some time after the transition to a holo-heterotrophic lifestyle – at least in this particular clade. According to ASR, loss of housekeeping genes (ribosomal protein genes and structural RNAs) occur rather late and is either specific to individual clades or taxa. The plastid-encoded polymerase complex (PEP, *rpo* genes) transcribing the majority of photosynthesis-related genes may be an exception. Pseudogenization of PEP was inferred for the broomrape root node, which is in line with the observation of high mutational rates (large indels, nucleotide subsitutions) of *rpo* genes in *Schwalbea*. Compensation of lacking PEP-activity by a nuclear-encoded polymerase (NEP) has been demonstrated in both autotrophs and heterotrophs (Lusson et al. 1998; Hajdukiewicz et al. 1997; Krause et al. 2003; Berg et al. 2004). Relaxed pressure on

the rapid assembly of photosynthesis complexes due to an early dark vegetative phase during host-root attachment may render *rpo*-gene function dispensable at early stages of heterotrophy in Orobanchaceae.

Significantly, a notable nucleotide compositional shift towards increased A/T-use in both plastid coding and non-coding regions accompanies the reductive evolution under relaxed selectional pressures in parasites. We detected a slight drift in codon usage for the distinct holoparasitic sister taxa *Epifagus* and *Conopholis*, and corroborated a minor codon bias related to deleted plastid tRNA-genes (Wolfe et al. 1992a). However, the majority of Orobanchaceae holoparasites utilize codons mostly similar to non-parasites irrespective of A/T content, and regardless of the number of preserved tRNA genes. Although statistically not significant, there are some cases of over- and underuses of particular codons in parasites relative to autotrophs. Generally biased codon usage preferences due to unavailable tRNA-isoacceptors do not to exist in the *Phelipanche*/*Myzorrhiza* and *Orobanche*/*Boulardia* clades. However, *Epifagus* and *Conopholis* retain fewer tRNAs than the remainder holoparasites. Thus, it is feasible that selection of codons based upon the availability of isoacceptors has not yet set in or is currently underway in Orobanchaceae holoparasites.

3.3 Factors influencing pseudogenization and segmental deletions

3.3.1. Role of the plastid operon structure and essential neighboring elements on functional gene loss

In most cases, we observed that pseudogenes or potentially functional photosynthesis-related genes were located close to elements of housekeeping function. We have convincingly demonstrated that the survival rate of dispensable plastid genes clearly relates to its distance to an essential element. The protective "neighboring-gene" effect seems to weigh heavier than the protection from deletion by the operon-structure. In the latter case, we could show that genes encoded in multifunctional operons tend to survive longer after the loss of photosynthesis than those arranged in units of similar or equal function. These findings corroborate evidences derived from plastid architectural features (see section 3.1.1) that deletion of dispensable regions by improper recombination and/or impaired DNA repair processes comes also from the fact that essential genes. Genes under selection, supply a certain degree of protection from rapid loss. There is also a slight tendency that within-operon losses may be "guided" by regulating signals, but the current data is not sufficient to allow statistical testing. Although not tested in the current study, we may hypothesize that neighboring-gene protection also takes effect at other essential elements, such as transcription promoting, terminating, and processing signals as well as at replication origins. Simulation analyses of segmental deletions in scenarios of different

dependencies from essential element distributions may contribute to answering this question. In some cases, expansion of the IR-segment to include more (essential) genic regions may contribute to prevent from gene deletion. The relocation of the *trnK-matk* region into the IR in *Cistanche phelypaea* as well as complete IR-inclusion of *ycf1* in *Schwalbea* may be further examples. In particular, the case of *Cistanche* is very interesting, because we find the *trnK*-intron lost in its closest relatives *Epifagus* and *Conopholis*. Moreover, in both species, *MatK* shares a long N-terminal truncation of ca. 65 amino acids compared to *Nicotiana* (Ems et al. 1995). *Cistanche* does not possess these mutations (including *trnK*-loss). The IR-genes evolve slower than genes of other plastid segments in that they exhibit a strikingly smaller amount of nucleotide substitutions in both coding and non-coding DNA compared to the plastid single copy regions (e.g. Wolfe et al. 1987; Perry and Wolfe 2002). The *matK* and *ycf1* relocations of *Cistanche* and *Schwalbea*, respectively, may thus represent precedents of an "IR-effect. By translocation of genes into the IR, protein function might be assured due to the reduction of mutational rates in the relocated gene. A thorough analysis of the evolution of substitution rates and, particularly, changes of purifying selection will have to follow up to test this specifically.

3.3.2. Role of nuclear factors and evidence of increased intracellular gene transfer

Independent losses of different subunits classified as essential ribosomal proteins from holoparasites offer space for speculation of an increased (functional) transfer of plastid DNA to the nuclear genomes in parasites. Normally, organellar DNA transfers are non-functional incorporations (Bock and Timmis 2008; Kleine et al. 2009; Sheppard and Timmis 2009), and those organellar DNA transfers happened frequently during the evolution of plants (reviewed in Martin 2003). However, an overall increased rate in heterotrophs would also raise chances of functional gene transfers by coupling transferred genic regions to essentials of plastid proteins import (e.g. transit peptides and eukaryotic transcription signals). In the course of this work, we gathered evidence for increased rates of intracellular plastid DNA transfer. *Phelipanche* species appear to harbor an extraordinarily high amount of nuclear-encoded plastid DNA in the nuclear genome (Fig. IV-11). Fosmid end sequencing revealed that more than 90 % of positive hybridization signals from a heterologous probe cocktail of plastid genes originated from nuclear DNA fragments. In contrast, a homologous probe cocktail covering PCR-amplified plastid genes and spacer elements detected only ~ 40 % of non-plastid originating fragments. Hybridization experiments of the fosmid library with single gene probes have pinpointed the putatively transferred region around a ribosomal protein gene. We are currently validating these results by shotgun paired-end sequencing in our group (G. M. Schneeweiss, S. Wicke, T. Eder, and T. Rattei, unpubl. data)

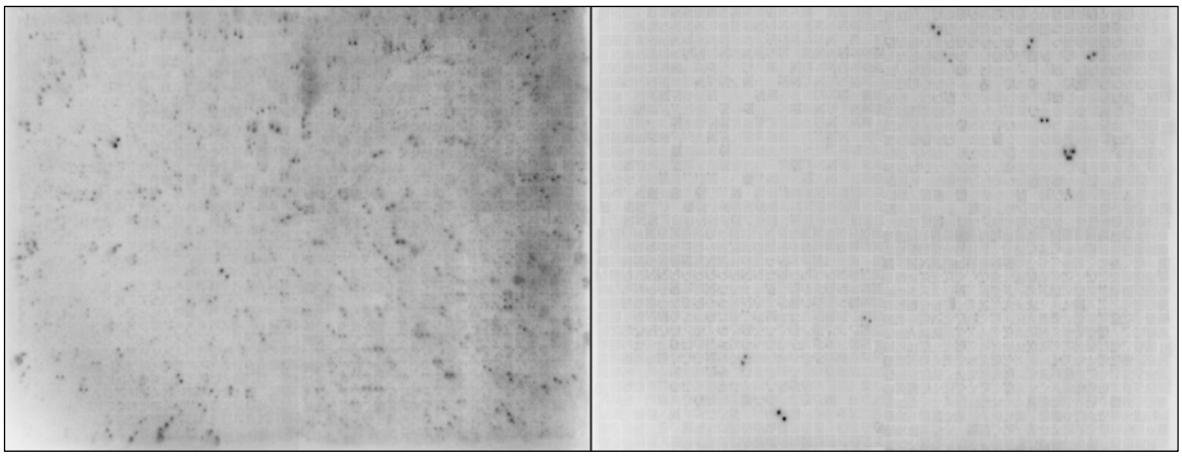


Fig. IV-11 **Nuclear plastid DNA fragments in the fosmid library of *Phelipanche purpurea*.** Approximately 18,000 redundantly spotted fosmid clones have been hybridized under the same conditions with a cocktail of ^{32}P-labeled heterologous plastid probes (left) and homologous probes (right). The heterologous probes bound to numerous clones with apparently high homology to plastid genes. Fosmid end-sequencing later on verified selected clones as nuclear DNA-fragments. In contrast to this, homologous probes yield only few clones of a non-plastid origin.

3.3.3. Rene retention due to a secondary and photosynthesis-decoupled function in nonphotosynthetic plants

Preservation of *atp* genes in a subset of holoparasites may be related to protection by either conserved neighboring genes or the operon structure of the plastid chromosome. At least four out the six plastid *atp*-genes (*atpA, F, H, I*) are localized in a multifunctional operon together with *rps2* encoding the ribosomal protein S2. *Rps2* has been shown to be essential for plastid ribosome function (Rogalski et al. 2008), and may therefore provide protection from rapid deletion. The remainder subunits form a separate transcription unit that is retained equally frequent in broomrape plastomes. The genes *rbcL* and *trnM* flank this operon in land plants; in *Prototheca*, *atpB*/*E* is bordered by *rps4* and *trnQ*. Thus, a "neighboring-gene-effect" for *atpB*/*E* plus an "operon-effect" in case of *atpA*/*F*/*H*/*I* may contribute to the preservation of all subunits in both lineages. Alhtough both the neighboring gene effect and the operon effect are not mutually exclusive, they may not account for their long survival as potentially functional genes in two distinct broomrape lineages and that of other non-broomrape holoparasites (Knauf and Hachtel 2002, Funk et al. 2007; McNeal et al. 2007 Wickett et al. 2008). Such an unusual long retention of gene functionality after the loss of photosynthetis has also been described for *rbcL* (reviewed in Wicke et al. 2011), leading to speculations of a putative involvement of RuBisCO in another photosynthesis-independent pathway. Eventually, RuBisCO was shown to contribute to a glycolysis bypassing reaction in the seeds of white turnip, *Brassica rapa* (Schwender et al. 2004). Similarly, the long retention of *atp*-genes may indicate another, yet unrecognized, function of the thylakoid ATP-Synthase, i.e. ATP-synthesis decoupled of the photosynthesis-driven, light dependent, ATP-generation. In photosynthetic plants, the

ATP-synthase utilizes a proton gradient originating from redox reactions during photoelectron transfer. This leads to conformational changes in the subcomplex resulting in the release of ATP. It is well known that certain substances alter the membrane permeability of protons and may decouple the ATP-release from the photoelectronic flux (Haraux and de Kouchkovsky 1998; McCarty et al. 2000). In this respect, it is conceivable that (some) parasites evolved (or employ) a mechanism for effective ATP-release that exploits a transmembrane gradient originating from sources other than a photoelectron transfer. Besides energy production, the thylakoid ATP-Synthase complex also functions as an ATPase (i.e. mediates ATP-hydrolysis) under certain conditions in photosynthetic plants. However, studies have shown that ATPase-activity is particularly low in the dark (McCarty et al. 2000). This function may thus be less likely to be the selective force mediating the retention of *atp*-genes in root parasitic non-photosynthetic plants like Orobanchaceae. Another interesting hypothesis offers the discovery from crystallization experiments revealing that the spinach plastid ATP-synthase complex apparently associates with pigment co-factors (Varco-Merth et al. 2008). In photosynthetic plants, carotenoids as well as chlorophyll prevent oxidative damage by non-photochemical quenching as a response to excessive light (e.g. Müller et al. 2001; Pascal et al. 2005). It may thus be conceivable that the retention of the ATP-synthase complex functions not only as an energy producing system, but also as a side effect may provide protection from light-induced damage in ancestral broomrape lineages.

4. CONCLUSION AND OUTLOOK

Using a model system of tropically distinguished plants, our study contributes significantly to understanding processes of reductive genome evolution under relaxed selective pressures. The analysis of a broad set of closely related species identified principal patterns of functional and physical plastome reduction. Reductive evolution of the plastid chromosome accelerates apparently shortly after the transition to a heterotrophic lifestyle in most, yet not all holoparasitic lineages. Although proceeding at alternating modes and tempos, reductive evolution exhibits recurrent patterns of gene loss and structural evolution of the plastid chromosome in holoparasites. We were able to demonstrate that major architectural changes accompany the process of reductive evolution under relaxed selective pressures. On the one hand side, we encounter a strong bias in nucleotide composition towards A/T-content throughout the plastome without an alteration of codon usage of most retained functional elements. However, a minor codon bias with respect to deleted plastid tRNA-genes could be shown. On the other hand, relaxation and the eventual loss of selective constraints introduce destabilizing elements

seeding major structural changes such as large inversions or IR-loss. The segmental deletion of dispensable DNA fragments proceeds at different tempos leading to an enormous diversity of structurally distinct plastid chromosomes even within a set of most closely related taxa. Major tempo-delimiting factors appear to be manifold with protection of neighboring essential elements probably being one important local influence in the plastid genome. Our likelihood reconstruction of putative states of gene function strongly suggests that the ancestor of the broomrape clade had already completed the transition to holoparasitism. However, the unexpectedly large plastome of *Myzorrhiza* may point towards a possible retention of photosynthetic ability along the backbone of the broomrape clade. Alternatively, *Myzorrhiza* may evolve several orders of magnitude slower than other broomrape holoparasites. An even denser sampling will be necessary to resolve the question regarding the preservation of semi-autotrophy along the backbone of this Orobanchaceae-clade (sensu Bennett and Mathews 2006). This will be essential to resolve the issue of how *fast* reductive evolution progresses after the complete loss of selective constraints. The marginally reduced plastome of holoparasitic suggests that several more as yet unknown factors severely influence or slow down gene loss. Potential functions of some photosynthesis complexes beyond photosynthesis may probably be another determining pressure. The current study provides reasonable evidence that the retention of *all atp*-genes in holoparasites may not be a coincident. Evidently, the long retention of those genes needs extra attention. So far, experiments are carried out in our labs in order to examine the extent of *atp*-genes expression (plastid and nuclear encoded) in *Phelipanche, Myzorrhiza* and *Orobanche*-species. Likewise, we are screening other non-photosynthetic lineages for putatively intact ATP-complexes and expression patterns during different stages in the life cycle of hemi- and holoparasitic plants. Besides this, a thorough and powerful analysis on the nucleotide level is currently underway. This will include the systematic analysis of the evolution of nucleotide substitution rates, especially focusing on purifying selection of coding regions in broomrape photosynthetic and non-photosynthetic parasites. The discovery that reductive evolutionary processes accelerate with the transition to heterotrophy brings into light the necessity to investigate plastid-related evolutionary processes in the fascinating group of *photosynthetic* parasitic plants.

5. MATERIAL AND METHODS

5.1 Taxon sampling

Two photosynthetic as well as eight non-photosynthetic parasitic members of the broomrape family have been subjected to complete plastome sequencing (*Lindenbergia philippensis, Schwalbea americana, Conopholis americana, Cistanche phelypaea, Boulardia latisquama, Myzorrhiza californica, Phelipanche purpurea, P. ramosa, Orobanche crenata, and O. gracilis*). Voucher information for all species examined here is summarized in Table IV-F. Sequences of the complete plastomes have been determined via two different approaches using fosmid libraries (McNeal et al. 2006), and shotgun-pyrosequencing from total genomic DNA. The dataset was complemented with the plastid DNA sequence of *Epifagus virginiana* (Wolfe et al. 1992a), which was obtained from NCBI Genbank.

Table IV-F **Plant material used for plastome sequencing**. Information of the source of plant material and the respective voucher information summarized for all newly sequenced Orobanchaceae species sorted in alphabetical order. Information on the applied sequencing method per taxon is also provided. [Abbreviations: WGSP – whole genome shotgun pyrosequencing, FSS – fosmid clone shotgun Sanger-sequencing, FPS – fosmid-clone pyrosequencing]

Taxon name	Source and voucher	Sequencing method
Boulardia latisquama	Collected at the Cap de la Nao, Spain, voucher deposited as S. Wicke s.n. 22.04.2008 at the Vienna University Herbarium)	fosmid libraries : FSS and FSP
Cistanche phelypaea	Collected between Murcia and Calasparra, Spain, voucher deposited as S. Wicke s.n. 25.04.2008 at the Vienna University Herbarium)	fosmid libraries: FSS
Conopholis americana	Collected in PA, USA, voucher deposited as dePamphilis s.n., in the the in the private herbarium of C.W. dePamphilis.	fosmid libraries: FSS
Lindenbergia philippensis	Cultivated at PennState University, S. Wicke #LP60/LP61, 29. Sept. 2009, desposited at the in the private herbarium of C.W. dePamphilis.	WGSP
Myzorrhiza californica	Cultivated at PennState University, S. Wicke #Oc54, 18. Aug. 2009, desposited at the in the private herbarium of C.W. dePamphilis.	fosmid libraries: FSS
Orobanche crenata	Cultivated on *Vicia faba* at at the Botanical Garden Bonn; voucher deposited as S. Wicke #OC41 at the Bonn University Herbarium.	fosmid libraries: FSS; WGSP
Orobanche gracilis	Collected in Lower Austria , parasitizing *Chamaecytisus sp.*, voucher deposited as G. Schneeweiss 7, Vienna University Herbarium	WGSP
Phelipanche purpurea	Cultivated on *Achillea millefolium* at the Botanical Garden Bonn; voucher deposited as S. Wicke Op38/39 at the Bonn University Herbarium.	fosmid libraries : FSS and FSP; WGSP
Phelipanche ramosa	Cultivated on tomatoe plants at the Botanical Garden Bonn; voucher deposited as S. Wicke Pr52/53 at the Bonn University Herbarium.	fosmid libraries : FSS and FSP; WGSP
Schwalbea americana	Collected by Kay Kirkman in Newton, GA, USA. 4. Aug., 2009; voucher deposited as Wicke/ dePamphilis #Sa57 in the private herbarium of C.W. dePamphilis.	WGSP

5.2 Fosmid library construction and library sorting

DNA extraction and purification was performed following the CTAB-based protocol of McNeal et al. (2006) using fresh and young flower tissue. Representative fosmid libraries were constructed for *Boulardia latisquama, Cistanche phelypaea, Conopholis americana, Myzorrhiza californica, Orobanche crenata, Phelipanche purpurea* using the CopyControl™ (HTP) Fosmid Library Production Kit (EPICENTRE® Biotechnologies) following the manufacturer's instructions with slight modifications. Five to 10 μg of freshly extracted total genomic and unsheared DNA was size selected by electrophoretic separation on a 30 cm long, 1% low melting point agarose gel. The gel was run overnight for a minimum of 16 hours at 8–12°C in freshly prepared 1× TAE buffer at 65 V. The gel was post-stained with GelStar® Nucleic Acid Gel Stain (Lonza), and the DNA was detected on a blue-light transilluminator. DNA fragments larger than undigested lambda DNA were excised using sterilized cover slips. Gel extraction and DNA purification was carried out as instructed in the Fosmid Library Production Kit using agarase and ethanol/sodium acetate precipitation. At least 0.5–1 μg of pure genomic DNA were ligated into the fosmid vector pcc1FOS or pcc2FOS, respectively, for a minimum of 2 hours at room temperature followed by a subsequent overnight incubation at 4–8°C. After inactivation of ligase (10–15 min at 71°C), the vectors-DNA concatemers were directly used for phage packaging and transduction. Size selection of genomic DNA was omitted for *Myzorrhiza*, where purified DNA was directly ligated into pcc1FOS. Selection and titering of fosmid-carrying clones was carried out using LB media and agar supplemented with 12.5 μg/ml chloramphenicol and 10 μg/ml cycloheximid. Between 2,000 and 3,000 fosmid clones were plated and grown on 24×24 cm LB-agar trays for 8-10 hours at 37°C. Libraries were sorted into 384-well plates filled with LB-freezing medium (Sambrook and Russell 2001) using a QPIX colony picker. Fosmid clones were redundantly arrayed on 22×22 cm positively-charged nylon membranes (Performa Nylon Filters, Genetix Ltd.) using a gridding robot (QPIX II, MicroGrid II) in 3×3, 4×4, or 5×5 offset double spotting pattern. Colony lysis/denaturation, neutralization, and fixation of DNA onto the filters were performed as suggested by the membrane manufacturer (Genetix Ltd.).

5.3 Fosmid library screening, probe preparation, end-sequencing

Plastid probes of all protein-coding genes and of selected tRNA-gene regions known to be present in the plastome of *Epifagus virginiana* have been PCR-amplified from *Nicotiana tabacum* using custom primers (available upon request). Probes were designed to be 0.2 to 1.5 kb long. GoTaq®Flexi polymerase system (Promega) was employed for PCR with reactions (25 μl) typically containing 1× Flexi reaction buffer, 20 mM MgCl, 0.1 M betaine, 0.20 mM of each dNTP, 10 mM of each amplification primer, 0.1 U *Taq*

polymerase, and 10–20 ng of template DNA. Cycling conditions were as follows: 3 minutes of pre-denaturation; 35 cycles with 30 s of denaturation, 20 s primer annealing at 50–53° C (i.e. 2–3° below calculated primer annealing temperatures), 60-150 s (according to the length of the expected product: 60 s per 1 kb) elongation at 68° C; 10 minutes of final elongation at 72° C. All probes were agarose gel-purified. Probes were sequenced at Macrogen Inc. (Seoul/ South Korea) prior to fosmid library screening.

Southern hybridizations of fosmid filters with plastid gene probes were performed following Sambrook & Russel (2001) with some modifications: Filters were pre-washed in 6x SSPE at 50°C for 30 minutes and subsequently pre-hybridized at 60 °C overnight in hybridization-buffer (5x SSPE + 5x Denhardt's solution + 0.2 % SDS) containing 10 µg/ml sheared and denatured herring sperm DNA. The probe cocktail (100 ng of 4–6 plastid gene probes mixed at equimolar ratio) was radiolabeled with ^{32}P-dATP using the Prime-a-Gene® Labeling System (Promega); labeling time was extended to two hours. Labeled probes were purified with custom Sephadex G-50 Superfine columns, eluting the samples stepwise with 100 µl 0.5x TE-buffer. The first three consecutive fractions with the highest emission of radiation were pooled, and denatured at 99°C for 10 minutes. Hybridization was carried out overnight at 61° C in fresh hybridization medium. Filters were washed twice with pre-warmed washing buffer (2x SSC + 0.2 % SDS) at 61°C for 10 min each, followed by a third 5 min washing step with room-temperate washing buffer; subsequently filters were briefly rinsed with 6x SSPE. Detection of positive signals was performed using a Typhoon 9200 Phosphoimager (GE Healthcare). Positive clones were prepped via alkali lysis (Sambrook and Russell 2001) or using the QIAprep Spin Miniprep Kit (Qiagen) following the manufacturer's instructions. End-sequencing of potential plastid-DNA carrying fosmid clones was performed at Macrogen (Seoul, South Korea) or at GATC Biotech (Konstanz, Germany), respectively.

5.4 Shotgun Sanger sequencing and pyrosequencing

Between three and five fosmids were selected for shotgun sequencing. Enriched fosmid DNA was isolated and purified using the NucleoBond® Xtra Midi Kit (Macherey-Nagel). Between 3 and 5 µg of freshly eluted fosmid DNA were precipitated with isopropanol, briefly washed twice with 70% ethanol, and resolved in 1 ml shearing buffer (0.5× TE, pH 8.3 + 10 % glycerol). Fosmid DNA was sheared to fragments of 2–3 kb length. Sheared DNA was precipitated by NaCl/ethanol precipitation, washed twice, and resolved in 50 µl 10 mM Tris-HCl (pH 8.0). DNA was end-repaired using the NEBNext® End Repair Module (New England Biolabs Inc.); DNA was additionally A-tailed if subsequent A/T-cloning was used. Subcloning libraries of each fosmid were generated using either the CloneJET™ PCR Cloning Kit (Fermentas) or the pGEM®-T Easy Vector System I

(Promega), respectively, in a 3:1 ratio of DNA to cloning vector. Positive clones were sorted into 384-well plates of LB+10% glycerol medium supplemented with the corresponding antibiotics. A minimum of two of these plates were sequenced bidirectionally via Sanger-sequencing at Macrogen (Seoul, South Korea). Alternatively, fosmid DNA was tagged and shotgun-pyrosequenced (12 fosmids in 1/8th of a picotiter plate) at the Center for Medical Research, Medical University Graz, Austria.

In addition to the fosmid-based approach, plastid un-enriched DNA of *L. philippensis*, *S. americana*, *O. gracilis*, *P. ramosa* and *P. purpurea* was 454-pyrosequenced at the Center for Medical Research, Medical University Graz, Austria employing standard GS FLX Titanium series' protocols. For shotgun-pyrosequencing, total DNA was extracted as above. After complete resuspension in 10 mM Tris-buffer, DNA was column-purified using the NucleoSpin® gDNA Clean-up kit (Macherey-Nagel).

5.5 Sequence assembly, finishing and contig verification

End-sequences of fosmid-clones were trimmed off the vector sequence using SeqMan I (DNA Star) or Geneious v5.5 (Drummond et al. 2009), respectively. Subsequently, a BLAST-search was conducted against custom plastid genes and genome databases as well as plastid protein databases employing the NCBI local BLAST package. Only blast hits with e-values of less than $10e^{-25}$ for BLASTn or $10e^{-10}$ for BLASTx, respectively, and a minimal contiguous alignment length of 100 bp were listed. Sanger shotgun reads were trimmed off vector sequences and cleaned by clipping off regions of Phred-values lower than Q17. Contaminant sequences (e.g. *E. coli*) were removed by blasting against custom contaminant databases. Sanger shotgun reads were assembled using CAP3 (Huang and Madan 1999) with a minimal overlap of 80 bp between two reads, a 95% identity cut-off and default gap penalties. Maximal length of gaps was reduced to 7 bp or less. Shotgun 454 fosmid reads were similarly assembled. A post-assembly of contigs was performed manually using the Phylogenetic Data Editor (PhyDE ®), and/or by using the SeqMan I software or the Geneious Software Package, respectively. Gaps and regions of uncertainty were subsequently closed and verified by PCR amplification and sequencing. Overlapping regions between different fosmids as well as junctions between single copy and the inverted repeat regions were verified via PCR. A list of primers is available from the first author upon request. Plastid genomes sequenced from total genomic DNA were reconstructed from two independent assemblies: 454 raw reads were extracted from the sff-file using a MIRA 3rd party script (Chevreux et al. 1999). Adapters plus the next 10 bp were clipped off. Trimmed reads were assembled de novo under the "accurate" assembly mode. Subsequently, contigs were blasted against custom plastid databases as described above. Positive matches were extracted and post-assembled gene-

by-gene with high stringency in SeqMan I. Gene contigs were pre-annotated in DOGMA and then overlapped manually. Contig joints, regions of uncertainty including microsatellite-like fragments were PCR-verified and confirmed by Sanger sequencing. A second approach involved 454 reads of potential plastid origin were sorted and extracted using NCBI Blast-suite and assembled independently from the remaining read pool using CAP3 (identity cutoff 97%, minimal overlap 90 bp, maximum gap length 5 bp). CAP3 contigs were pre-annotated using DOGMA (Wyman et al. 2004), manually joined, and checked for incongruences with MIRA contigs. Cases of incongruence (mainly satellite regions) were verified by PCR and Sanger-sequencing. Annotation of finished plastid chromosome sequences was carried out in DOGMA with some manual refinement. Reference-assisted assembly of Phelipanche plastomes was performed in MIRA using P. purpurea-fosmids (covering IR and LSC) as reference/backbone sequence.

5.6 Plastid genome analysis and ancestral genome reconstruction

Descriptive statistics (G/C-content, coding/non-coding ratios etc.) of Orobanchaceae plastid genomes were obtained via the PhyDE-plugin SeqState (Kai F. Müller 2005). Selfplots have been computed using the Geneious Software package in order to obtain a rough overview of repeated elements. The REPuter Software package (Kurtz et al. 2001) was employed to find and quantify all forward, reverse, complement and palindromic repeats longer than 20 bp applying a hamming distance of 3; repeats with an Evalue > 0.1 were not considered. We determined the ratio of repeats per genome as the total number of repeats divided by the length of the plastid chromosome in order to assess whether functional genome reduction coincides with an increase in repeats.

The program "CodonW" (http://codonw.sourceforge.net/) was employed for the analyses of codon usage and base pair composition in plastid coding regions. CodonW results have been analyzed using custom R- and Perl scripts. Unless mentioned otherwise, *Nicotiana tabacum* (NC_001879, K. Shinozaki et al. 1986), *Mimulus guttatus* (Mimulus Genome Project, DoE Joint Genome Institute), and *Lindenbergia* were standardly used as core autotrophic groups for analysis of sequence data; *N. tabacum* and *Mimulus* were also employed as outgroup for tree-based analyses. Statistical analyses and hypothesis testing were performed in "R" with appropriate packages available at the CRAN repository. Unless mentioned otherwise, we used nonparametric gamma statistics to analyze and evaluate correlations. Pairwise Wilcoxon tests with sequential alpha-error correction were employed to evaluate differences in A, T, C, and G-distribution, total G/C-content, and G/C-content at different codon positions. In order to evaluate differences among gene specific G/C-contents at different codon positions, unpaired Wilcoxon tests were performed between parasites and an equally sized set of closely related autotrophic taxa.

The latter dataset of autotrophs included eleven lamiid species: *Antirrhinum majus* (GQ996966-GQ997048, Moore et al. 2010), *Atropa belladonna* (NC_004561, Schmitz-Linneweber et al. 2002), *Coffea arabica* (NC_008535, Samson et al. 2007), *Jasminum nudiflorum* (NC_008407, Hae-Lim Lee et al. 2007), *Nerium oleander* (GQ997630-GQ997712, Moore et al. 2010), *Olea europaea* (NC_013707, Mariotti et al. 2010), *Solanum lycopersicon* (DQ347959, Daniell et al. 2006) plus *N. tabacum, Mimulus* and *Lindenbergia*, and *Aucuba japonica* (GQ997049-GQ997131, Moore et al. 2010). *Aucuba japonica*, sister to the remainder lamiid species, was also used as a reference to build genewise distances for boxplots of A/T-variation in photosynthetic taxa and Orobanchaceae. As the genes *infA* and *rps12* were missing in the original *Olea* plastid genome annotation, we successfully BLAST-searched and extracted them from the genome reference sequence (localizations of *infA* in refseq: c82247-82480; localization of *rps12*-exons in refseq: c72298-72411, c99915-99941, and c100477-100708). We removed all introns from *Aucuba, Nerium* and *Antirrhinum* sequences. "CodonW" was used to determine codon-specific nucleotide distributions of all taxa.

For all tree-based analysis, phylogenetic relationships among herein analyzed species were inferred based upon a concatenated dataset of nearly all plastid ribosomal protein genes (except for *rpl22; 23*). Species-specific absence of a marker gene was treated as an indel event and included as binary coded data. Using PAUP 4.0b, a maximum likelihood (ML) tree was reconstructed using the GTR+G+I model selected by the AICc in ModelTest 3.7 (Posada and Crandall 2001). We used four rate categories, and the gamma shape parameters, proportion of invariable sites, nucleotide frequencies and transition rates of the GTR-model were optimized via ML; 1000 bootstrap replicates were run with the same settings. In addition, two runs of each one million generations were run with eight chains each in MrBayes. Chain temperature was set to 0.2, and each chain was sampled every 1000th generations. 10% of trees were discarded as burn-in fraction. Treegraph 2 (Stöver and Müller 2010) was employed to visualize the results.

Based upon the ML-consensus tree, we reconstructed the ancestral genome structure for all Orobanchaceae nodes. In contrast to plant nuclear genomes, the plastid chromosome mainly evolves in a rather prokaryotic fashion where structural mutations by inversion dominate, and duplication and deletion rounds converge to nearly zero. As mentioned above, evolution of plastomes in non-photosynthetic plants is mainly driven by substantial gene loss. Currently available algorithms and software cannot cope with this particular scenario. In order to assess both the history of pseudogenization/gene losses and large-scale reorganization events among Orobanchaceae, we applied a novel two-way strategy based upon two independent likelihood reconstruction methods. First, we reconstructed the gene content and pseudogenization patterns at all internal nodes in the tree for all plastid genes (79 protein-coding genes + 30 tRNAs) using BayesTraits (Pagel et al. 2004). Therefore, we coded all plastid genes in a multistate matrix as functionally

present, pseudogenized or deleted. The probabilities of whether a gene was intact, a pseudogene, or absent at any of the internal nodes have been computed with an unconstrained 6-parameter model; frequencies of state changes have been estimated over the entire dataset. Likelihood ratio tests have been performed to test whether the 6-parameter model is more suitable over 3 or 4-parameter models in which we constrained either reversible rates or all forward parameters to occur at an equal frequency. The 6-parameter model retrieved significant better results than the forward constrained one; no significant differences existed in restraining back changes.50 likelihood attempts were run on the consensus tree using the most-recent-common-ancestor mode (mrca). Treegraph II (Stöver and Müller 2010) was employed to visualize the results.

In a subsequent step of our analysis, the history of large-scale rearrangements among Orobanchaceae-plastomes was traced using Badger v1.02 (Larget et al. 2005). As the raw gene data cannot be used as a primary input for Badger because of dissimilar gene contents, we determined the maximum amount of locally co-linear blocks (LCBs) among all sequenced Orobanchaceae genomes. Therefore, we employed progressiveMauve 2.3.1 (Darling et al. 2010) with a seed weight of 21, and a custom gap-open penalty of -200 to account for "small" gaps caused by localized gene loss. Based upon this alignment, a permutation matrix was constructed and transformed as input for the Badger-program, which was run in the MCMCMC-mode with eight parallel chains of one million cycles each, sampling every tenth tree and ancestral permutations.

6. ACKNOWLEDGEMENTS

The authors would like to thank Monika Ballmann, Karola Maul (University of Bonn), Thomas Münster (MPIZ) and Lena Landherr-Schaeffer (Penn State University) for excellent technical lab support as well as Norman Wickett (Chicago Botanic Garden) for support and discussion of bioinformatic issues. Sincere thanks is due to Klaus Bahr, Wolfram Lobin (Botanical Garden Bonn), Barbara Ditsch (Botanical Garden Dresden), Mats Hjertson (Uppsala University), Kay Kirkman (Joseph W. Jones Ecological Research Center) for cultivation or providing plant material. We thank Volker Knoop (University of Bonn) and Christoph Neinhuis (TU Dresden) for assistance with the realization of experiments. This study received financial support from the Austrian Science Fund (FWF grant 19404 to G.M.S), the German Science Foundation (DFG grant MU2875/2, to K.F.M., QU153/2 to D.Q. and SRXX/X to S.S.R.), and the US National Science Foundation (N.S.F. grants DEB-0120709 and DBI-0701748 to C.W.D.). Financial support to S.W. from the University of Vienna (Austria) and the Botanical Society of America is gratefully acknowledged.

7. AUTHORS' CONTRIBUTIONS

S.W. designed and coordinated the study, conducted all experiments, analyzed the data and drafted the manuscript. G.M.S. conceived of the study, assisted data analysis, contributed to the conceptual layout of this manuscript, and critically revised it. D.Q., K.F.M., and C.W.D. contributed to the design of the study, assisted experiments, and data analyses, and critically revised the manuscript. N.J.W. and Y.Z. contributed wet lab and dry lab work. S.S.R. critically revised the manuscript.

This chapter will be published in a modified version as a research article in a peer-reviewed journal. The tentative author list and title are as follows:

> *Wicke S, Müller KF, Quandt D, dePamphilis CW, Wickett NJ, Zhang Y, Renner SS, and Schneeweiss GM. Broomrape plastid genomes reveal distinct patterns of functional and physical gene deletion under relaxed selective constraints.*

8. REFERENCES

Ahlert D, Ruf S, and Bock R. 2003. Plastid protein synthesis is required for plant development in tobacco. *Proc. Natl. Acad. Sci. USA* **100**: 15730 -15735.

Bennett JR, and Mathews S. 2006. Phylogeny of the parasitic plant family Orobanchaceae inferred from phytochrome A. *Am. J. Bot.* **93**: 1039-1051.

Berg S, Krause K, and Krupinska K. 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.

Bock R. 2007. Structure, function, and inheritance of plastid genomes. In *Cell and Molecular Biology of Plastids*, Vol. 19 of *Topics in Current Genetics*, pp. 29-63, Springer Berlin / Heidelberg. http://dx.doi.org/10.1007/4735_2007_0223.

Bock R, and Timmis JN. 2008. Reconstructing evolution: Gene transfer from plastids to the nucleus. *BioEssays* **30**: 556-566.

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

Chevreux B, Wetter T, and Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics* **99**: 45-56.

Chris Blazier J, Guisinger-Bellian MM, and Jansen RK. 2011. Recent loss of plastid-encoded *ndh* genes within *Erodium* (Geraniaceae). *Plant Mol. Biol.* **76**: 1-10.

Chumley TW, Ferraris JD, Mower JP, Fourcade H. Matthew, Calie PJ, Boore Jeffrey L., and Jansen RK. 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. *Mol. Biol. Evol.* **23**: 2175-2190.

Cosner ME, Raubeson LA, and Jansen RK. 2004. Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. *BMC Evol. Biol.* **4**:27.

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

Darling AE, Mau B, and Perna NT. 2010. progressiveMauve: Multiple genome alignment with gene gain, loss, and rearrangment. *PLoS ONE* **5**: e11147.

Delannoy E, Fujii S, des Francs CC, Brundrett M, and Small I. 2011. Rampant gene loss in the underground orchid *Rhizanthella gardneri* highlights evolutionary constraints on plastid genomes. *Mol. Biol. Evol.* **28**: 2077-2086.

Delavault PM, Russo NM, Lusson NA, and Thalouarn P. 1996. Organization of the reduced plastid genome of *Lathraea clandestina*, an achlorophyllous parasitic plant. *Physiol Plant* **96**: 674 - 682.

Delavault PM, Sakanyan V, and Thalouarn P. 1995. Divergent evolution of two plastid genes, *rbcL* and *atpB*, in a non-photosynthetic parasitic plant. *Plant Mol. Biol.* **29**: 1071 - 1079.

dePamphilis CW, Young ND, and Wolfe AD. 1997. Evolution of plastid gene *rps2* in a lineage of hemiparasitic and holoparasitic plants: Many losses of photosynthesis and complex patterns of rate variation. *Proc. Natl. Acad. Sci. USA* **94**: 7367-7372.

Downie SR, and Palmer JD. 1992. Restriction site mapping of the chloroplast DNA inverted repeat - a molecular phylogeny of the Asteridae. *Ann. Mo. Bot. Gard.* **79**: 266-283.

Drummond AJ, Ashton B, Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T, and Wilson A. 2009. Geneious v5.5. *Available from http://www.geneious.com/*.

Ems SC, Morden CW, Dixon CK, Wolfe KH, dePamphilis CW, and Palmer JD. 1995. Transcription, splicing and editing of plastid RNAs in the nonphotosynthetic plant *Epifagus virginiana*. *Plant Mol. Biol.* **29**: 721 - 733.

Fejes E, Engler D, and Maliga P. 1990. Extensive homologous chloroplast DNA recombination in the pt14 *Nicotiana* somatic hybrid. *Theor. Appl. Genet.* **79**: 28-32.

Fleischmann TT, Scharff LB, Alkatib S, Hasdorf S, Schöttler MA, and Bock R. 2011. Nonessential plastid-encoded ribosomal proteins in tobacco: A developmental role for plastid translation and implications for reductive genome evolution. *Plant Cell*. Doi. 10.1105/tpc.111.088906

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

Gray B, Ahner B, and Hanson M. 2009. Extensive homologous recombination between introduced and native regulatory plastid DNA elements in transplastomic plants. *Transgenic Res.* **18**: 559-572.

Guisinger MM, Kuehl Jennifer V., Boore Jeffrey L., and Jansen RK. 2011. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. *Mol. Biol. Evol.* **28**: 583-600.

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

Hajdukiewicz PTJ, Allison LA, and Maliga P. 1997. The two RNA polymerases encoded by the nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids. *EMBO J* **16**: 4041-4048.

Haraux F, and de Kouchkovsky Y. 1998. Energy coupling and ATP synthase. *Photosynth. Res.* **57**: 231-251.

Hiratsuka J, Shimada Hiroaki, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun C-R, Meng B-Y, et al. 1989. The complete sequence of the rice (*Oryza sativa*) chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. *Mol. Gen. Genet.* **217**: 185-194.

Jackson KE, Habib S, Frugier M, Hoen R, Khan S, Pham JS, de Pouplana LR, Royo M, Santos MAS, Sharma A, et al. 2011. Protein translation in *Plasmodium* parasites. *Trends Parasitol.* **27**: 467-476.

Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, Leebens-Mack JH, Muller KF, Guisinger-Bellian MM, Haberle RC, Hansen AK, et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc. Natl. Acad. Sci. USA* **104**: 19369 - 19374.

Kleine T, Maier UG, and Leister D. 2009. DNA Transfer from Organelles to the Nucleus: The Idiosyncratic Genetics of Endosymbiosis. *Ann. Rev. Plant Biol.* **60**: 115-138.

Knauf U, and Hachtel W. 2002. The genes encoding subunits of ATP synthase are conserved in the reduced plastid genome of the heterotrophic alga *Prototheca wickerhamii*. *Mol. Genet. Genom.* **267**: 492-497.

de Koning AP, and Keeling PJ. 2004. Nucleus-encoded genes for plastid-targeted proteins in *Helicosporidium*: functional diversity of a cryptic plastid in a parasitic alga. *Eukaryot. Cell* **3**: 1198 - 1205.

de Koning AP, and Keeling PJ. 2006. The complete plastid genome sequence of the parasitic green alga *Helicosporidium sp.* is highly reduced and structured. *BMC Biology* **4**: 12.

Krause K. 2011. Piecing together the puzzle of parasitic plant plastome evolution. *Planta* **234**: 647-656.

Krause K, Berg S, and Krupinska K. 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.

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, and Giegerich R. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucl. Acids Res.* **29**: 4633 - 4642.

Larget B, Kadane JB, and Simon DL. 2005. A Bayesian approach to the estimation of ancestral genome arrangements. *Mol Phyl Evol* **36**: 214-223.

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

Logacheva MD, Schelkunov MI, and Penin AA. 2011. Sequencing and analysis of plastid genome in mycoheterotrophic orchid *Neottia nidus-avis*. *Genome Biol. Evol.* **3**: 1296-1303.

Lusson NA, Delavault PM, and Thalouarn P. 1998. The *rbcL* gene from the non-photosynthetic parasite *Lathraea clandestina* is not transcribed by a plastid-encoded RNA polymerase. *Curr. Genet.* **34**: 212-215.

Magee AM, Aspinall S, Rice DW, Cusack BP, Sémon M, Perry AS, Stefanović S, Milbourne D, Barth S, Palmer JD, et al. 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Res.* **20**: 1700-1710.

Maréchal A, and Brisson N. 2010. Recombination and the maintenance of plant organelle genome stability. *New Phytol* **186**: 299-317.

Mariotti R, Cultrera N, Munoz Diez C, Baldoni L, and Rubini A. 2010. Identification of new polymorphic regions and differentiation of cultivated olives (*Olea europaea* L.) through plastome sequence comparison. *BMC Plant Biol.* **10**: 211.

Martin W. 2003. Gene transfer from organelles to the nucleus: Frequent and in big chunks. *Proc. Natl. Acad. Sci. USA* **100**: 8612-8614.

McCarty RE, Evron Y, and Johnson EA. 2000. The chloroplast ATP synthase: A rotary enzyme? *Ann. Rev. Plant Physiol. Plant Mol. Biol.* **51**: 83-109.

McNeal JR, Kuehl J, Boore J, and de Pamphilis C. 2007. Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus *Cuscuta*. *BMC Plant Biol.* **7**: 57.

McNeal JR, Leebens-Mack JH, Arumuganathan K, Kuehl J. V., Boore J. L., and dePamphilis CW. 2006. Using partial genomic fosmid libraries for sequencing complete organellar genomes. *Biotechniques* **41**: 69 - 73.

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

Müller A., Kamisugi Y, Grüneberg R, Niedenhof I, Hörold R., and Meyer P. 1999. Palindromic sequences and A+T-rich DNA elements promote illegitimate recombination in *Nicotiana tabacum*. *J Mol Biol* **291**: 29-46.

Müller KF. 2005. SeqState: Primer design and sequence statistics for phylogenetic DNA datasets. *Appl. Bioinformatics* **4**: 65-69.

Müller P, Li X-P, and Niyogi KK. 2001. Non-photochemical quenching. A response to excess light energy. *Plant Physiol.* **125**: 1558 -1566.

Ogihara Y, Terachi T, and Sasakuma T. 1988. Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species. *Proc. Natl. Acad. Sci. USA* **85**: 8573 -8577.

Pagel M, Meade A, and Barker D. 2004. Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology* **53**: 673 - 684.

Park J-M, Manen J-F, Colwell A, and Schneeweiss GM. 2008. A plastid gene phylogeny of the non-photosynthetic parasitic *Orobanche* (Orobanchaceae) and related genera. *J. Plant Res.* **121**: 365-376.

Pascal AA, Liu Z, Broess K, van Oort B, van Amerongen H, Wang C, Horton P, Robert B, Chang W, and Ruban A. 2005. Molecular basis of photoprotection and control of photosynthetic light-harvesting. *Nature* **436**: 134-137.

Peltier G, and Cournac L. 2002. Chlororespiration. *Ann. Rev. Plant Biol.* **53**: 523-550.

Perry AS, and Wolfe KH. 2002. Nucleotide substitution rates in legume chloroplast DNA depend on the presence of the inverted repeat. J. Mol. Evol. **55**: 501-508.

Posada D, and Crandall KA. 2001. Modeltest v3.06: testing the model of DNA substitution. *Bioinformatics* **14**: 817-818.

Randle CP, and Wolfe AD. 2005. The evolution and expression of RBCL in holoparasitic sister-genera *Harveya* and *Hyobanche* (Orobanchaceae). *Am. J. Bot.* **92**: 1575-1585.

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

Rogalski M, Ruf S, and Bock R. 2006. Tobacco plastid ribosomal protein S18 is essential for cell survival. *Nucl. Acids Res.* **34**: 4537 -4545.

Rogalski M, Schöttler MA, Thiele W, Schulze WX, and Bock R. 2008. Rpl33, a nonessential plastid-encoded ribosomal protein in tobacco, is required under cold stress conditions. *Plant Cell* **20**: 2221-2237.

Sambrook J, and Russell DW. 2001. *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, N.Y.

Samson N, Bausher MG, Lee Seung-Bum, Jansen RK, and Daniell H. 2007. The complete nucleotide sequence of the coffee (*Coffea arabica* L.) chloroplast genome: organization and implications for biotechnology and phylogenetic relationships amongst angiosperms. *Plant Biotechnol. J* **5**: 339-353.

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

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

Schäferhoff B. 2011. Carnivory in Lamiales: Phylogeny, taxonomy, and chloroplast genome evolution. Dissertation, Westfälischen Wilhelms-Universität Münster, Münster.

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

Schneeweiss GM, Colwell A, Park J-M, Jang C-G, and Stuessy TF. 2004. Phylogeny of holoparasitic *Orobanche* (Orobanchaceae) inferred from nuclear ITS sequences. *Mol. Phylogenet. Evol.* **30**: 465-478.

Schwender J, Goffman F, Ohlrogge JB, and Shachar-Hill Y. 2004. Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds. *Nature* **432**: 779-782.

Sears BB, Stoike LL, and Chiu WL. 1996. Proliferation of direct repeats near the *Oenothera* chloroplast DNA origin of replication. *Mol. Biol. Evol.* **13**: 850 -863.

Segall AM, and Roth JR. 1989. Recombination between homologies in direct and inverse orientation in the chromosome of *Salmonella*: intervals which are nonpermissive for inversion formation. *Genetics* **122**: 737 - 747.

Sheppard AE, and Timmis JN. 2009. Instability of plastid DNA in the nuclear genome. *PLoS Genet.* **5**: e1000323.

Shinozaki K., Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, et al. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J.* **5**: 2043 - 2049.

Stöver B, and Müller KF. 2010. TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics* **11**: 7.

Suorsa M, Sirpio S, and Aro E-M. 2009. Towards Characterization of the chloroplast NAD(P)H dehydrogenase complex. *Mol. Plant.* ssp052.

Varco-Merth B, Fromme R, Wang M, and Fromme P. 2008. Crystallization of the c14-rotor of the chloroplast ATP synthase reveals that it contains pigments. *BBA Bioenergetics* **1777**: 605-612.

Wakasugi T, Tsudzuki J, Ito S, Nakashima K, Tsudzuki T, and Sugiura M. 1994. Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the Black Pine *Pinus thunbergii*. *Proc. Natl. Acad. Sci. USA* **91**: 9794 - 9798.

Westwood JH, Yoder JI, Timko MP, and dePamphilis CW. 2010. The evolution of parasitism in plants. *Trends Plant Sci.* **15**: 227-235.

Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, and Quandt D. 2011. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol. Biol.* **76**: 273-297.

Wickett NJ, Zhang Y, Hansen SK, Roper JM, Kuehl Jennifer V., Plock SA, Wolf PG, dePamphilis CW, Boore Jeffrey L., and Goffinet B. 2008. Functional Gene losses occur with minimal size reduction in the plastid genome of the parasitic liverwort *Aneura mirabilis*. *Mol. Biol. Evol.* **25**: 393-401.

Wolfe AD, and dePamphilis CW. 1997. Alternate paths of evolution for the photosynthetic gene *rbcL* in four nonphotosynthetic species of *Orobanche*. *Plant Mol. Biol.* **33**: 965 - 977.

Wolfe KH, Li WH, and Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. USA* **84**: 9054 - 9058.

Wolfe KH, Morden CW, and Palmer JD. 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proc. Natl. Acad. Sci. USA* **89**: 10648-10652.

Wolfe KH, Morden CW, Ems SC, and Palmer JD. 1992. Rapid evolution of the plastid translational apparatus in a nonphotosynthetic plant: loss or accelerated sequence evolution of tRNA and ribosomal protein genes. J. Mol. Evol. **35**: 304 - 317.

Wu F-H, Chan M-T, Liao D-C, Hsu C-T, Lee Y-W, Daniell H, Duvall M, and Lin C-S. 2010. Complete chloroplast genome of *Oncidium* Gower Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. *BMC Plant Biol.* **10**: 68.

Wyman SK, Boore J. L., and Jansen RK. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **20**: 3252 - 3255.

Young ND, and dePamphilis CW. 2005. Rate variation in parasitic plants: correlated and uncorrelated patterns among plastid genes of different function. *BMC Evol. Biol.* **5**: 16.

Zhang Q, and Sodmergen. 2010. Why does biparental plastid inheritance revive in angiosperms? *J Plant Res* **123**: 201-206.

Zoschke R, Nakamura M, Liere K, Sugiura M, Börner T, and Schmitz-Linneweber C. 2010. An organellar maturase associates with multiple group II introns. *Proc. Natl. Acad. Sci. USA* **107**: 3245-3250.

9. SUPPLEMENTAL MATERIAL

Figures

Tables

List of references cited in the supplemental material.

9.1 Figures

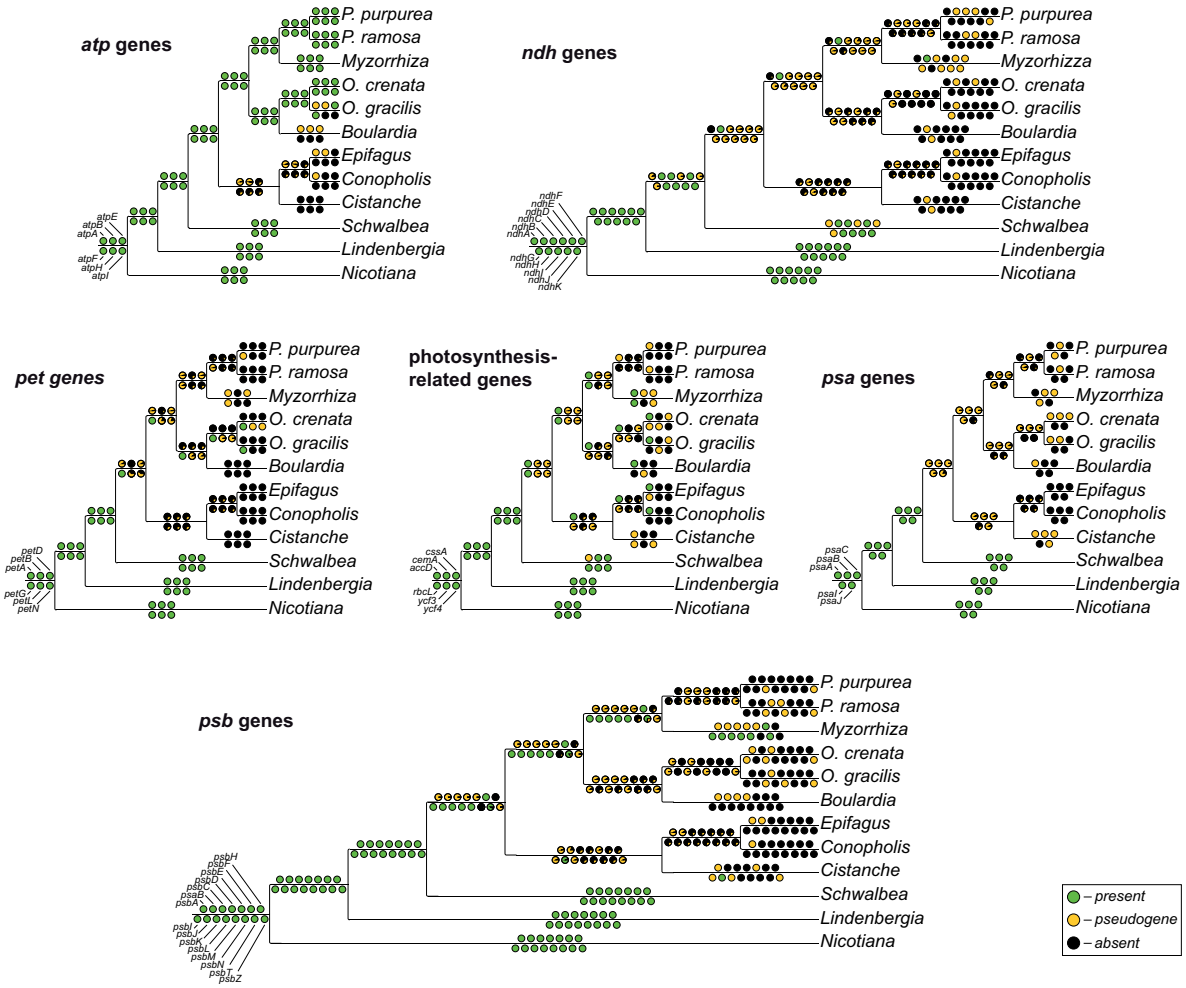


Fig. SIV-1 **Ancestral states for photosynthesis related protein-coding genes.** Pie charts above and below the branches reflect the probabilities of functionality of a certain protein-coding gene (green), pseudogenization (orange) or loss (black); gene IDs are indicated once at the root of each tree.

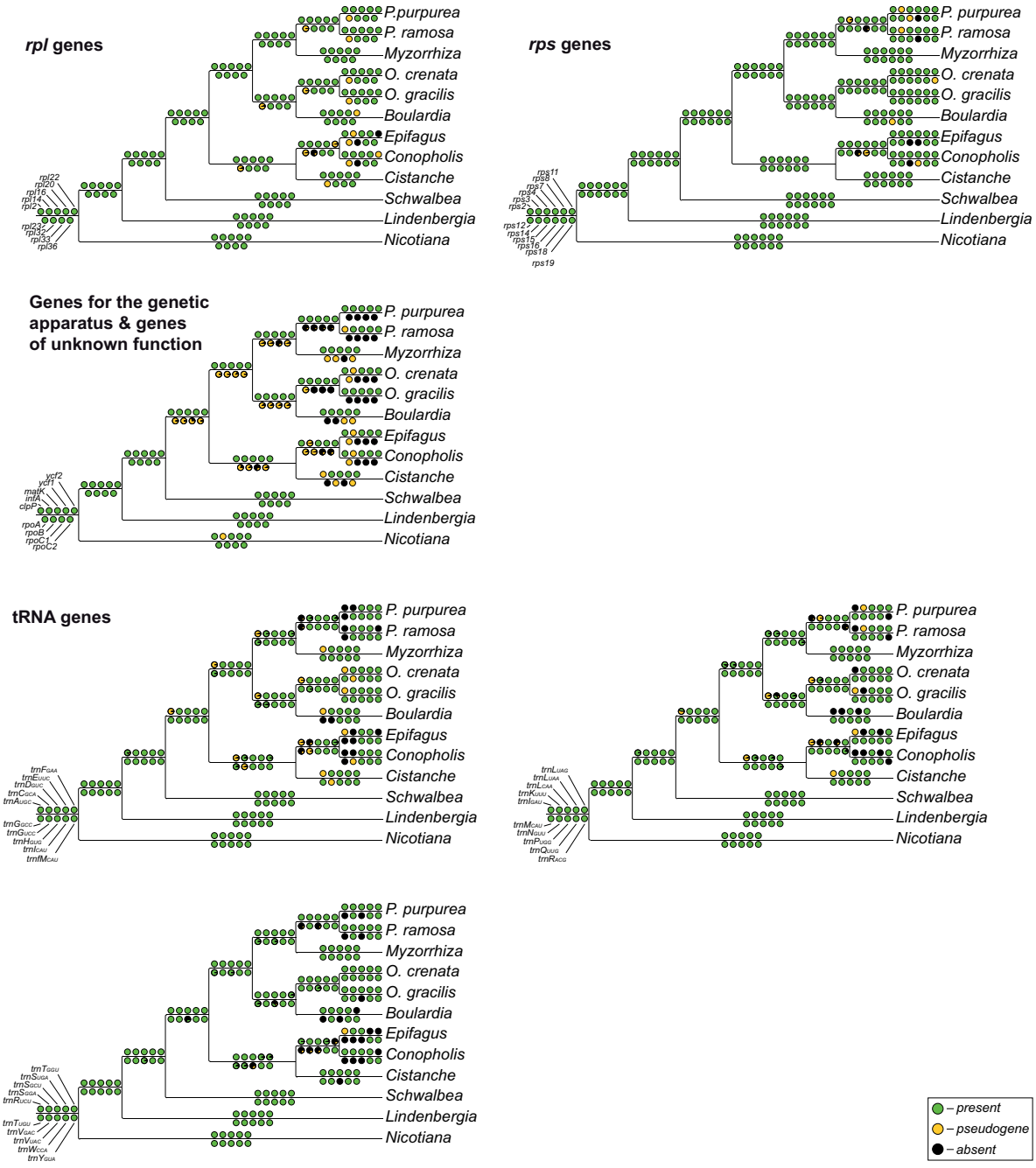


Fig. SIV-2 **Ancestral states for protein coding genes of housekeeping function and all plastid tRNA genes.** Pie charts above and below the branches reflect the probabilities of functionality of a certain tRNA (green), pseudogenization (orange) or loss (black); gene IDs are indicated once at the root of each tree.

9.2 Tables

Table SIV-A **Detailed overview of the gene content of 9 photosynthetic and non-photosynthetic Orobanchaceae.** Presence ("x", green), absence ("lost") or presence as pseudogene (Ψ, gray) of 79 protein coding plastid genes and is summarized for two photosynthetic Orobanchaceae, 8 newly sequenced non-photosynthetic broomrapes and Epifagus (Wolfe et al. 1992). Uncertain assignments of functionality are indicated in red. Classification of genes as putatively as pseudogenes bases solely on evidence from the DNA-level, validation on the RNA-level is currently underway in our labs. Reason for pseudogene assignment is provided accordingly. The table continues over five pages. Abbreviations: *Nic – Nicotiana tabacum, Lin – Lindenbergia philippensis, Sch – Schwalbea americana, Epi – Epifagus virginiana, Con – Conopholis americana, Cis – Cistanche phelypaea, Bou – Boulardia latisquama, Ogr – O. gracilis, Ocr – O. crenata, Myz – Myzorrhiza californica, Ppu – Phelipanche purpurea, Pra – P. ramosa,* prem. stop – premature stop codon, indel – insertions/deletions (>6 bp).

Gene ID	*Lin.*	*Sch*	*Epi*	*Con*	*Cis*	*Bou*	*Ocr*	*Ogr*	*Myz*	*Ppu*	*Pra*
Protein-coding genes											
accD	x	Ψ (5'-truncated)	x	x	Ψ (large indels, several prem. stops)	x	x	x	x	Ψ (5'-truncated)	Ψ (5'-truncated)
atpA	x	x	Ψ	Ψ (5'truncated, large indels, frame shift)	lost	Ψ (5'and 3' truncated, high sequence divergence, frame shifts)	x	Ψ (several prem. stops)	1 prem. stop	x	x
atpB	x	x	Ψ	lost	lost	Ψ (5' and 3' truncated, large indels, prem. stops)	x	Ψ (several prem. stops)	1 prem. stop	1 prem. stop	x
atpE	x	x	lost	lost	lost	Ψ (7 prem. stops, 3'truncated)	x	x	x	x	x
atpF	x	x	lost	lost	lost	lost	stop may be 3'-truncated	1 prem. stop	x	1 prem. stop at 3' (maybe true stop codon)	1 prem. stop at 3' (maybe true stop codon)
atpH	x	x	lost	lost	lost	lost	x	lost	x	x	x
atpI	x	x	lost	lost	lost	lost	x	lost	x	x	x
ccsA		5 prem. stop codons	lost	lost	lost	lost	Ψ (several indels, high sequence divergence, framshifts)	Ψ (several prem. stops, frameshifts, high sequence divergence)	Ψ (large deletions)	lost	lost
cemA	x	x	lost	lost	lost	lost	lost	lost	Ψ (several prem. stops)	lost	lost
clpP	x	one intron lost, no stop codon	x	x	Ψ –(intron 1 too short)	x	x (loss of both introns)	x (loss of both introns)	x (one intron lost)	x (loss of both introns)	Ψ - 20 prem. stops, many indels, introns lost
infA	1 prem. stop	1 prem. stop	x	Ψ (3' and 5' truncated, high sequence divergence)	x	x	Ψ (unclear start/high sequence divergence)	Ψ (unclear start/high sequence divergence	x	x	x
matK	x	x	x	x	unclear start/stop	regulatory elements missing or newly co-transcribed with *rps16*	x	1 prem. stop	x	x	1 prem. stop
ndhA	x	Ψ (truncated -> one exon missing, ends directly in *ndhF*)	lost	lost	lost	lost	lost	lost	lost	lost	lost
ndhB	x	x	Ψ	Ψ (truncated, one exon missing, no intron)	Ψ (large indels, prem. stops, frame shift)	Ψ (3'truncated; several large indels, prem. stops)	Ψ (exon 2 missing, several prem. stops and indels)	Ψ (truncated, one exon missing)	x	Ψ (truncated, exon missing)	lost
ndhC	x	x	lost	lost	lost	lost	lost	lost	Ψ (high sequence divergence, 3' and 5' truncated)	Ψ (5'-truncated)	Ψ (3' and 5' truncated, large indel, truncated operon)

Gene ID	*Lin*	*Sch*	*Epi*	*Con*	*Cis*	*Bou*	*Ocr*	*Ogr*	*Myz*	*Ppu*	*Pra*
ndhD	x	Ψ (several prem. stops, 5'-truncated , several small indels)	lost	lost	lost	lost	Ψ (5'-truncated, prem. stops)	lost	lost	Ψ (3' and 5' truncated)	Ψ (3' and 5' truncated, large indel, truncated operon)
ndhE	x	x	lost	lost	lost	lost	lost	lost	Ψ (multiple prem. stop codons)	lost	lost
ndhF	x	Ψ (multiple fragments, disrupted)	lost	lost	lost	lost	lost	lost	Ψ (large indels, prem. stops)	lost	lost
ndhG	x	Ψ (start codon mutated to atc, several prem. stops, small indels)	lost	lost	lost	lost	lost	Ψ (truncated)	Ψ (multiple prem. stop codons, large deletion)	lost	lost
ndhH	x	x	lost	lost	Ψ (truncated)	Ψ (truncated)	lost	lost	lost	lost	lost
ndhI	1 prem. stop	1 prem. stop	lost	lost	lost	lost	lost	lost	Ψ (5'-truncated ORF)	lost	lost
ndhJ	x	x	lost	lost	lost	lost	lost	lost	Ψ (several frame shifts)	lost	lost
ndhK	x	1 prem. stop	lost	lost	lost	lost	lost	lost	Ψ (5'-truncated)	Ψ (several prem. stops, multiple indels, 5'truncated)	lost
petA	x	x	lost	lost	lost	lost	lost	lost	Ψ (several prem. stops)	lost	lost
petB	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost
petD	x	x	lost	lost	lost	lost	lost	lost	Ψ (lacking/undetectable 5'-exon; 3'truncated)	lost	lost
petG	x	x	lost	lost	lost	lost	x	x	Ψ (high sequence divergence, truncated)	Ψ (large indel, no start codon)	lost
petL	x	x	lost	lost	lost	lost	Ψ (5'truncated)	lost	lost	lost	lost
petN	x	x	lost	lost	lost	lost	Ψ (start codon mutated, prem. stop)	lost	lost	lost	lost
psaA	x	x	lost	lost	Ψ (truncated)	Ψ (large deletions, truncated)	Ψ (large indels, truncated)	Ψ (large indels, truncated)	lost	lost	lost
psaB	x	x	lost	lost	Ψ (truncated)	lost	Ψ (large indels, truncated)	Ψ (large indels, truncated)	Ψ (3' and 5' truncated)	Ψ (3' and 5' truncated)	Ψ (3' and 5' truncated)
psaC	x	x	lost	lost	Ψ (truncated)	lost	Ψ (5' truncated)	lost	Ψ (prem. stop codons, lies in 3'-truncated operon)	lost	lost
psaI	x	x	lost	lost	lost	lost	lost	lost	Ψ (prem. stop, indel. high sequence divergence)	Ψ (5'-truncated)	lost
psaJ	x	x	lost	lost	Ψ (high sequence divergence, unclear stop)	lost	lost	lost	lost	lost	lost
psbA	x	x	Ψ	Ψ (truncated, large indel)	Ψ (large deletions)	Ψ (large deletions, truncated)	Ψ (large deletions, truncated)	lost	Ψ (numerous prem. stops)	lost	lost
psbB	x	x	Ψ	lost	lost	Ψ (truncated)	lost	lost	Ψ (3'-truncated)	lost	lost
psbC	x	x	lost	lost	lost	Ψ (5' and 3'-truncated)	Ψ (5' and 3'-truncated)	Ψ (frame shifted, large indels, prem. stops and unclear start/stop)	Ψ (frame shifts, several prem. stops)	lost	Ψ (truncated and frame shifted)

Gene ID	*Lin*	*Sch*	*Epi*	*Con*	*Cis*	*Bou*	*Ocr*	*Ogr*	*Myz*	*Ppu*	*Pra*
psbD	x	x	lost	lost	lost	Ψ (5' and 3' truncated, 3' ends in *trnE*)	lost	lost	Ψ (several prem. stops, frame shift, small indels)	lost	Ψ (3' and 5' truncated, large indel)
psbE	x	x	lost	lost	Ψ (frame shift, deletion (7 aa)	lost	lost	lost	Ψ (prem. stop, indel. high sequence divergence)	lost	lost
psbF	x	x	lost	lost	lost	lost	lost	lost	x	lost	lost
psbH	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost
psbI	x	x	lost	lost	Ψ (3' truncated, without stop codon --> truncated by trnS-GCU)	lost	Ψ (3'-truncated after prem. stop)	lost	x	lost	lost
psbJ	x	x	lost	lost	1 prem. stop	lost	lost	lost	x	lost	lost
psbK	x	x	lost	lost	Ψ (5'-truncated)	lost	Ψ (5'-truncated; several prem. stops towards 3')	Ψ (unclear start frame shifts and several indels + pre-stops)	x	Ψ (large indels, no 5'-3' truncated, high sequence divergence)	Ψ (3' and 5' truncated, high sequence divergence)
psbL	x	x	lost	lost	lost	lost	lost	lost	x	lost	lost
psbN	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost
psbM	x	x	lost	lost	lost	lost	lost	Ψ (high sequence divergence, may be functional)	1 prem. stop	lost	Ψ (3' and 5' truncated, high sequence divergence)
psbT	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost
psbZ	x	x	lost	lost	Ψ (3' truncated, start codon missing)	lost	Ψ (indels, unclear high sequence divergence)	Ψ (several indels, 3' truncated, prem. stops)	Ψ (5'-3' truncated)	Ψ (5'-and 3' truncated)	Ψ (5'-3' truncated, indels)
rbcL	x	x	Ψ	lost	Ψ (large indels, 3'-truncated)	lost	Ψ (large deletions)	lost	x	lost	lost
rpl14	x	x	Ψ	x	x	x	x	x	x	1 prem. stop, maybe true stop.	x
rpl16	x	x	x	x	x	x	x	x	x	x	x
rpl2	x	x	x	x	x	x	x	x	x	x (loss of intron)	x (loss of intron)
rpl20	x	x	x	1 prem. stop	x	x	x	x	x	x	x
rpl22	x	x	lost	Ψ (multiple large)	x	Ψ (several prem. stops)	x	x	high sequence divergence, many indels; maybe a pseudogene	high sequence divergence, many indels; maybe a pseudogene	high sequence divergence, many indels; maybe a pseudogene
rpl23	x	x	Ψ (Wolfe et al. 1992)	Ψ (start/stop unclear; multiple indels and prem. stop codons)	Ψ (truncated)	x	Ψ (5'-truncated)	Ψ (5'-truncated)	x	Ψ (5'-truncates, prem. stop)	Ψ (3' and 5' truncated)
rpl32	x	x	lost	lost	x	x	x	x	1 prem. stop	x	x
rpl33	x	x	x	x	x	x	x	x	x	x	x
rpl36	x	x	x	x	x	x	x	x	x	x	x
rpoA	x	putative Ψ: high sequence divergence, unclear stop codon	Ψ	Ψ (3' truncated)	lost	lost	Ψ (large deletions, truncated)	lost	Ψ (multiple prem. stop codons)	lost	lost
rpoB	x	x	lost	lost	Ψ (truncated)	lost	lost	lost	Ψ (3' and 5' truncated)	lost	lost
rpoC1	x	putative Ψ: smaller indels, high sequence divergence, prem. stop	lost	lost	lost	Ψ (5' truncated, large indels, several prem. stops)	lost	lost	lost	lost	lost
rpoC2	x	x	lost	lost	Ψ (truncated)	Ψ (5' and 3' truncated)	lost	lost	Ψ (truncated)	lost	lost

Gene ID	*Lin*	*Sch*	*Epi*	*Con*	*Cis*	*Bou*	*Ocr*	*Ogr*	*Myz*	*Ppu*	*Pra*
rps11	x	x	x	x	x	x	x	x	x	x	x
rps12	x	x	x	x	x	x	1 prem. stop	1 prem. stop	x	x (loss of intron)	x (loss of intron, aberrant divergence at 3'-end)
rps14	x	x	x	x	x	x	x	x	x	x	x
rps15	x	x	lost	lost	x	x	x	x	x	2 copies of different length, large indels; high sequence divergence, one copy is truncated	x
rps16	x	x	lost	Ψ (3' and 5' truncated, intron + 5' exon missing)	1 prem. stop	Ψ (large indel, prem. stops codons)	1 prem. stop	Ψ (truncated --> exon 2 missing)	1 prem. stop	lost	lost
rps18	x	x	x	x	x	x	x	x	x	x	x
rps19	x	x	x	x	x	x	Ψ (5'truncated)	x	x	x	x
rps2	x	x	x	x	x	1 prem. stop	x	x	x	x	x
rps3	x	x	x	x (gene start/stop unclear)	x	x	x	x	x	Ψ (large deletion,)	Ψ (large indel,)
rps4	x	x	x	x	x	x	x	x	x	2 prem. stops	x
rps7	x	x	x	x	x	x	x	x	x	x	very large indel)
rps8	x	x	x	x (gene start unclear)	x	x	x	x	x	x	x
ycf1	x	putative sequencing/ assembly error	x	x	putative sequencing/ assembly error	putative sequencing/ assembly error	putative sequencing/ assembly error	putative sequencing/ assembly error	1 prem. stop	putative sequencing/ assembly error	putative sequencing/ assembly error
ycf2	x	x	x	1 prem. stop	x	x	x	18 prem. stops accumulating at 3'; Seq/assembly-error	x	One copy may be functional but several prem. stops. One copy 5'-truncated at IR-LSC-junction	x
ycf3	x	x	lost	lost	lost	Ψ (two exons missing)	Ψ (exon 1 missing, intron too short, large deletions)	Ψ (2 exons missing)	lost	lost	lost
ycf4	x	x	lost	lost	Ψ (3'-truncated)	lost	lost	lost	Ψ (several prem. stops)	lost	lost
Structural RNA genes											
rrn16	x	x	x	x	x	x	x	x	x	x (partial duplication)	x (partial duplication)
rrn23	x	x	x	x	x	x	x	x	x	x	x
rrn4.5	x	x	x	x	x	x	x	x	x	x	x
rrn5	x	x	x	x	x	x	x	x	x	x	x
trnA-UGC	x	x	Ψ	lost	putative Ψ: intron only ~350 bp	Ψ (exon1/ intron missing)	Ψ (intron too short)	Ψ (exon 2 lost)	2 copies: 1 intact; 1 copy is Ψ	lost	lost
trnC-GCA	x	x	lost	lost	x	x	x	x	x	lost	x
trnD-GUC	x	x	x	x	x	x	x	x	x	x	x
trnE-UUC	x	x	x	x	x	x	x	x	x	x	x
trnF-GAA	x	x	lost	x	x	x	x	x	x	x	lost
trnfM-CAU	x	x	x	x	x	x	x	x	x	x	x
trnG-GCC	x	x	lost	lost	x	lost	x	x	x	lost	lost
trnG-UCC	x	x	lost	Ψ (2nd exon missing, no traces of intron)	Ψ (exon 2 lost)	lost	Ψ (intron too short)	x (loss of intron)	x	x (loss of intron)	x (loss of intron)

Gene ID	*Lin*	*Sch*	*Epi*	*Con*	*Cis*	*Bou*	*Ocr*	*Ogr*	*Myz*	*Ppu*	*Pra*
trnH-GUG	x	x	x	x	x	x	x	x	x	x	x
trnI-CAU	x	x	x	x	x	x	x	x	x	x	x
trnI-GAU	x	x	Ψ	lost	Ψ (exon1 and intron missing)	lost	lost	Ψ (intron too short)	2 copies -> 1 apparently intact; 1 copy is Ψ	lost	lost
trnK-UUU	x	x	lost	lost	x	lost	x	lost	x	Ψ (exon1 and intron missing)	Ψ (lacking 3'-exon)
trnL-CAA	x	x	x	x	x	x	x	x	x	x	x
trnL-UAA	x	x	lost	lost	x	lost	x	x	x	x	x
trnL-UAG	x	x	x	x	x	x	x	x	x	x	x
trnM-CAU	x	x	x	x	x	x	x	x	x	x	x
trnN-GUU	x	x	x	x	x	x	x	x	x	x	x
trnP-UGG	x	x	x	x	x	x	x	x	x	x	x
trnQ-UUG	x	x	x	x	x	x	x	x	x	x	x
trnR-ACG	x	x	x	lost	x	x	x	x	x	lost	lost
trnR-UCU	x	x	Ψ	x	x	x	x	x	x	x	x
trnS-GAA	x	x	x	x	x	x	x	x	x	x	x
trnS-GCU	x	x	x	x	x	x	x	x	x	x	x
trnS-UAG	x	x	lost	x	x	x	x	x	x	x	x
trnT-GGU	x	x	lost	lost	x	lost	x	x	x	x	x
trnT-UGU	x	x	lost	lost	x	lost	x	x	x	lost	lost
trnV-GAC	x	x	lost	lost	x	x	x	x	x	x	x
trnV-UAC	x	x	lost	lost	lost	lost	x	lost	x	lost	lost
trnW-CCA	x	x	x	x	x	x	x	x	x	x	x
trnY-GUA	x	x	x	x	x	x	x	x	x	x	x

Further remarks:

Expression of genes is currently validated in our labs, and must therefore be treated as preliminary. The current classification of genes as putatively functional or as pseudogenes bases solely on evidence from the DNA-level.

Schwalbea: *yc1* – extreme sequence divergence and therefore excluded from codon usage analyses (CU). *ccsA* - putative seq/assembly-error; excluded from CU.

Epifagus: matK – if annotated as in Young & dePamphilis 2000 it contains but would contains 37; annotated as in Wolfe et al. 1992 no stops, but ORF ~ 200bp shorter than in other clades of holoparasitic Orobanchaceae. --> "Wolfe-"sequences was used for CU-analysis.

Conopholis: matK – if annotated as suggested in Young & dePamphilis 2000 it contains 37; annotated as in Wolfe et al. 1992 it contains 6 premature stops accumulating at 3'-end after homopolymeric stretch, but ORF ~ 146bp shorter than in other clades of holoparasitic Orobanchaceae. The "Wolfe-"sequence was used for CU-analysis. Remaining stops may be due to sequencing/assembly errors; amino acid sequence is highly similar to *Epifagus*.

Cistanche: *matK* – contains 5 premature stops accumulating at 3'end maybe alternative gene end or may be due to sequencing/assembly errors after homopolymeric stretch. ORF lengths according to either Young and dePamphilis 2000 or Wolfe etal. 1992 may be possible, which distinguishes *Cistanche* from *Epifagus*/*Conopholis*. CU has been performed with the "long" ORF.

Boulardia: ycf1/ycf2 may be pseudogenes – several premature stop codons detected although re-sequenced with Sanger, but still sequencing/assembly cannot be excluded.

Myzorrhiza: *rpl22* – maybe a pseudogene due to large indels and high substitution rate, assigned functional in ASR and included in CU.

Phelipanche purpurea: *atpA*: contains 3 untranslatable codons; *atpH*: 1 untranslatable codon; *matK* – 3 untranslatable codons.

P. ramosa: *matK* – 1 premature stop, 3 untranslatable codons; *atpF* – unclear stop codon 3 potential stops in a row, current annotation assume the first stop as the true stop codon; *rpl2* – 4 premature stops; *ycf1* – fragmented, gene start/stop is unclear.

Table SIV-B **Results of Mann-Whitney-U tests evaluating differences in GC-content of coding regions between non-parasites and parasites.** Differences of GC-content at the first, second, and the third codon position (GC1, GC2, GC3) were evaluated by pairwise Wilcoxon tests between photosynthetic plants and Orobanchaceae. Differences of nucleotide composition per codon position were compared in the same way. Asterisks mark the significance levels (≤0.05*, <0.01**, <0.001***). The table continues the next page. Abbreviations: *Auc – Aucuba japonica, Nic – Nicotiana tabacum, Lin – Lindenbergia philippensis, Sch – Schwalbea americana, Epi – Epifagus virginiana, Con – Conopholis americana, Cis – Cistanche phelypaea, Bou – Boulardia latisquama, Ogr – O. gracilis, Ocr – O. crenata, Myz – Myzorrhiza californica, Ppu – Phelipanche purpurea, Pra – P. ramosa,* prem. stop – premature stop codon; NA – not tested.

	Nic **p-value**	***Mim*** **p-value**	***Lin*** **p-value**	***Schw*** **p-value**	***Epi*** **p-value**	***Con*** **p-value**	***Cis*** **p-value**	***Bou*** **p-value**	***Ocr*** **p-value**	***Ogr*** **p-value**	***Myz*** **p-value**	***Ppu*** **p-value**	***Pra*** **p-value**
#GC-content													
Auc	1.000	0.331	1.000	1.000	0.007**	0.001**	0.007**	0.001**	<0.001***	0.001**	0.007**	<0.001***	<0.001***
Nic	NA	0.034*	1.000	1.000	0.006**	0.001**	0.002**	0.001**	<0.001***	<0.001***	0.004**	<0.001***	<0.001***
Mim	NA	NA	0.042*	0.001**	0.004**	0.001**	0.004**	0.001**	<0.001***	<0.001***	0.029*	<0.001***	<0.001***
Lin	NA	NA	NA	0.360	0.008**	0.002**	0.003**	0.003**	0.003**	<0.001***	0.085	0.003**	0.003**
Sch	NA	NA	NA	NA	0.002**	0.001**	0.001**	0.001**	<0.001***	<0.001***	0.001**	<0.001***	<0.001***
#GC1													
Auc	0.035	1.000	1.000	1.000	0.003**	0.001**	0.073	0.001**	0.001**	0.001**	0.093	<0.001***	<0.001***
Nic	NA	0.063	0.063	0.063	0.003**	0.001**	0.040*	0.002**	0.001**	<0.001***	0.008**	<0.001***	0.001**
Mim	NA	NA	0.811	0.811	0.001**	0.001**	0.007**	0.001**	0.001**	0.001**	0.028*	<0.001***	<0.001***
Lin	NA	NA	NA	0.375	0.001**	<0.001***	0.005**	<0.001***	<0.001***	0.001**	0.005**	<0.001***	<0.001***
Sch	NA	NA	NA	NA	0.004**	<0.001***	0.048*	0.004**	0.001**	<0.001***	0.048*	<0.001***	<0.001***
#GC2													
Auc	0.309	0.309	0.309	0.097	0.001**	0.002**	0.003**	0.003**	<0.001***	<0.001***	0.010*	<0.001***	<0.001***
Nic	NA	1.000	1.000	1.000	0.003**	0.003**	0.004**	0.005**	<0.001***	<0.001***	0.059	<0.001***	<0.001***
Mim	NA	NA	0.891	0.891	0.001**	0.001**	0.003**	0.001**	<0.001***	<0.001***	0.151	<0.001***	<0.001***
Lin	NA	NA	NA	0.280	0.001**	0.001**	0.001**	0.005**	<0.001***	<0.001***	0.022*	<0.001***	<0.001***
Sch	NA	NA	NA	NA	0.004**	0.003**	0.030*	0.030*	<0.001***	<0.001***	0.174	<0.001***	<0.001***
#GC3													
Auc													
Nic	NA	<0.001***	0.662	0.170	0.011*	0.076	0.170	0.021*	0.002**	0.003**	0.294	0.003**	0.015*
Mim	NA	NA	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***
Lin	NA	NA	NA	0.041	0.002**	0.005**	0.046*	0.003**	0.001**	0.007**	0.087	0.001**	0.003**
Sch	NA	NA	NA	NA	0.002**	0.002**	0.004**	0.002**	<0.001***	0.001**	0.002**	<0.001***	<0.001***
#A1													
Nic	NA	0.048*	0.072	0.040*	0.002**	0.002**	0.088	0.008**	0.006**	0.006**	0.072	<0.001***	0.002**
Mim	NA	NA	0.260	0.547	0.002**	0.002**	0.157	0.006**	0.010*	0.002**	0.207	<0.001***	0.001**
Lin	NA	NA	NA	0.300	0.002**	0.001**	0.009**	0.001**	0.002**	0.002**	0.040*	<0.001***	<0.001***
Sch	NA	NA	NA	NA	0.011**	0.004**	0.016*	0.011*	0.003**	0.002**	0.165	<0.001***	<0.001***
#A2													
Nic	NA	1.000	1.000	1.000	0.153	0.067	0.512	0.722	0.004**	0.012*	1.000	<0.001***	<0.001***
Mim	NA	NA	1.000	0.615	0.010*	0.014*	0.268	0.615	0.005**	0.004**	1.000	<0.001***	<0.001***
Lin	NA	NA	NA	0.714	0.027*	0.011*	0.199	0.956	0.005**	0.005**	0.956	<0.001***	<0.001***
Sch	NA	NA	NA	NA	0.562	0.764	1.000	1.000	0.002**	0.014*	1.000	0.004**	0.004**
#A3													
Nic	NA	0.742	1.000	0.126	0.024	0.434	0.411	0.104	0.031*	0.008**	1.000	0.009**	0.015*
Mim	NA	NA	0.108	0.007**	0.108	0.477	0.477	0.063	0.051	0.003**	0.798	0.002**	0.005**
Lin	NA	NA	NA	0.058	0.016*	0.050*	0.050*	0.008**	0.004**	0.001**	0.395	<0.001***	0.001**
Sch	NA	NA	NA	NA	0.022	0.116	0.116	0.012*	0.004**	0.002**	0.116	0.001**	0.002**
#T1													
Nic	NA	1.000	0.813	0.465	0.038*	0.021*	0.550	0.071	0.022*	0.092	0.550	0.020*	0.047*
Mim	NA	NA	0.337	0.767	0.007**	0.007**	0.119	0.012*	0.003**	0.024*	0.119	0.003**	0.007**
Lin	NA	NA	NA	0.631	0.015*	0.016*	0.577	0.070	0.015*	0.034*	1.000	0.020*	0.066
Sch	NA	NA	NA	NA	0.061	0.054	0.305	0.305	0.143	0.085	0.305	0.143	0.231

	Nic p-value	*Mim* p-value	*Lin* p-value	*Schw* p-value	*Epi* p-value	*Con* p-value	*Cis* p-value	*Bou* p-value	*Ocr* p-value	*Ogr* p-value	*Myz* p-value	*Ppu* p-value	*Pra* p-value
#T2													
Nic	NA	1.000	1.000	1.000	0.005**	0.015*	0.053	0.009**	0.001**	0.144	0.155	0.224	0.167
Mim	NA	NA	1.000	1.000	0.003**	0.024*	0.010*	0.003**	0.003**	0.029*	0.153	0.124	0.024*
Lin	NA	NA	NA	0.961	0.004**	0.018**	0.020*	0.003**	0.002**	0.020*	0.069	0.096	0.038*
Sch	NA	NA	NA	NA	0.008**	0.008**	0.018*	0.002**	0.001**	0.008**	0.048*	0.018*	0.018*
#T3													
Nic	NA	0.247	0.757	0.757	0.027*	0.357	0.757	0.357	0.005**	0.238	0.757	0.247	0.357
Mim	NA	NA	0.426	0.004	0.027*	1.000	1.000	1.000	0.027*	0.937	1.000	0.937	0.937
Lin	NA	NA	NA	0.124	0.008**	0.125	0.125	0.096	<0.001***	0.037*	0.125	0.125	0.125
Sch	NA	NA	NA	NA	0.003**	0.004**	0.013*	0.004**	<0.001***	0.004**	0.013*	0.005**	0.013*
#C1													
Nic	NA	0.256	0.583	1.000	0.020*	0.013*	0.256	0.025*	0.026*	0.044*	0.004**	<0.001***	0.001**
Mim	NA	NA	0.084	0.337	0.007**	0.005**	0.337	0.009**	0.337	0.078	0.034*	<0.001***	0.001**
Lin	NA	NA	NA	0.557	0.004**	0.001**	0.017*	0.001**	0.013*	0.017*	0.003**	<0.001***	<0.001***
Sch	NA	NA	NA	NA	0.015*	0.015*	0.056	0.049*	0.049*	0.015*	0.015*	<0.001***	<0.001***
#C2													
Nic	NA	1.000	1.000	0.346	0.005**	0.005**	0.069	0.209	0.037*	0.041*	0.219	0.006**	0.004**
Mim	NA	NA	0.373	0.089	0.001**	0.001**	0.022*	0.022*	0.002**	0.001**	0.020*	0.002**	0.002**
Lin	NA	NA	NA	0.047*	0.001**	0.001**	0.010*	0.020*	0.003**	0.003**	0.010*	0.002**	0.003**
Sch	NA	NA	NA	NA	0.002**	0.004**	0.399	0.322	0.006**	0.009**	0.399	0.001**	0.003**
#C3													
Nic	NA	0.136	0.909	0.909	0.148	0.178	0.909	0.909	0.012*	0.025*	0.498	0.001**	0.178
Mim	NA	NA	0.711	<0.001***	0.131	0.421	1.000	1.000	0.015*	0.112	1.000	0.001**	0.421
Lin	NA	NA	NA	0.034*	0.010*	0.034*	0.039*	0.193	0.001**	0.002**	0.193	0.001**	0.016**
Sch	NA	NA	NA	NA	0.001**	0.006**	0.006**	0.008**	<0.001***	0.002**	0.006	<0.001***	0.002**
#G1													
Nic	NA	0.236	0.236	0.011*	0.014*	0.006**	0.100	0.006**	0.006**	0.006**	0.236	0.011*	0.006**
Mim	NA	NA	0.463	0.016*	0.004**	0.005**	0.005**	0.001**	0.001**	0.001**	0.088	0.001**	0.001**
Lin	NA	NA	NA	0.137	0.006**	0.004**	0.009**	0.003**	0.002**	0.003**	0.349	0.002**	0.002**
Sch	NA	NA	NA	NA	0.167	0.134	0.209	0.167	0.007**	0.010*	0.747	0.002**	0.001**
#G2													
Nic	NA	1.000	1.000	1.000	0.035*	0.024*	0.021*	0.134	0.001**	0.001**	0.473	0.016*	0.020*
Mim	NA	NA	1.000	1.000	0.038*	0.026*	0.045*	0.005**	0.003**	<0.001***	0.924	0.001**	0.002**
Lin	NA	NA	NA	0.875	0.030*	0.027*	0.030*	0.041*	0.002**	0.001**	0.875	0.002**	0.005**
Sch	NA	NA	NA	NA	0.099	0.050*	0.062	0.099	0.001**	0.001**	0.555	0.001**	0.001**
#G3													
Nic	NA	1.000	1.000	0.173	0.030	0.087	0.354	0.011*	0.019*	0.031*	1.000	0.034*	0.022*
Mim	NA	NA	0.146	0.002**	0.049	0.146	0.254	0.016*	0.032*	0.045*	0.964	0.124	0.003**
Lin	NA	NA	NA	0.070	0.007**	0.010*	0.070	0.001**	0.003**	0.010*	0.185	0.007**	0.001**
Sch	NA	NA	NA	NA	0.010	0.008**	0.067	0.004**	0.001**	0.003**	0.067	0.004**	0.001**

Table SIV-C **Codon Usage in photosynthetic and non-photosynthetic Orobanchaceae.** The proportion (in %) of used codons for all amino acids (AA) and stop codons is summarized for photosynthetic and non-photosynthetic Orobanchaceae and *Nicotiana*. The preferred codon is highlighted in blue. The table continues the next page. Abbreviations: *Nic – Nicotiana tabacum, Lin – Lindenbergia philippensis, Sch – Schwalbea americana, Epi – Epifagus virginiana, Con – Conopholis americana, Cis – Cistanche phelypaea, Bou – Boulardia latisquama, Ogr – O. gracilis, Ocr – O. crenata, Myz – Myzorrhiza californica, Ppu – Phelipanche purpurea, Pra – P. ramosa,* prem. stop – premature stop codon; NA – not tested; TER – stop codon.

		Proportion of codon usage (in %)													
AA	**+ Codon**	***Nic***	***Mim***	***Lin***	***Sch***	***Epi***	***Con***	***Cis***	***Bou***	***Ocr***	***Ogr***	***Myz***	***Phe***	***Pra***	**remarks**
Phe	**UUU**	66.7	68.8	67.9	64.6	78.6	76.8	67.7	79.6	67.2	74.5	66.7	82.1	70.2	
	UUC	33.3	31.2	32.1	35.4	21.4	23.2	32.3	20.4	32.8	25.5	33.3	17.9	29.8	lost in *Epi, Pra*
Leu	**UUA**	32.5	32.5	31.7	29.8	35.6	34.7	27.5	38.0	32.7	38.6	30.1	41.7	36.4	lost/pseudogenized in *Epi, Con, Bou*
	UUG	20.1	19.9	19.7	22.2	17.5	17.8	23.0	20.2	21.8	23.0	19.9	20.2	18.8	
	CUU	21.6	21.3	21.3	20.3	21.1	19.5	22.4	18.4	21.8	17.8	22.8	18.4	20.4	
	CUC	6.9	5.9	5.7	6.8	6.3	5.7	5.5	4.0	4.3	2.2	6.1	2.7	5.7	
	CUA	12.1	13.1	13.8	12.8	10.7	12.6	12.8	9.4	12.0	10.6	12.5	9.4	11.6	
	CUG	6.2	6.5	6.8	7.4	5.6	6.2	6.7	6.5	5.6	4.4	6.9	4.9	5.0	
Ile	**AUU**	49.5	49.8	49.9	50.8	43.0	42.7	48.7	52.8	51.8	49.0	49.6	53.3	50.2	
	AUC	20.1	19.3	20.3	20.6	12.3	14.8	18.5	12.2	15.5	11.8	19.0	7.8	12.3	lost/pseudogenized in all holoparasites but *Myz*
	AUA	30.4	30.9	29.8	28.6	44.7	42.5	32.8	35.1	32.7	39.2	31.4	38.9	37.5	
Met	**AUG**	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	
Val	**GUU**	37.1	38.0	38.6	36.0	37.9	37.8	35.6	36.0	39.3	36.4	35.7	36.3	41.7	
	GUC	11.9	11.6	11.4	12.2	12.1	9.1	12.0	12.2	9.1	9.7	11.0	9.7	10.2	lost in *Epi, Con*
	GUA	38.1	37.7	37.1	37.5	38.2	35.7	34.1	37.6	40.1	41.0	40.7	42.3	35.5	lost/pseudogenized in all holoparasites but *Myz, Ocr*
	GUG	13.0	12.7	13.0	14.3	11.8	17.4	18.4	14.3	11.5	12.8	12.7	11.6	12.7	
Tyr	**UAU**	80.3	81.4	82.0	81.6	82.5	79.8	83.6	86.9	85.3	86.6	81.4	82.2	82.7	
	UAC	19.7	18.6	18.0	18.4	17.5	20.2	16.4	13.1	14.7	13.4	18.6	17.8	17.3	
His	**CAU**	77.2	77.5	77.9	77.4	74.7	80.2	79.3	80.2	79.2	80.6	77.1	78.8	77.4	
	CAC	22.8	22.5	22.1	22.6	25.3	19.8	20.7	19.8	20.8	19.4	22.9	21.3	22.6	
Gln	**CAA**	75.6	76.9	76.3	73.7	81.2	79.9	77.9	85.3	79.5	82.2	79.4	82.7	76.4	
	CAG	24.4	23.1	23.7	26.3	18.8	20.1	22.1	14.7	20.5	17.8	20.6	17.3	23.6	
Asn	**AAU**	77.0	78.7	77.4	76.4	80.3	79.6	74.7	75.9	81.3	83.0	78.5	79.4	78.0	
	AAC	23.0	21.3	22.6	23.6	19.7	20.4	25.3	24.1	18.7	17.0	21.5	20.6	22.0	
Lys	**AAA**	75.5	76.8	76.3	74.9	81.1	80.3	75.5	80.1	78.3	81.4	77.4	84.1	79.6	lost/pseudogenized in *Epi, Con, Bou, Ogr, Ppu, Pra*
	AAG	24.5	23.2	23.7	25.1	18.9	19.7	24.5	19.9	21.7	18.6	22.6	15.9	20.4	
Asp	**GAU**	79.6	81.2	81.1	78.7	80.2	82.4	82.0	79.8	83.2	76.7	80.4	83.4	82.3	
	GAC	20.4	18.8	18.9	21.3	19.8	17.6	18.0	20.2	16.8	23.3	19.6	16.6	17.7	
Glu	**GAA**	75.6	77.7	77.2	73.8	75.9	76.2	73.6	77.9	73.1	75.6	73.6	78.4	73.9	
	GAG	24.4	22.3	22.8	26.2	24.1	23.8	26.4	22.1	26.9	24.4	26.4	21.6	26.1	
Ser	**UCU**	29.7	29.2	29.4	29.1	26.1	22.1	27.5	26.7	30.1	31.5	27.6	34.2	31.3	
	UCC	15.1	14.7	15.7	17.1	13.4	14.8	17.1	13.9	15.4	14.1	16.8	12.1	13.0	
	UCA	19.5	19.1	19.1	17.5	24.3	25.7	21.3	19.4	21.5	20.4	19.8	17.9	21.0	lost in *Epi*
	UCG	9.4	9.8	9.5	10.5	8.9	9.1	9.8	5.9	9.1	5.6	9.6	8.6	8.5	
	AGU	20.7	21.7	21.2	20.2	23.0	23.4	19.5	29.5	20.5	25.6	21.6	24.6	23.9	
	AGC	5.8	5.5	5.2	5.5	4.3	4.8	4.8	4.5	3.3	3.0	4.5	2.6	2.3	

Proportion of codon usage (in %)

AA + Codon		*Nic*	*Mim*	*Lin*	*Sch*	*Epi*	*Con*	*Cis*	*Bou*	*Ocr*	*Ogr*	*Myz*	*Phe*	*Pra*	**remarks**
Pro	**CCU**	38.9	38.7	38.4	37.9	35.5	39.2	31.0	39.7	40.8	41.4	38.7	37.6	35.8	
	CCC	18.7	19.6	18.7	20.2	23.8	27.4	23.8	27.3	22.4	19.3	21.2	21.3	23.7	
	CCA	28.9	29.0	28.2	25.5	31.9	23.2	31.4	28.1	26.0	32.1	27.6	30.9	28.9	
	CCG	13.4	12.8	14.6	16.4	8.9	10.3	13.8	5.0	10.8	7.1	12.5	10.1	11.6	
Thr	**ACU**	39.3	42.1	40.6	39.7	38.4	36.8	38.1	38.0	40.8	38.1	39.6	43.4	37.7	
	ACC	19.6	18.3	18.6	21.2	18.7	18.4	18.8	22.9	17.6	20.3	20.2	20.6	23.0	lost in *Epi, Con, Bou*
	ACA	30.5	29.2	28.6	27.8	32.9	31.9	31.2	29.1	29.8	32.0	29.7	28.9	28.4	lost in *Epi, Con, Bou, Ppu, Pra*
	ACG	10.6	10.4	12.2	11.3	10.0	12.9	11.9	10.1	11.8	9.6	10.5	7.0	10.9	
Ala	**GCU**	44.7	44.6	44.2	41.9	36.7	37.7	37.7	33.5	43.9	38.0	40.5	40.8	40.8	
	GCC	17.2	15.4	17.0	17.7	16.7	19.4	17.2	18.8	15.8	17.6	19.3	18.7	21.7	
	GCA	28.0	28.2	27.4	27.4	36.3	35.6	35.8	38.2	29.8	36.6	28.0	32.0	29.6	lost/pseudogenized in all holoparasites but *Myz*
	GCG	10.0	11.9	11.4	12.9	10.4	7.3	9.3	9.4	10.5	7.9	12.2	8.5	8.0	
Cys	**UGU**	75.2	76.4	76.5	77.3	74.5	73.5	77.0	72.0	83.0	79.6	75.6	79.2	70.8	
	UGC	24.8	23.6	23.5	22.7	25.5	26.5	23.0	28.0	17.0	20.4	24.4	20.8	29.2	lost in *Epi, Con, Ppu*
Trp	**UGG**	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	
Arg	**CGU**	22.0	22.0	21.1	21.4	19.5	17.0	19.0	23.0	22.4	26.6	20.1	23.8	20.8	lost in *Con, Ppu, Pra*
	CGC	6.3	6.8	7.3	8.7	6.9	7.0	5.8	7.1	5.0	4.3	6.1	6.7	7.6	
	CGA	25.1	21.6	22.4	23.2	16.8	18.4	21.5	20.8	21.8	22.0	19.3	19.0	19.6	
	CGG	7.3	8.1	8.4	8.7	7.1	6.4	7.7	5.0	7.0	6.6	7.9	7.0	6.9	
	AGA	28.9	31.4	31.2	28.5	36.5	36.2	32.3	36.0	33.5	31.5	35.2	35.0	33.7	pseudogenized in *Epi*
	AGG	10.4	10.1	9.6	9.4	13.2	14.9	13.7	8.1	10.4	9.0	11.5	8.4	11.4	
Gly	**GGU**	32.3	32.7	32.5	32.2	33.4	30.8	28.8	30.6	32.8	32.6	31.4	36.3	33.1	
	GGC	11.7	10.8	10.5	11.5	7.3	5.9	7.5	9.9	8.5	8.4	8.5	8.9	10.4	
	GGA	38.8	39.3	39.3	37.6	40.7	40.3	41.8	41.9	39.8	39.6	40.3	40.9	39.5	lost/pseudogenized in *Epi, Con, Cis, Bou, Ocr*
	GGG	17.1	17.1	17.7	18.7	18.5	23.0	21.9	17.6	18.9	19.4	19.8	13.9	17.1	
TER	**UAA**	53.8	61.5	53.1	47.3	57.1	63.3	64.5	82.6	70.6	73.1	54.2	75.0	76.7	
	UAG	23.1	17.9	24.7	31.1	28.6	30.0	22.6	17.4	14.7	23.1	20.8	15.6	10.0	
	UGA	23.1	20.5	22.2	21.6	14.3	6.7	12.9	0.0	14.7	3.8	25.0	9.4	13.3	

Table SIV-D **Vicinity of non-essential broomrape plastid genes to conserved genic elements in photosynthetic and non-photosynthetic Orobanchaceae ancestors.** Given a circular genome map with trnH-GUG as origin (+1), the distance of non-essential gene is estimate to the next conserved gene in counterclockwise (CCW) and clockwise (CW) direction. Essential genes (i.e. those that are universally present in broomrape plastomes) are indicated in red. Loss of a gene in an ancestor is indicated by an "x". If a non-essential gene was deleted between two nonessentials, the size of the deleted gene plus an additional 50bp (25bp for each CW and CCW) was substracted assuming that gene deletion also affects surrounding regions. The gene distances from the *Lindenbergia* plastome were used as reference for the putative Orobanchaceae ancestor. The table continues over three pages. Abbreviations: Lin – *Lindenbergia philippensis*, Holo – ancestor of the broomrape-clade (sensu Bennett and Mathews 2006) , EpCoCis – *Epifagus*/*Conopholis*/*Cistanche* -ancestor, OroBou – *Boulardia*/*Orobanche*-ancestor, MyzPhe – *Myzorrhiza*/*Phelipanch*-ancestor, *EpCo* – *Epifagus*/*Conopholis-ancestor*, *Oro* – *Orobanche*-ancestor, *Phe* – *Phelipanche*-ancestor.

	Distance to next conserved gene (in bp)															
	Lin		*Holo*		*EpCoCis*		*OroBou*		*MyzPhe*		*EpCo*		*Oro*		*Phe*[#]	
Gene-ID	CCW	CW	CCW	CW	CCW	CW	CCW	CW	CCW	CW	CCW	CW	CCW	CW	CCW	CW
								trnH-GUG								
psbA	631	501	631	501	631	501	631	501	631	501	631	476	631	501	x	x
trnK-UUU	1,929	236	1,929	236	1,929	236	1,929	236	1,929	236	x	x	1,929	236	871	236
								matK								
trnK-UUU	706	3,183	706	3,183	706	3,183	706	3,183	706	3,183	x	x	706	3,183	706	2,062
rps16	1,656	1,148	1,656	1,148	1,656	1,148	1,656	1,148	1,656	1,148	1,620	1,148	1,656	1,148	x	x
								trnQ-UUG								
psbK	336	634	336	634	336	634	336	634	336	634	x	x	336	634	336	524
psbI	933	118	933	118	933	118	933	118	933	118	x	x	933	118	x	x
								trnS-GCU								
trnG-UCC	755	5,933	755	5,933	755	2,231	755	5,933	755	5,933	755	2,231	755	5,933	x	x
trnR-UCU	1,723	5,651	1,723	5,651	1,723	1,949	1,723	5,651	1,723	5,651	1,723	1,949	1,723	5,651	966	5,651
atpA	1,899	4,023	1,899	4,023	1,899	321	1,899	4,023	1,899	4,023	1,899	321	1,899	4,023	1,142	4,023
atpF	3,490	2,661	3,490	2,661	x	x	3,490	2,661	3,490	2,661	x	x	3,490	2,661	2,733	2,661
atpH	5,150	2,050	5,150	2,050	x	x	5,150	2,050	5,150	2,050	x	x	5,150	2,050	4,393	2,050
atpI	6,449	253	6,449	253	x	x	6,449	253	6,449	253	x	x	6,449	253	5,692	253
								rps2								
rpoC2	257	10,005	257	10,005	257	5,994	257	6,793	257	7,081	x	x	x	x	x	x
rpoC1	4,530	7,012	4,530	7,012	x	x	4,530	3,800	x	x	x	x	x	x	x	x
rpoB	7,392	3,773	7,392	3,773	4,557	2,597	x	x	4,557	3,684	x	x	x	x	x	x
trnC-GCA	11,779	2,520	11,779	2,520	8,944	1,344	8,567	2,520	8,944	2,431			1,459	2,520	1,432	2,431
petN	12,648	1,640	12,648	1,640	x	x	9,436	1,640		x	x	x	2,328	1,640	x	x
psbM	13,720	553	13,720	553	x	x	10,508	553	10,796	553	x	x	3,400	553	3,284	553
								trnD-GUC								
								trnY-GUA								
								trnE-UUC								
trnT-GGU	626	4,983	626	4,983	626	2,553	626	4,983	626	4,983	626	2,051	626	3,922	626	4,983
psbD	1,914	2,705	1,914	2,705	x	x	1,914	2,705	1,914	2,705	x	x	x	x	1,914	2,705
psbC	2,901	1,336	2,901	1,336	x	x	2,901	1,336	2,901	1,336	x	x	1,840	1,336	2,901	1,336
trnS-UGA	4,584	1,005	4,584	1,005	2,154	1,005	4,584	1,005	4,584	1,005	2,154	503	3,523	1,005	4,584	1,005
psbZ	5,014	478	5,014	478	2,584	478	5,014	478	5,014	478	x	x	3,953	478	5,014	478
trnG-GCC	5,446	164	5,446	164	3,016	164	5,446	164	5,446	164	x	x	4,385	164	5,446	164
								trnfM-CAU								
								rps14								
psaB	116	5,720	116	5,720	116	3,768	116	5,720	116	791	x	x	116	5,720	116	791
psaA	2,346	3,442	2,346	3,442	2,346	1,490	2,346	3,442	x	x	x	x	2,346	3,442	x	x
ycf3	5,323	765	5,323	765	x	x	5,323	765	x	x	x	x	5,323	765	x	x
								trnS-GGA								
								rps4								
trnT-UGU	407	5,935	407	5,935	407	3,610	407	4,257	407	5,935	x	x	407	4,257	x	x
trnL-UAA	1,190	4,658	1,190	4,658	1,190	2,333	1,190	2,980	1,190	4,658	x	x	1,190	2,980	1,118	3,535
trnF-GAA	1,584	4,279	2,063	4,279	2,063	1,954	2,063	2,601	2,063	4,279	714	1,954	2,063	2,601	1,991	3,156
ndhJ	2,797	3,141	2,797	3,141	x	x	x	x	2,797	3,141	x	x	x	x	x	x
ndhK	3,379	2,181	3,379	2,181	x	x	x	x	3,379	2,181	x	x	x	x	2,831	1,534
ndhC	4,113	1,939	4,113	1,939	x	x	x	x	4,113	1,939	x	x	x	x	3,565	1,292
trnV-UAC	5,602	165	5,602	165	x	x	3,924	165	5,602	165	x	x	3,924	165	x	x
								trnM-CAU								

	Distance to next conserved gene (in bp)															
	Lin		*Holo*		*EpCoCis*		*OroBou*		*MyzPhe*		*EpCo*		*Oro*		*Phe*[‡]	
Gene-ID	CCW	CW	CCW	CW	CCW	CW	CCW	CW	CCW	CW	CCW	CW	CCW	CW	CCW	CW
							trnM-CAU									
atpE	216	13,603	4,566	13,603	x	x	2,888	7,962	4,566	13,508	x	x	2,888	7,962	3,443	8,424
atpB	614	12,110	4,964	12,110	2,238	9,761	3,286	6,469	4,964	12,015	2,238	5,794	3,286	6,469	3,841	6,931
rbcL	2,900	9,887	7,250	9,887	4,524	7,538	5,572	4,246	7,250	9,792	4,524	3,571	5,572	4,246	6,127	4,708
accD	4,963	7,728	9,313	7,728	6,587	5,379	7,635	2,087	9,313	7,633	6,587	1,412	7,635	2,087	8,190	2,549
psaI	7,149	6,964	11,499	6,964	x	x	x	x	11,499	6,869	x	x	x	x	10,376	1,785
ycf4	7,706	5,960	12,056	5,960	9,223	3,718	x	x	12,056	5,865	x	x	x	x	x	x
cemA	9,093	4,438	13,443	4,438	x	x	x	x	13,443	4,343	x	x	x	x	x	x
petA	10,001	3,257	14,351	3,257	x	x	x	x	14,351	3,162	x	x	x	x	x	x
psbJ	12,009	2,089	16,359	2,089	11,656	1,717	x	x	16,359	1,994	x	x	x	x	x	x
psbL	12,265	1,839	16,615	1,839	x	x	x	x	16,615	1,744	x	x	x	x	x	x
psbF	12,405	1,696	16,755	1,696	x	x	x	x	16,755	1,601	x	x	x	x	x	x
psbE	12,539	1,430	16,889	1,430	11,927	1,317	x	x	16,889	1,335	x	x	x	x	x	x
petL	13,707	418	18,057	418	13,095	305	10,738	418			x	x	10,738	418	x	x
petG	13,984	123	18,334	123	x	x	11,015	123	18,239	123	x	x	11,015	123	12,032	123
							trnW-CCA									
							trnP-UGG									
psaJ	413	449	413	449	413	449	x		413	449	x	x	x	x	413	449
							rpl33									
							rps18									
							rpl20									
							rps12									
clpP	134	7,004	134	4,842	134	3,620	134	3,620	134	4,842	134	3,620	134	2,093	134	2,419
psbB	2,578	5,031	2,578	2,869	2,578	1,647	2,578	1,647	2,578	2,869	2,578	1,647			2,578	446
psbT	4,264	4,761	x	x	x	x	x	x	x	x	x	x	x	x	x	x
psbN	4,447	4,557	x	x	x	x	x	x	x	x	x	x	x	x	x	x
psbH	4,688	4,226	x	x	x	x	x	x	x	x	x	x	x	x	x	x
petB	5,039	2,709	x	x	x	x	x	x	x	x	x	x	x	x	x	x
petD	6,631	1,282	4,469	1,282	x	x	x	x	4,469	1,282	x	x	x	x	x	x
rpoA	8,044	81	5,882	81	4,660	81	4,660	81	5,882	81	4,660	81	3,133	81	x	x
							rps11									
							rpl36									
infA	68	125	68	125	68	125	68	125	68	125	68	125	68	125	68	125
rps8																
rpl14	177	126	177	126	177	126	177	126	177	126	177	126	177	126	177	126
							rpl16									
rps3	152	814	152	814	152	814	152	814	152	814	152	814	152	814	152	814
rpl22	830	409	830	409	830	409	830	409	830	409	830	409	830	409	830	409
rps19	1,286	64	1,286	64	1,286	64	1,286	64	1,286	64	1,286	64	1,286	64	1,286	64
							rpl2									
rpl23	19	166	19	166	x	x	x	x	19	166	x	x	x	x	19	166
							trnI-CAU									
							ycf2									
							trnL-CAA									
ndhB	549	275	549	275	549	275	549	275	549	275	549	275	549	275	549	275
							rps7									
							rps12-3end									
trnV-GAC	1,619	228	1,619	228	1,619	228	1,619	228	1,619	228	x	x	1,619	228	1,619	228
							rrn16									
trnI-GAU	299	1,129	299	1,129	299	1,129	299	1,129	299	1,129	299	1,129	299	1,129	x	x
trnA-UGC	1,384	179	1,384	179	1,384	179	1,384	179	1,384	179	1,384	179	1,384	179	x	x
							rrn23									
							rrn4.5									
							rrn5									
trnR-ACG	245	571	245	571	245	571	245	571	245	571	245	571	245	571	x	x
							trnN-GUU									
ndhF	1,485	1,444	1,485	1,444	x	x	x	x	1,485	1,444	x	x	x	x	x	x
rpl32	4,192	816	4,192	816	1,952	816	1,952	816	4,192	816	x	x	1,952	816	1,952	816
							trnL-UAG									

	Distance to next conserved gene (in bp)															
	Lin		*Holo*		*EpCoCis*		*OroBou*		*MyzPhe*		*EpCo*		*Oro*		*Phe*[#]	
Gene-ID	**CCW**	**CW**	**CCW**	**CW**	**CCW**	**CW**	**CCW**	**CW**	**CCW**	**CW**	**CCW**	**CW**	**CCW**	**CW**	**CCW**	**CW**
							trnL-UAG									
ccsA	106	8,520	106	7,390	x	x	106	6,579	106	6,209	x	x	106	5,399	x	x
ndhD	1,348	6,723	1,348	4,522	x	x	1,348	3,711	1,348	3,341	x	x	1,348	2,531	383	936
psaC	2,985	6,361	2,985	4,160	223	2,285	2,985	3,349	2,985	2,979	x	x	2,985	2,169	x	x
ndhE	3,515	5,771	3,515	3,570	x	x	x	x	3,515	2,389	x	x	x	x	x	x
ndhG	4,012	5,049	4,012	2,848	x	x	3,707	2,342	4,012	1,667	x	x	3,707	1,162	x	x
ndhI	4,884	4,201	4,884	2,000	x	x	x	x	4,884	819	x	x	x	x	x	x
ndhA	5,465	1,925	x	x	x	x	x	x	x		x	x	x	x	x	x
ndhH	7,668	742	5,467	742	830	742	4,656	742	x		x	x	x	x	x	x
rps15	8,946	373	6,745	373	2,108	373	5,934	373	5,564	373	x	x	4,754	373	2,194	373
							ycf1									

putative rearrangements in the common ancestor were not considered.

Table SIV-D **Plastid transcription units.** The table provides a summary of plastid gene arrangement in transcription units. For operon-like transcription units, the operon type is providided regarding the function of the respective genes. The studied organisms providing transcriptional evidence is given as well as the respective reference. The table continues the next page. Abbreviations: EQ – equal function, M – multi-functional, SIM – similar function.

Plastid gene or transcription unit	transcription	operon type	model organism	Remarks and reference
psbA	monocistronic		*Nicotiana*	Sugita and Sugiura 1996, - dicistronic transcribt observed in mustard by Nickelsen & Link 1991
trnK-matK	mono-/dicistronic		*Nicotiana*	Sugita et al. 1985
rps16	monocistronic		*Nicotiana*	Shinozaki et al. 1986
psbK-I	polycistronic	EQ	*Nicotiana*	Sugita and Sugiura 1996
trnSgcu	monocistronic		*Nicotiana*	Sugita and Sugiura 1996
trnGucc	monocistronic		*Nicotiana*	Sugita and Sugiura 1996,
rps2-atpA-F-H-I	polycistronic	M	*Nicotiana* *Spinacia*	Ohto et al. 1988, Miyagi et al. 1998 - *atpH/I* also transcribed independently after Sugita and Sugiura 1996,
trnCgca	monocistronic		*Nicotiana* *Oryza*	Sugita and Sugiura 1996 Kanno and Hirai 1993
rpoB-C1-C2	polycistronic		*Nicotiana*	Shinozaki et al. 1986
petN	monocistronic	EQ	*Nicotiana* *Oryza*	Sugita and Sugiura 1996, Legen 2002 Kanno and Hirai 1993
psbM	monocistronic		*Nicotiana*	Wakasugi et al. 1992
trnEuuc-Ygua-Dguc	polycistronic	EQ	*Nicotiana*	Ohme et al. 1985
trnTggu	monocistronic		*Nicotiana*	Sugita and Sugiura 1996
trnSuag	monocistronic		*Spinacia*	Gruissem et al. 1986
psbD-psbC-trnSuag-psbZ	polycistronic	EQ	*Nicotiana*	Sugita and Sugiura 1996 - *psbC* also transcribed independently - t*rnSuag* transcribed independently
trnGgcc	monocistronic		*Nicotiana*	Sugita and Sugiura 1996
trnfMcau	monocistronic		*Oryza*	Kanno and Hirai 1993
rps14-psaA-psaB	polycistronic	M	*Nicotiana*	Ohto et al. 1998
trnSgga	monocistronic		*Nicotiana* *Oryza*	Ohto et al. 1998 Kanno and Hirai 1993 - maybe co-transcribed with *ycf3*-operon
ycf3-trnSgaa-rps4-trnTugu	polycistronic	M	*Nicotiana* *Oryza*	Ohto et al. 1998 Kanno and Hirai 1993 - unclear how many genes transcribed may include *trnSgga*
trnLuaa	monocistronic		*Oryza*	Kanno and Hirai 1993 - maybe cotranscribed with *trnFgaa*
trnFgaa	monocistronic		*Oryza*	Kanno and Hirai 1993 - maybe cotranscribed with *trnLuua*
ndhC-K-J	polycistronic	EQ	*Nicotiana*	Matsubayashi et al. 1987
trnVuac	monocistronic		*Nicotiana* *Oryza*	Sugita and Sugiura 1996 Kanno and Hirai 1993
trnMcau	monocistronic		*Oryza*	Kanno and Hirai 1993
atpB-E	dicistronic	EQ	*Nicotiana*	Sugita and Sugiura 1996
rbcL	monocistronic		*Nicotiana*	Sugita and Sugiura 1996
accD	monocistronic		*Nicotiana*	Hajdukiewicz et al. 1997
psaI-ycf4-cemA-petA	polycistronic	SIM	*Nicotiana*	Shinozaki et al. 1986, Świątek 2002

Plastid gene or transcription unit	transcription	operon type	model organism	Remarks and reference
psbE-F-L-J	polycistronic		*Nicotiana*	Sugita and Sugiura 1996
trnWcca	monocistronic		*Oryza*	Kanno and Hirai 1993
trnPuug	monocistronic		*Oryza*	Kanno and Hirai 1993
petL-G-trnWcca-Pugg-psaJ-rpl33-rps18	polycistronic	M	*Nicotiana*	Sugita and Sugiura 1996, Ohto et al. 1998
clpP-rps12-rpl20	polycistronic	SIM	*Nicotiana*	Ohto et al. 1998
psbN	monocistronic		*Nicotiana* *Oryza*	Wakasugi et al. 1992 Kanno and Hirai 1993
psbB-T-N-H-petB-D	polycistronic	SIM	*Nicotiana* *Oryza*	Sugita and Sugiura 1996 Kanno and Hirai 1993 - *psbN* also transcribed from own promoter
rpoA-rps11-rpl36-infA-rps8-rpl14-16-rps3-rpl22-rps19-rpl2-rpl23	polycistronic	SIM	*Nicotiana*	Tanaka et al. 1986, Ohto et al. 1998
trnIcau	monocistronic		*Nicotiana*	Sugita and Sugiura 1996
ycf2	monocistronic		*Nicotiana*	Hajdukiewicz et al. 1997, Drescher et al. 2000
trnLcaa	monocistronic		*Oryza*	Kanno and Hirai 1993
ndhB	monocistronic		*Nicotiana*	Matsubayashi et al. 1987
rps7-12	dicistronic	EQ	*Nicotiana*	Ohto et al. 1998
trnVgac-rrn16-trnIgau-Augc-rrn23-rrn4.5-rrn5	polycistronic	SIM	*Nicotiana*	Shinozaki et al. 1986
trnRacg	monocistronic		*Oryza*	Kanno and Hirai 1993 - co-transcribed with rDNA operon in *Brassica napus* by Leal-Klevezas et al. 2000
trnNguu	monocistronic		*Oryza*	Kanno and Hirai 1993
ndhF	monocistronic		*Nicotiana*	Sugita and Sugiura 1996
trnLuag	monocistronic		*Nicotiana*	Sugita and Sugiura 1996
rpl32-trnLuag-cssA	polycistronic	M	*Nicotiana*	Sugita and Sugiura 1996
rps15-ndhH-A-I-G-E-psaC-ndhD	polycistronic	M	*Nicotiana*	Ohto et al. 1998
ycf1	monocistronic		*Nicotiana*	Drescher et al. 2000, Drescher 2003

9.3 References cited in the supplemental material

Bennett JR, and Mathews S. 2006. Phylogeny of the parasitic plant family Orobanchaceae inferred from phytochrome A. *Am. J. Bot.* **93**: 1039-1051.

Drescher A. 2003. ycf1, ycf14 und RNA-Edierung: Untersuchungen an im Lauf der Plastidenevolution neu hinzu gewonnenen Genen und Eigenschaften. Dissertation, Ludwig-Maximilian-Universität München, München. Available at http://edoc.ub.uni-muenchen.de/1369/.

Drescher A, Ruf S, Calsa T, Carrer H, and Bock R. 2000. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Plant J. 22: 97-104.

Hajdukiewicz PTJ, Allison LA, and Maliga P. 1997. The two RNA polymerases encoded by the nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids. EMBO J 16: 4041-4048.

Kanno A, and Hirai A. 1993. A transcription map of the chloroplast genome from rice (Oryza sativa). Curr. Genet. 23: 166 - 174.

Legen J. 2002. Gene expression in plastids of higher plants: evolutionary and functional aspects of different RNA polymerases – coordinated assembly of multiprotein complexes. Dissertation, Ludwig-Maximilian-Universität München, München. Available at http://edoc.ub.uni-muenchen.de/973/1/Legen_Juliana.pdf.

Leal-Klevezas DS, Martínez-Soriano JP, and Nazar RN. 2000. Cotranscription of 5S rRNA-tRNA-Arg$^{(ACG)}$ from *Brassica napus* chloroplasts and processing of their intergenic spacer. Gene 253: 303-311.

Matsubayashi T, Wakasugi T., Shinozaki Kazuo, Yamaguchi-Shinozaki K, Zaita N, Hidaka T, Meng B-Y, Ohto C, Tanaka M, Kato Akira, et al. 1987. Six chloroplast genes (*ndhA-F*) homologous to human mitochondrial genes encoding components of the respiratory chain NADH dehydrogenase are actively expressed: determination of the splice sites in *ndhA* and *ndhB* pre-mRNAs. Mol. Gen. Genet. 210: 385-393.

Miyagi T, Kapoor S, Sugita M, and Sugiura M. 1998. Transcript analysis of the tobacco plastid operon *rps2/atpI/H/F/A* reveals the existence of a non-consensus type II (NCII) promoterupstream of the atpI coding sequence. Mol. Gen. Genet. 257: 299-307.

Nickelsen J, and Link G. 1991. RNA-protein interactions at transcript 3′ ends and evidence for *trnK-psbA* cotranscription in mustard chloroplasts. Mol. Gen. Genet. 228: 89-96.

Ohme Masaru, Kamogashira Takashi, Shinozaki Kazuo, and Sugiura M. 1985. Structure and cotranscription of tobacco chloroplast genes for $tRNA^{Glu(UUC)}$, $^{tRNATyr(GUA)}$ and $tRNA^{Asp(GUC)}$. Nucl. Acids Res. 13: 1045 -1056.

Ohto C, Torazawa K, Tanaka M, Shinozaki Kazuo, and Sugiura M. 1988. Transcription of ten ribosomal protein genes from tobacco chloroplasts: A compilation of ribosomal protein genes found in the tobacco chloroplast genome. Plant Mol. Biol. 11: 589-600.

Shinozaki K., Ohme M., Tanaka M, Wakasugi Tatsuya, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, et al. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J. 5: 2043 - 2049.

Sugita M, and Sugiura M. 1996. Regulation of gene expression in chloroplasts of higher plants. Plant Mol. Biol. 32: 315.

Sugita M, Shinozaki Kazuo, and Sugiura M. 1985. Tobacco chloroplast $tRNA^{Lys(UUU)}$ gene contains a 2.5-kilobase-pair intron: An open reading frame and a conserved boundary sequence in the intron. Proc. Natl. Acad. Sci. USA 82: 3557 -3561.

Świątek M. 2002. Functional analysis of plastid-encoded genes. Application of reverse genetics on *Nicotiana tabacum*. Dissertation, Ludwig-Maximilian-Universität München, München. Available at http://edoc.ub.uni-muenchen.de/168/.

Tanaka M, Obokata J, Chunwongse J, Shinozaki K., and Sugiura M. 1987. Rapid splicing and stepwise processing of a transcript from the *psbB* operon in tobacco chloroplasts - Determination of the intron sites in *petB* and *petD*. Mol. Gen. Genet. 209: 427-431.

Wakasugi Tatsuya, Meng B-Y, Matsubayashi T, Vera A, Torazawa K, and Sugiura M. 1992. Transcription map of the tobacco chloroplast genome. In Research in Photosynthesis, proceedings of the IXth International Congress on Photosynthesis, Nagoya, Japan, August 30-September 4, 1992, pp. 263-266, Kluwer Academic Publishers, Dordrecht, The Netherlands.

– CHAPTER V –

PLASTID GENOME-WIDE ANALYSES OF SUBSTITUTION RATES UNCOVER RELAXATION OF SELECTIVE CONSTRAINTS IN HEMIPARASITES AND REVEAL PURIFYING SELECTION IN ATP-SYNTHASE SUBUNITS OF HOLOPARASITES

ABSTRACT. Plastid genes evolve under strong purifying selection as they encode essential subunits for photosynthesis, thus being essential for the autotrophic way of life. Nevertheless, elevation of nucleotide substitutions rates and relaxation of gene-specific selectional pressure occurs in several distinct land plant lineages, most prominently so in non-photosynthetic heterotrophic plants. Here, we examine plastid-genome-wide patterns of substitution rates and the evolution of selective pressures along the transition from an autotrophic stage via various forms of heterotrophy towards a complete parasitic way of life. Different broomrape lineages exhibit a great diversity of relative substitution rates in the set of retained plastid genes. Photosynthetic heterotrophs (hemiparasites) possess significantly higher rates of non-synonymous substitutions in both housekeeping and photosynthesis genes than autotrophs leading to a notable relaxation of selective constraints in photosynthesis related elements. Substitution rates are significantly increased in highly reconfigured plastomes of hemiparasites, whereas no such correlation exists in holoparasites. Genes for the plastid encoded thylakoid ATP-Synthase complex retained in a subset of non-photosynthetic broomrapes evolve at significantly lowered non-synonymous substitution rates and appear to underlie purifying selection.

KEYWORDS. Substitution rate evolution, selection pressure, plastid genome, hemiparasitic plants, non-photosynthetic plants, likelihood ratio tests

CONTENTS.

This chapter contains approx. 13,000 words, 6 figures, 2 tables, plus 5 pages of supplemental information.

1. INTRODUCTION

In general, the plastid chromosome evolves in a highly conservative mode concerning nucleotide substitutional rates, microstructural changes, and chromosomal architecture. In plastomes, nucleotide substitutional rates are about five times lower as those in the plant nuclear genome (1). Plastid genes typically evolve under strong purifying selection, which is reflected in mostly moderate mutational rates across a great diversity of land plants. Nevertheless, an elevation of relative plastid substitution rates occurs in several lineages independently and has been reported for lineages with aberrant lifestyles and short generation times implying a shaping role of macro-evolutionary features in rate evolution (2, 3). Furthermore, elevated plastid substitutional rates correlate significantly with severe departures from the normally highly conserved plastid chromosomal structure (4–9).

In particular, plants with certain degrees of a heterotrophic nutritional mode tend to possess higher nucleotide substitution rate elevation compared to autotrophs (10). For instance, significantly elevated rates have been observed in representatives of carnivorous Lentibulariaceae (5, 11) as well as in non-photosynthetic parasitic and myco-heterotrophic plants (e.g. 12–15). However, their cellular mechanisms and basis are barely understood. Shifts in nucleotide composition or mutations in transcription and repair systems are likely to contribute substantially to mutational rate variations (16). The latter may be considered as lineage specific effects (16). Moreover, an interrupted respiratory chain and high levels of reactive oxygen species (ROS) have been brought into discussion as another possible cause of rate elevation in the case of carnivorous Lentibulariaceae (17). Synonymous rates might be attributed to lineage effects and/or generation time, sharing a rather uniform pattern across different genes of the same genome (16). The rate of non-synonymous changes (dN) depends strongly upon selectional evolutionary constraints. Increased rates in overall nucleotide substitutions could thus indicate relaxation of selective constraints, for instance in parasitic plants, caused by increasing levels of heterotrophy (13, 14, 18). Selective evolutionary constraints can be evaluated by *omega* (ω), which describes the ratio of non-synonymous to synonymous changes (dN/dS). In order to account for gene-specific and lineage-specific effects, a direct comparison of changes in ω between closely related taxa provides a suitable approach for evaluating the evolution of selective pressures in genes and genomes along branches (19, 20).

In addition to many gene losses, the elevation of substitution rates in the remaining plastid coding regions is a well-known phenomenon in heterotrophic (in particular parasitic) plants (e. g. 13, 21). The loss of photosynthesis entails the relaxation of selective constraints acting on photosynthesis genes and genes of the genetic apparatus (12, 22–25,

21, 26). Relaxation of selection pressures allows for a significant increase in the amount of relative non-synonymous changes. Nevertheless, some retained plastid genes have been demonstrated to underlie purifying selection in non-photosynthetic plants (12, 15, 19, 20), implying that maintenance of transcriptional and translational activity seem to be of *some* importance in the life cycle of non-photosynthetic plants. Interestingly, significantly higher mutational rates have also been described for single plastid genes (e.g. *rps2*) in semi-heterotrophic (hemiparasitic) plants, i.e. plants that retain photosynthetic capability (13, 14). The complex pattern of plastid gene loss accompanying the transition from autotrophy to a heterotrophic way of life was illustrated in the preceding chapter (IV). Gene loss can be considered a consequence of relaxed selective constraints leading to the assumption that parasitism itself severely relaxes evolutionary constraints upon the photosynthesis apparatus. If this holds true we would expect significant mutational rate changes in the majority of plastid genes in hemiparasites compared to autotrophic relatives. If selective constraints were affected by parasitism, this should be measurable by ω.

We showed recently for a group of closely related holoparasites that reductive evolution of the plastid chromosome takes alternative evolutionary paths regarding patterns of gene loss and pseudogenization after the loss of selective pressures upon photosynthesis elements (Chapter IV). Broomrapes (Orobanchaceae) differ remarkably in plastid chromosome size reflecting both different gene contents as well as structural reconfigurations. Plastid chromosomal architecture may likely relate to substantial differences regarding the evolution of mutational rates among holoparasites. The strong correlation of non-canonical structural evolution and rate acceleration observed in autotrophic plants (4, 6, 8, 9, 29) lead us to test whether rate acceleration does also coincide with the occurrence of significant genomic rearrangements in parasites. If rate elevation correlates with a more rapid gene loss this would consequently imply that strongly reconfigured genomes with high relative substitution rates retain fewer genes than their closest relatives with structurally inconspicuous plastomes. Even more, the retention of several photosynthesis-related genes in holoparasites (30–32, 33; Chapter IV) implies that relaxation of evolutionary constraints upon plastid photosynthesis genes may not occur uniformly after the transition to holoparasitism. Retention of individual genes from distinct functional complexes (e.g. *atp* genes in *Phelipanche* and *Myzorrhiza,* Chapter IV) could indicate that they potentially carry out currently unknown functions beyond photosynthesis. In this case, we may find those genes still evolving under purifying selection, even though total substitution rates are elevated. In consequence, assessing evolutionary patterns of mutational rate changes will contribute substantially to our understanding of reductive evolution of the plastid chromosome under relaxed evolutionary constraints.

The broomrape family represents an excellent model system for investigating molecular evolution of plastid chromosomes for numerous reasons. As a major aspect distinguishing the family from other parasitic plants, Orobanchaceae span the complete set of evolutionary transition forms (34). Besides non-parasitic autotrophs, the family includes a great diversity of parasitic autotrophs (facultative and obligate hemiparasites with different degrees of photosynthetic capabilities) as well as non-photosynthetic heterotrophs (holoparasites) that evolved more than three times independently (35). Above that, the phylogenetic relationships within Orobanchaceae are comparatively well established (21, 35) and provide a solid basis for testing hypotheses on the evolution of rate patterns and selective constraints.

In the present study, we will examine patterns of substitution rates and selective pressures in hemi- and holoparasitic plants of the broomrape family using a maximum likelihood approach. Using the complete set of plastid genes from several representatives, we will thoroughly examine differences in substitution rates and selection among autotrophs, hemiparasites and holoparasites. Selective constraints and variation of selection pressures in plastid genes among parasitic species will be assessed via likelihood ratio tests (LRT). We will examine the correlation of substitution rate variation with the pattern of gene losses and pseudogenization. Above that, we will address the putative interrelation of rate acceleration and non-canonical structural plastid chromosome evolution.

2. RESULTS

2.1. Broomrape lineages exhibit a great diversity in relative nucleotide substitution rates.

We performed a plastid genome-wide analysis of 15 autotrophic and heterotrophic species of the broomrape family, four of which were newly sequenced here (see Chapter IV for the remainder). Out of the maximally 35 genes retained in the majority of Orobanchaceae plastomes, 25 genes are related to the function of the plastid genetic apparatus (*rpl*/*rps* genes, *clpP*, *infA*, *matK*; for details see Chapter IV and supplemental material: Table SV-A). The remaining genes carry out functions in lipid synthesis (*accD*), and photosynthesis-coupled ATP synthesis (*atp*-genes). The latter unexpectedly survived completely as potentially functional reading frames in the plastomes of 5 out of 11 investigated holoparasitic species; a reduced subset of *atp*-genes exists in two more holoparasitic species (Chapter IV for details). The hemiparasites *Striga* and *Schwalbea* have independently lost subunits of the plastid *Ndh* dehydrogenase complex. In *Striga*, extreme

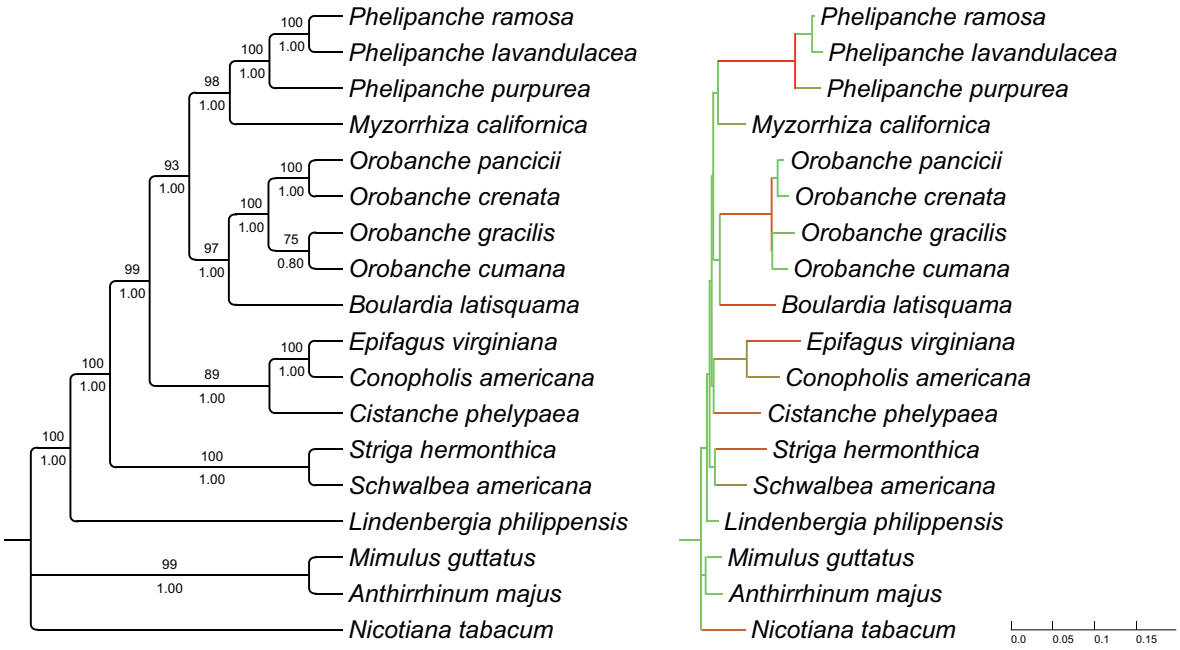


Fig. V-1 **Phylogenetic relationships among Orobanchaceae based upon plastid coding regions.** Left: Topology inferred with maximum likelihood and Bayesian inference, ML bootstrap proportions are depicted above, and posterior probabilities below branches. Right: phylogram from ML. Colors indicate the relative amounts of substitutional changes. Rates increase from green to red.

sequence divergence of plastid genes encoding the plastid polymerase suggests pseudogenization of at least some subunits. Moreover, there is evidence for non-functionalization of the *clpP* and *accD* genes in both *Striga* and *Schwalbea*. Both genes possess very large indels (relative to *Lindenbergia* and close lamiid taxa) and diverge extremely in terms of nucleotide substitutions. The two hemiparasites also differ remarkably with respect to plastid chromosome structure. While only two smaller inversions exist in *Schwalbea* (Chapter IV), *Striga* shows an extreme reconfiguration around the large inverted repeat region (IR), and exhibits several large-scale inversions (36; S. Wicke et al., unpubl. data). The gene contents and chromosomal structure of herein newly sequenced holoparasitic species (*O. cumana, O. pancicii, P. lavandulacea*) are mostly consistent with those analyzed in an earlier study (Chapter IV; Supplemental Material: Table SV-A). *Phelipanche lavandulacea* shares the plastid genome reconfiguration with *P. ramosa* distinguishing both taxa from *P. purpurea*. Unlike in *O. gracilis*, the plastomes of *O. cumana* and *O. pancicii* possess a very conservative structure, similar to *O. crenata* (Chapter IV).

We inferred the relationships among Orobanchaceae based upon the coding regions from 15 fully sequenced plastid genomes and three outgroup taxa. The aligned dataset with 77 plastid protein-coding genes comprises 57,930 characters. Using maximum likelihood (ML) and Bayesian inference under the GTR+Γ+I model, we inferred a well-resolved and statistically well supported tree (Fig. V-1). *Lindenbergia* is sister to the

remainder Orobanchaceae. The two hemiparasites *Schwalbea* and *Striga* form one clade that is sister to the holoparasitic broomrape clade. The relationships between the different non-photosynthetic lineages match those inferred in previous works with a much denser taxon sampling (e.g. 37, 38).

The phylogenetic reconstruction suggests a great diversity of nucleotide substitution rates in both the hemiparasites and among holoparasites (Fig. V-1). Compared to *Lindenbergia* and closely related lamiid taxa, both hemiparasitic species show elevated substitution rates. The magnitude of mutational changes is slightly higher in *Striga*. In holoparasites, nucleotide substitution rates vary greatly. *Myzorrhiza* appears to exhibit the smallest degree of substitution rate elevation compared to photosynthetic plants. Its sister group *Phelipanche*, however, seems to be most divergent among broomrapes. The plastid chromosomes in this clade display the most dramatic structural reconfigurations among holoparasites described so far (see Chapter IV for details).

2.2. Hemiparasites possess significantly higher relative nucleotide substitution rates in housekeeping genes and elements of the photosynthesis apparatus than *Lindenbergia*

In order to analyses the effect of parasitism on mutational rates, we at first analyzed the relative nucleotide substitution rates for all plastid genes between different autotrophs using the GTR model. We find that the majority of plastid genes evolve widely at similar rates in the autotrophic *Lindenbergia* and outgroup species. Among 77 investigated plastid protein coding genes, we do find 22 genes that exhibit significantly different mutational rates between autotrophs (Supplemental material: Table SV-B). Only two genes (*accD, matK*) evolve at a higher rate in *Lindenbergia*. The remaining genes display significantly lower relative substitutional changes in *Lindenbergia*. Among those, we find many ribosomal genes as well as subunits for the ATP-synthase complex (*atpA, atpB*), photosystem genes (*rbcL, psbI,K,M*), and elements of the plastid *Ndh*-complex (*ndhB, D, H, K*). These results suggest that an elevation of substitution rates has not been present in the common ancestor of *Orobanche*, but appears to have evolved later on.

We evaluated whether parasitism correlates with rate acceleration via relative rate tests between *Lindenbergia* and parasitic Orobanchaceae. In both the hemiparasites *Schwalbea* and *Striga*, we observe an elevated nucleotide substitution rate in a number of plastid genes (Fig. V-2). Compared to *Lindenbergia*, rate acceleration accounts for genes of photosynthesis related pathways as well as for genes of the genetic apparatus. In hemiparasites, genes for the ATP-synthase (*atpA, B, E, F, H*), retained *ndh*-genes as well as

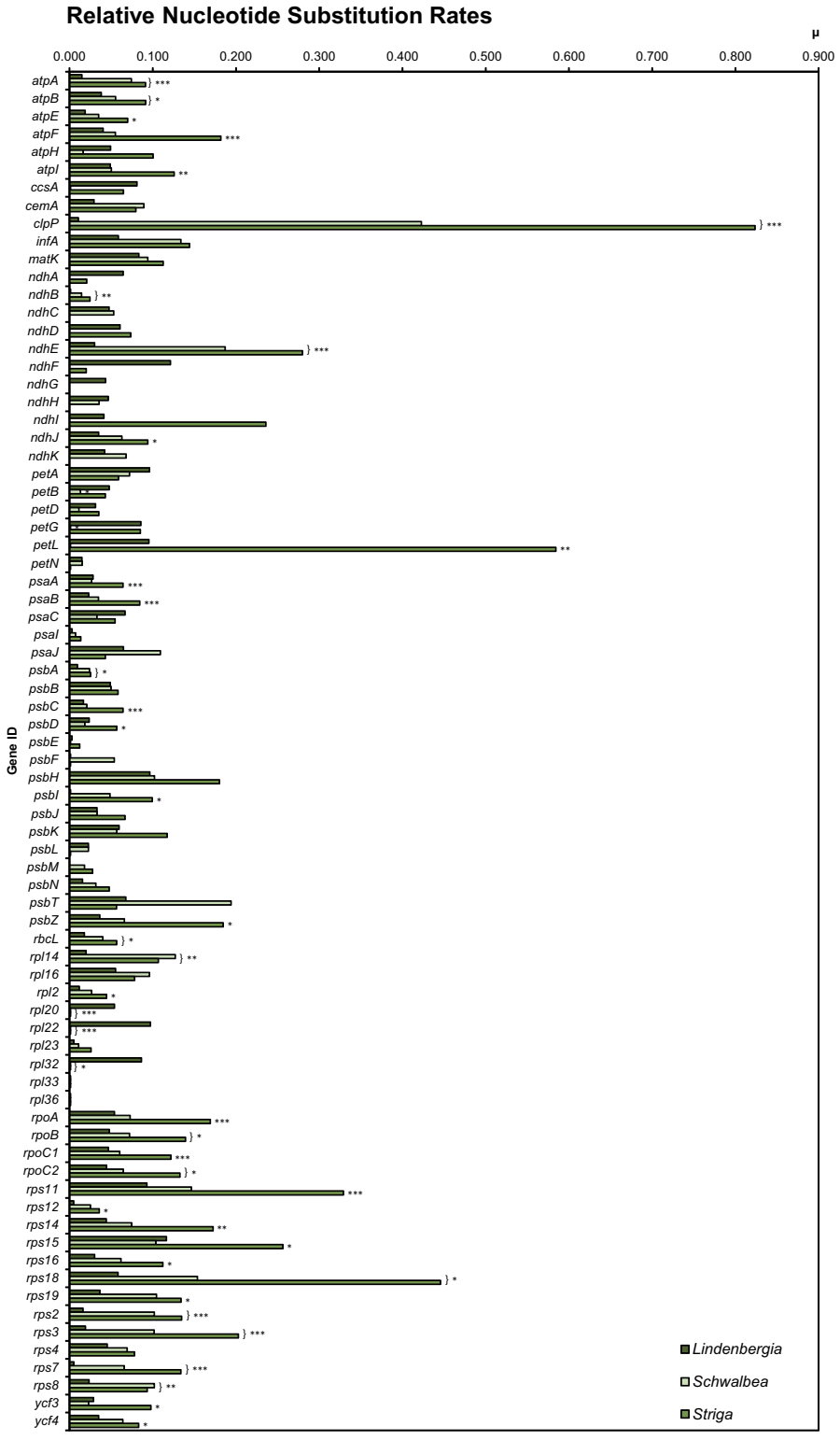


Fig. V-2 **Relative nucleotide substitution rates between autotrophic and hemiparasitic Orobanchaceae.** Relative rate tests between autotrophic *Lindenbergia* and two hemiparasites are illustrated for all plastid genes as vertical bars. Asterisks mark the significance level of rate changes in hemiparasites compared to *Lindenbergia* (<0.05*, <0.01**, <0.001***). Parentheses mark significant deviation of substitution rates of a particular gene in both hemiparasites relative to *Lindenbergia.* In this case, asterisks indicate the minimum common significance level. Complete lack of a bar indicates genes loss

rbcL show significantly increased nucleotide mutation rates. Also, substitutions increase in all genes of the plastid-encoded polymerase as well as the majority of ribosomal protein genes in both hemiparasites relative to *Lindenbergia*. Across all plastid genes, c*lpP* exhibits the highest rate of change. The number of genes sped up in *Striga* is much higher than in *Schwalbea.* This is particularly evident for the number of photosystem genes. While only *psbA* evolves faster in *Schwalbea* relative to *Lindenbergia*, we observe significant rate increases in the two largest of the five plastid encoded subunits of photosystem I (*psaA, B*); five subunits of photosystem II (*psbC, D, I, Z*) are also affected along with the two photosystem assembly factors *ycf3* and *ycf4*. Several more photosystem genes are sped up with marginal significance (Fig. V-2; Supplemental Material: Table SV-C). In *Schwalbea*, most of those genes found to evolve significantly different in *Striga* are also accelerated with marginal significance relative to *Lindenbergia* (Supplemental Material: Table SV-B). With the exception of *petL*, the lowest rate variation is found in subunits of the cytochrome complex. Even more, *petD* and *pet*G show significantly fewer substitutions in *Schwalbea* than in *Lindenbergia*.

Compared to hemiparasites with only few *ndh*-gene losses, holoparasites underwent extensive functional reduction of the plastid chromosome retaining on average only between 20 and 30 functional protein coding genes (Chapter IV). Across the retained genes, we observe extreme variation in relative nucleotide substitution rates between the different lineages (Fig. V-3). Even more, the pattern of rate variation is extremely complex showing not only significant rate elevations, but also significantly lower relative rates in holoparasites versus *Lindenbergia*. Highest rates occur in the *Epifagus*/*Conopholis*/*Cistanche* clade with *Cistanche* exhibiting on average higher mutational rates than *Epifagus* and *Conopholis*. All genes retained in the plastomes of this clade as well as those from *Boulardia* are sped up remarkably. The majority of those changes are significant accelerations of mutational rates.

The pattern of rate variation is most interesting in that we do not encounter a general elevation of relative nucleotide substitution in holoparasites. Relative rates are only partially accelerated in *Orobanche* species and the *Phelipanche*/*Myzorrhiza* clade. In the latter, molecular rates change substantially in the retained set of *atp* genes, although rate variation is only significant in a subset of those genes for the different taxa. Moreover, while we observe an increasing substitutional rate in *Myzorrhiza* and *Phelipanche purpurea*, *P. ramosa* and *P. lavandulacea* evolve at a notably reduced rate compared to *Lindenbergia*. A similar pattern applies to the majority of the ribosomal protein genes in the *Phelipanche* clade. Compared to the remainder holoparasites, plastid genes of *Orobanche* species evolve at rather moderate rates comparable to that of *Lindenbergia*. Significant changes are confined to few genes, including *accD, rpl22* and *33*, *rps2*, and *matK*, although some of the species exhibit only marginally significant differences.

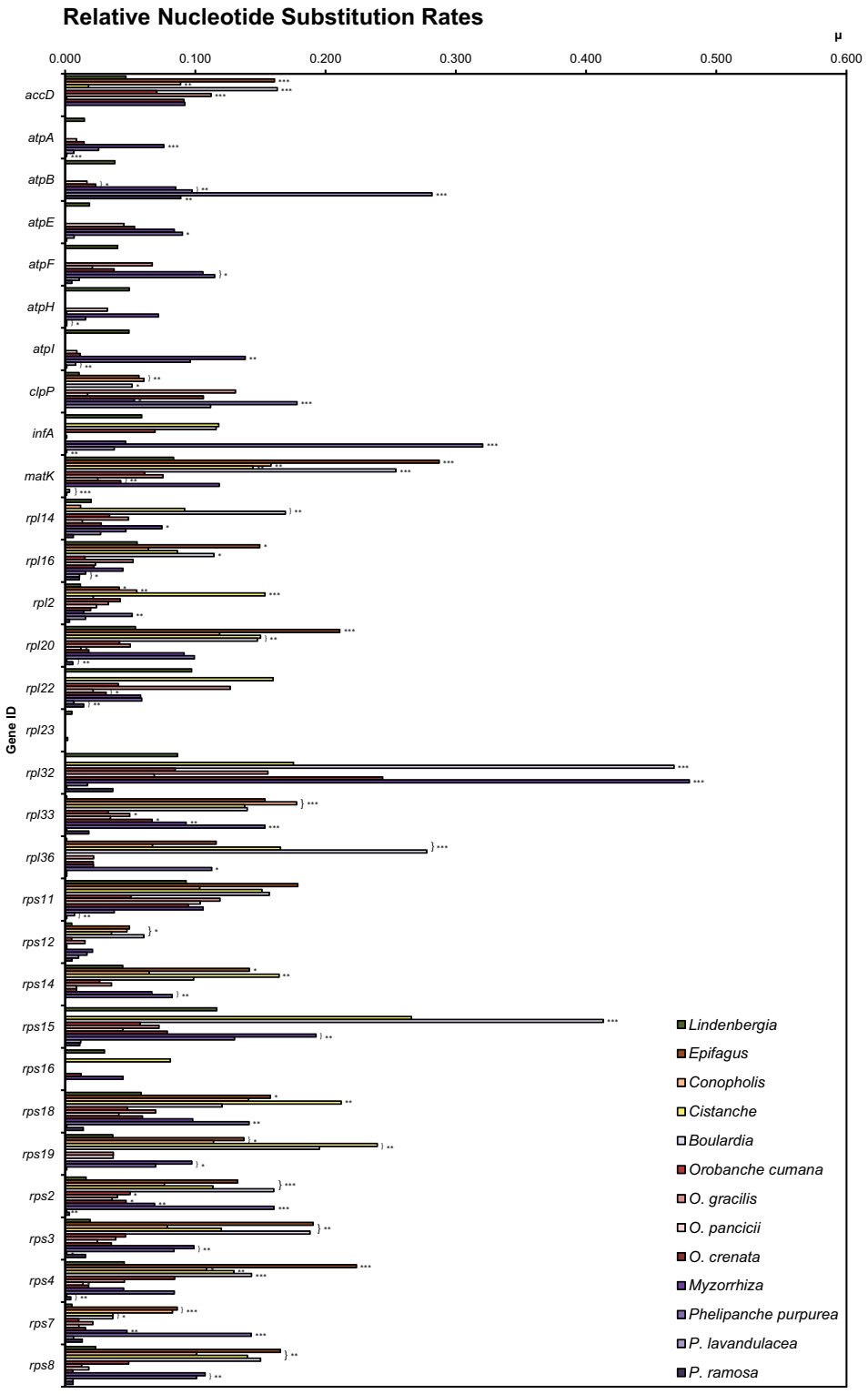


Fig. V-3 **Relative nucleotide substitution rates between autotrophic and holoparasitic Orobanchaceae.** Relative rate tests between autotrophic *Lindenbergia* and 12 holoparasites are illustrated for all retained plastid genes as vertical bars. Asterisks mark the significance level of rate changes in parasites compared to non-parasitic *Lindenbergia* (<0.05*, <0.01**, <0.001***). Parentheses mark a significant deviation of substitution rates of a particular gene in two or more holoparasites relative to *Lindenbergia* In this case, asterisks indicate the minimum common significance level. Species-specific lack of a bar indicates genes loss.

2.3. Substitution rates are significantly higher in reconfigured plastomes of hemiparasites, but decrease in plastid genes of holoparasites with structurally aberrant plastomes.

Severely reconfigured plastome are found in the hemiparasites *Striga,* and the holoparasites *Conopholis, Orobanche gracilis* and *Phelipanche*. Using LRT for all plastid genes, we tested whether these parasites differ remarkably in nucleotide substitution rate relative to their closest related sister species. The results of our tests regarding a correlation between elevated nucleotide substitution and an aberrant chromosome structure applies also for semi-heterotrophs, but pertains only to a minimal degree in holoparasites. In *Striga hermonthica*, evolutionary rates are extremely elevated compared to *Schwalbea americana* (Fig. V-4). Most notably, highly significant rate increases occur in all genes for the plastid-encoded polymerase (PEP, encoded by *rpoA-C2*; $p \ll 0.001$) as well as several ribosomal protein genes for the small ribosomal subunits and *clpP*. Besides housekeeping genes, several elements of photosynthesis-related complexes are significantly accelerated, including ATP-Synthase genes and subunits for the photosystems I and II. In contrast, *rpl* genes appear to evolve at a slower rate in *Striga*, although statistical support is weak (Supplemental material: Table SV-C). *Orobanche gracilis* exhibits slightly higher rates in its retained genes than its closest relative *Orobanche cumana*. However, the majority of gene-specific rate variations are insignificant in the *O.gracilis-O. cumana* comparison, except for the genes *rps16* ($p < 0.05$) and *rps18* ($p < 0.01$). Moreover, we observe slightly lower rates in some *rps* genes (*rps2, 3, 4, 8*) as well as in *rpl2*. In *Conopholis,* which has lost a complete segment of the IR, most plastid genes evolve on average at lower rates than in *Epifagus.* The difference is highly significant in *accD* and *matK* as well as *rps2* and *rps3*. Extreme reconfigurations are found in the plastid chromosome of *Phelipanche* species with large scale inversions and drastic IR diminution and eventual loss, as well as several large-scale inversions (see Chapter IV for details.). We encounter extreme deviation in relative nucleotide substitution rates in the *Myzorrhiza*/*Phelipanche* lineage. While *Phelipanche purpurea* shows hardly any significant difference to structurally highly conserved *Myzorrhiza*, extremely reduced mutation rates occur in *P. lavandulacea* and *P. ramosa.* We observe the greatest difference in substitution rates in *rpl32* and *infA*. Those rate elevations in *Myzorrhiza* may be indicative of a putative non-functionalization of both genes; the same accounts for *atpB* in *P. lavandulacea,* which evolves significantly faster than in both *Myzorrhiza* and other *Phelipanche*-species.

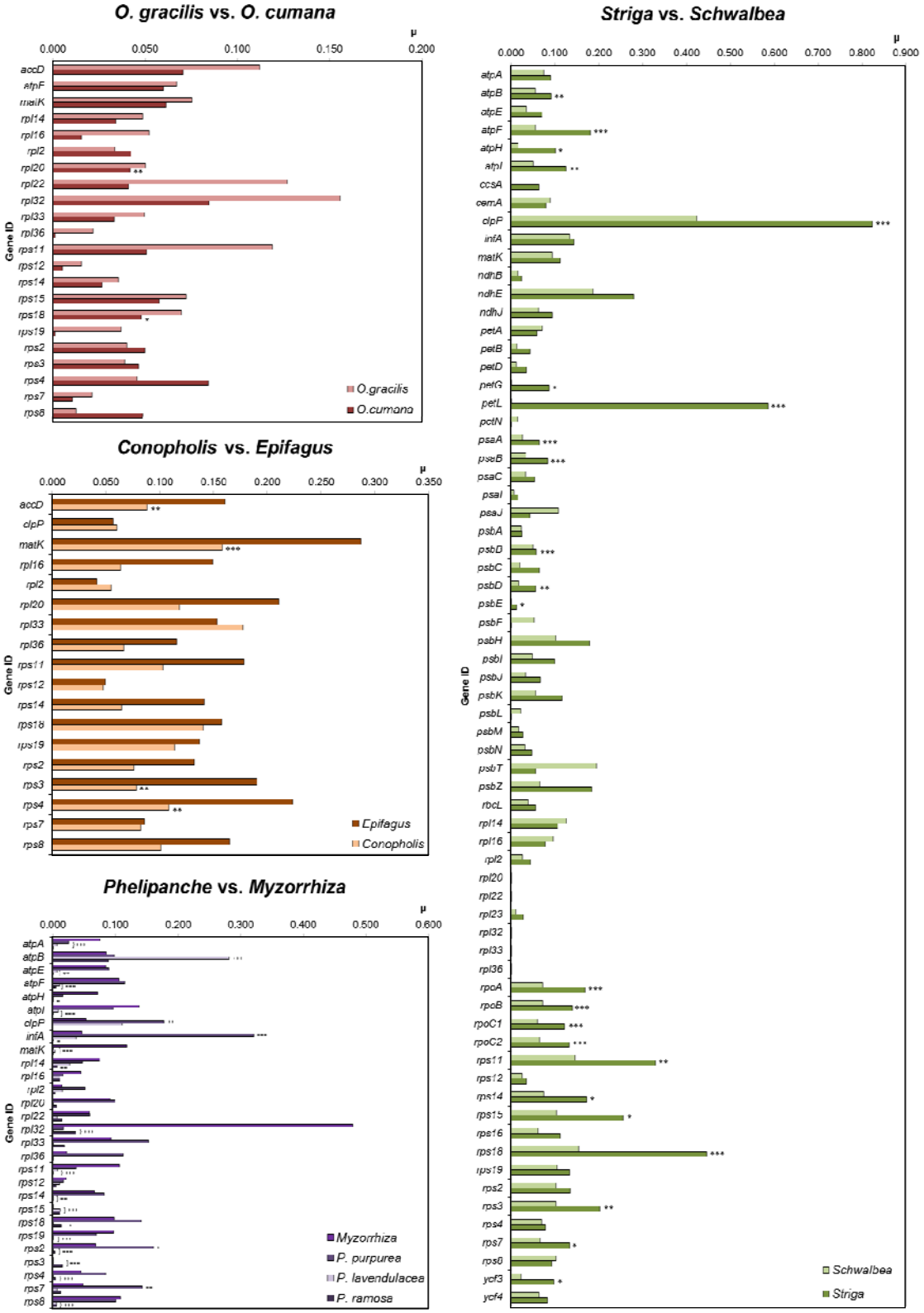


Fig. V-4 **Relative nucleotide substitution rates of parasites with anomalous plastome structure.** Results of relative rate tests comparing parasites with a non-canonical plastid chromosome structure to its closest relative with without genomic rearrangements are illustrated for all plastid genes as vertical bars. Asterisks mark the significance level of rate changes in parasites compared to *Lindenbergia* (<0.05*, <0.01**, <0.001***). Parentheses mark a significant deviation of substitution rates of a particular gene in two or more holoparasites relative to *Lindenbergia* In this case, asterisks indicate the minimum common significance level. Species-specific lack of a bar indicates genes loss.

2.4. Hemiparasites show an extreme increase of non-synonymous substitution rates in genes coding for subunits of both the plastid genetic apparatus and photosynthesis complexes.

We assessed changes in selective pressures upon plastid genes via LRTs on synonymous rats (dS), non-synonymous (dN) rates, and omega (ω), respectively. Compared to *Lindenbergia*, hemiparasites exhibit changes of synonymous rates in several elements of the genetic apparatus as well as in genes for photosynthetic complexes (Fig. V-5). In *Striga*, dS is elevated in genes for the ATP-synthase complex (*atpA, B, F*), some genes for the photosystem I (*psaA, B*), and in most ribosomal protein genes and subunits of the plastid polymerase (*rpoA, C1, C2*). In many cases, those genes are also accelerated in *Schwalbea* relative to *Lindenbergia*, but to a much lesser extent. Rate increases extremely in the *clpP* genes of hemiparasites. Significantly fewer synonymous substitutions occur in some of the *rpl* genes (*rpl20, 22*) of *Schwalbea*. Pronounced differences in non-synonymous substitutions mostly co-occur in genes with high synonymous rates. Furthermore, several housekeeping and photosynthesis-related genes exhibit dN-elevation independently of increases in dS (Fig. V-5). Except for the subunits of the cytochrome b_6f complex, we observe significant changes in *atp-, ndh-, psa-*, and *psb*-genes. Similar to previous results, increases relative to *Lindenbergia* are much stronger in *Striga*. In particular genes for the photosystems are seen to evolve primarily with higher non-synonymous substitution rates. Many of the housekeeping genes also exhibit significant changes in hemiparasites compared to *Lindenbergia* with differences being again much more pronounced in *Striga* than in *Schwalbea*. The genes that are found to evolve with significantly different synonymous and non-synonymous rate compared to *Lindenbergia* are widely congruent with those identified in overall rate tests. Prominently, in hemiparasites the majority of genes show significantly more changes in the non-synonymous rate changes. This suggests that the transition to hemiparasitism allows significant relaxation of purifying selection in plastid genes giving rise to pseudogenization and eventual gene loss. Protein-coding genes retained in plastid genomes of holoparasitic plants mainly encode subunits for the translation machinery. Those genes evolve with elevated rates of both synonymous and non-synonymous changes compared to autotrophs.

The analyses of differences in synonymous and non-synonymous rates between *Lindenbergia* on the one hand and all non-photosynthetic species on the other revealed great differences between the three holoparasitic clades (Fig. V-6). Both synonymous and non-synonymous rates of plastid genes are remarkably increased in the *Epifagus*/*Conopholis*/*Cistanche* lineages as well as in *Boulardia*, whereas rate deviation is rather moderate in *Phelipanche* and *Orobanche.*

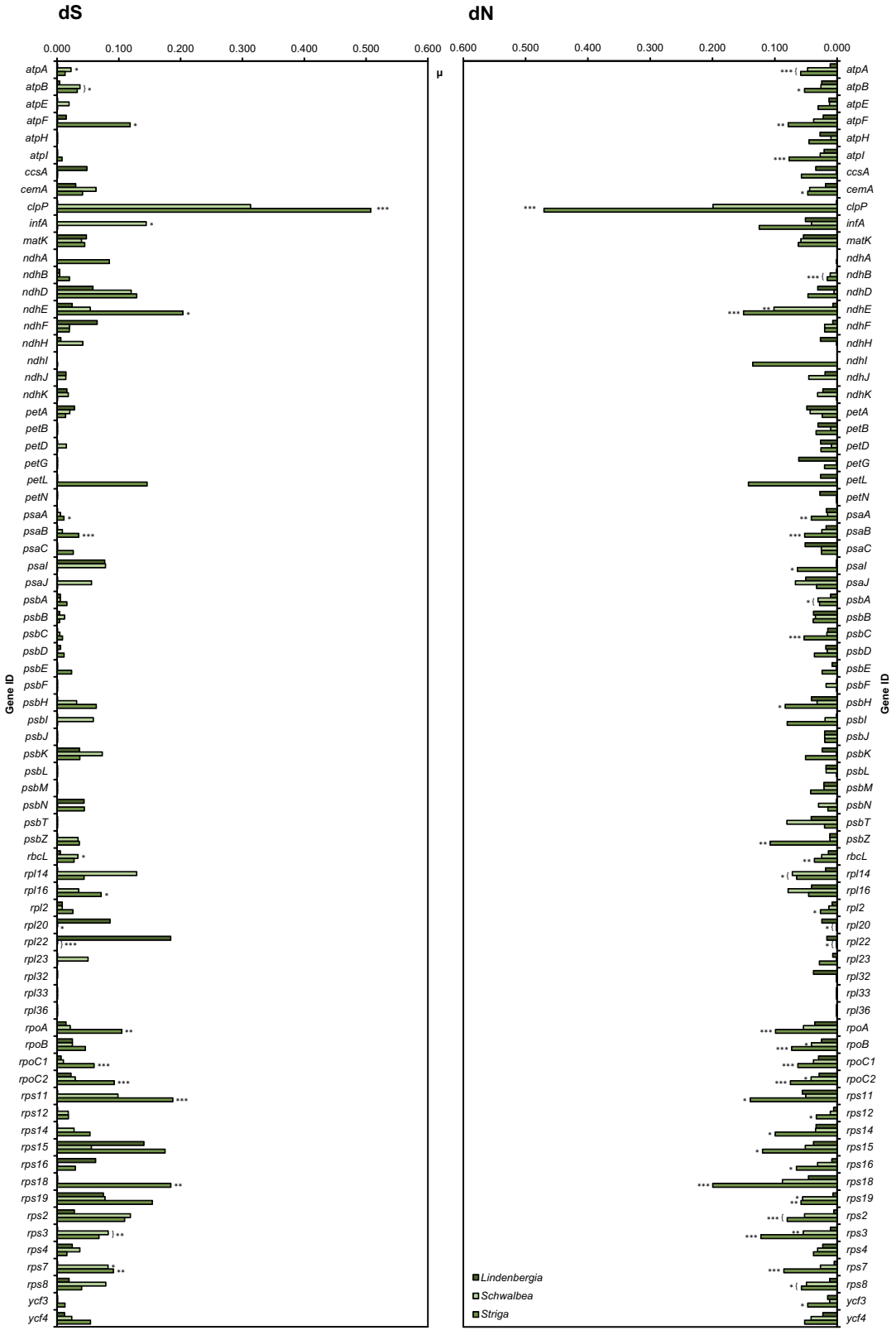


Fig. V-5 **Relative synonymous and non-synonymous substitution rates in hemiparasites.** Results of relative rate tests comparing synonymous and non-synonymous substitutions of hemiparasites to autotrophic *Lindenbergia* illustrated for all plastid genes as vertical bars. Asterisks mark the significance level of rate changes in hemiparasites (<0.05*, <0.01**, <0.001***). Parentheses mark a significant deviation of substitution rates of a particular gene in two or more holoparasites relative to *Lindenbergia*. In this case, asterisks indicate the minimum common significance level. Species-specific lack of a bar indicates genes loss.

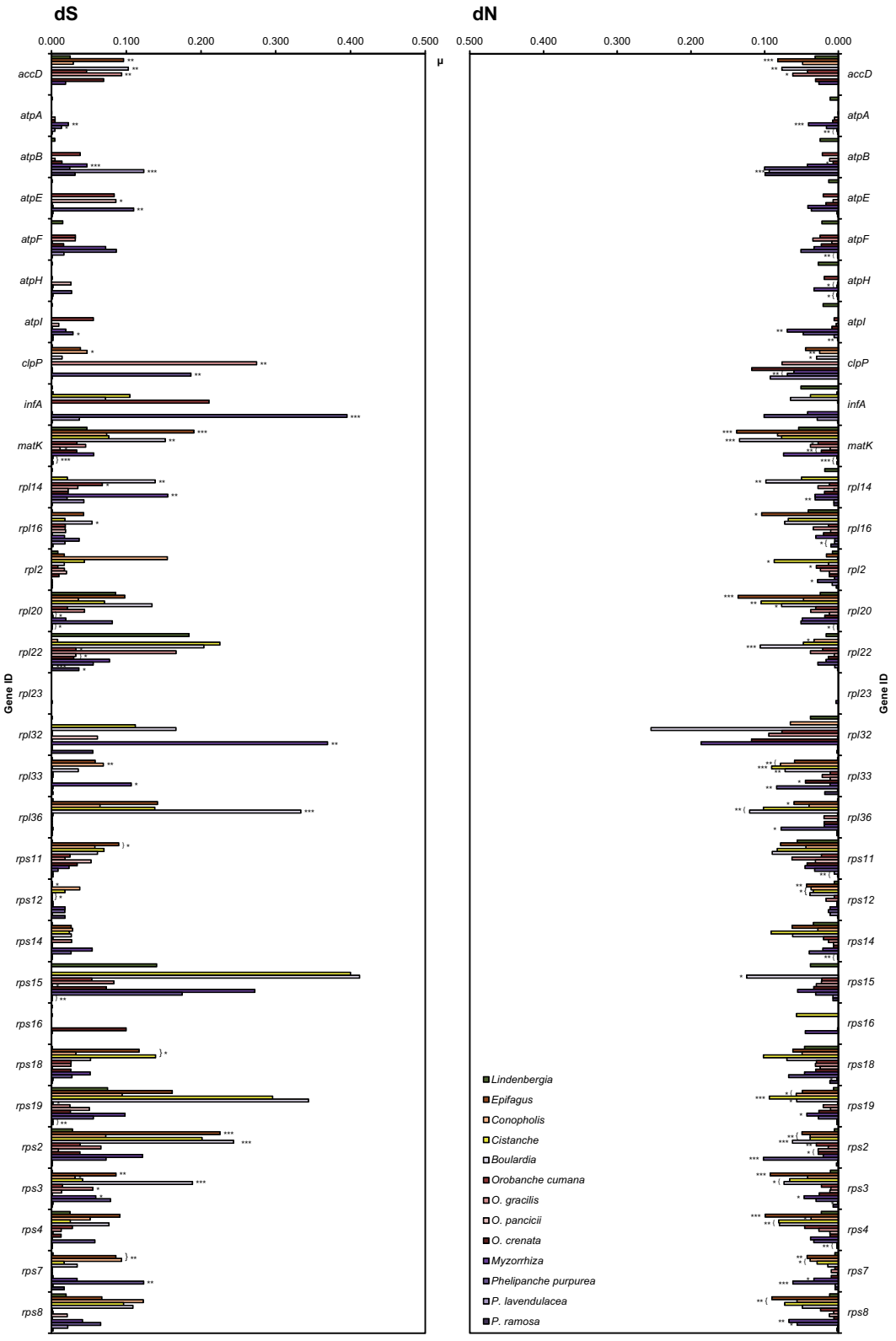


Fig. V-6 **Relative synonymous and non-synonymous substitution rates in holoparasites.** Results of relative rate tests comparing synonymous and non-synonymous substitutions of holoparasites to autotrophic *Lindenbergia* are illustrated for all plastid genes. Asterisks mark the significance level of rate changes in holoparasitic lineages (<0.05*, <0.01**, <0.001***). Parentheses mark a significant deviation of substitution rates of a particular gene in two or more holoparasites relative to *Lindenbergia*. In this case, asterisks indicate the minimum common significance level. Species-specific lack of a bar indicates genes loss.

Myzorrhiza also show notable elevation of synonymous rates, but in general resembles *Phelipanche* and *Orobanche* concerning non-synonymous substitutions. *Orobanche* species show the least rate differences compared to *Lindenbergia* across retained plastid genes, reflecting the moderate elevation of overall nucleotide substitutions in plastid genes of holo-heterotrophs. In contrast to *Epifagus*/*Conopholis*/*Cistanche*, *Boulardia* and *Myzorrhiza* where we observe a general rate elevation, both synonymous and non-synonymous rates are significantly lower in *Phelipanche*-species relative to *Lindenbergia,* which is most unexpected and surprising. Significant non-synonymous rate reduction occurs in many housekeeping genes of *Phelipanche* and *Orobanche*, whereas *Epifagus*/*Conopholis*/*Cistanche* and *Boulardia* display on average moderately to significantly elevated dN-values compared to *Lindenbergia*.

2.5. Plastid-encoded polymerase for the transcription of photosynthesis elements evolve under highly significantly relaxed purifying selection in hemiparasites.

We tested whether evolutionary constraints are significantly altered in hemiparasites by performing genewise LRTs comparing selective pressures by changes of ω between *Lindenbergia* and semi-heterotrophs. We analyzed 19 essential photosynthesis genes, all of which exhibited significant changes in non-synonymous substitution rates (see section 2.4.) that function either directly in any of the photosynthesis complexes or are responsible for the efficient transcription of photosynthesis-genes (*rpo*-genes). Genewise and single-species tests revealed that sequence variation is insufficient to allow proper estimates of the dN/dS-ratios, and in most cases local ω-estimates were either infinite or remained undefined (see 39, 40 for details). For that reason, we analyzed changes of selection pressure clade-wise by comparing the hemiparasite clade (*Striga* + *Schwalbea*) versus autotrophic species (*Lindenbergia* + *Antirrhinum* + *Mimulus* + *Nicotiana*) and also included the branch leading to the hemiparasite subtree. LRTs of models allowing for one, two and three possible selection paths revealed that *atpI, rbcL* as well as six out of seven photosystem genes (*psaA/I, psbA/C/H/Z*) did not show any alteration in selection pressures between hemiparasites and autotrophs (Table V-A, Supplemental material: Table SV-C). In case of *atpI, rbcL* and *psaA/I, psbA/C,* AIC selected a single-rate model assuming a global ω for all clades and branches of the tree. Relaxed purifying selection was detected in the chloroplast envelope protein involved in CO_2 uptake (*cemA*), subunits of the ATP-Synthase complex (*atpA/B/F*), remnants of the plastid *Ndh*-complex (*ndhB/E*), and all genes encoding the plastid polymerase (*rpoA/B/C1/C2*). For *atpA, atpB* and *ndhB,* the assumption of independent ω for autotrophs, hemiparasites and the hemiparasite root are most suitable to explain variation of selection pressures. Hemiparasites exhibit a significantly different dN/dS ratio compared to autotrophs indicating relaxed purifying selection in *atpA/B*

Table V-A **Changes of selection in selected plastid genes of hemiparasites.** Photosynthesis-related genes with elevated rate of non-synonymous substitution rates were analyzed regarding changes of selection between autotrophs and hemiparasites. Clade-specific ω-values from 5 different models (mid-lines without "<") are provided with their respective lower ("*nn* <") and upper ("< *nn*") confidence levels including their respective p-value from LRT against the null model assuming a global ω. The best fit-model according to AIC-ranks is highlighted in blue. Clade A corresponds to the hemiparasite subtree, clade B corresponds to autotrophic rest of the tree. Branch (T) denotes the root of the subtree. Asterisks indicate the significance level (<0.05*, <0.01**, <0.001***).

	Single rate model	2 selective paths									3 selective paths			
		Branch (T) vs. clade A/B			Clade A/branch vs. clade B			Clade A vs. clade B /branch			Clade A, Clade B, branch			
Gene	ω_{global}	ω_{A+B}	ω_T	p-value	ω_{A+T}	ω_B	p-value	ω_A	ω_{B+T}	p-value	ω_A	ω_B	ω_T	p-value
atpA	0.0683< 0.0884 < 0.1121	0.0613< 0.0805 < 0.1033	0.3042< 0.8236 < 1.7249	0.010*	0.1115< 0.1558 < 0.2105	0.0339< 0.0520 < 0.0757	<<0.001***	0.0993< 0.1432 < 0.1987	0.0418< 0.0617 < 0.0870	0.006**	0.0954< 0.1373 < 0.1901	0.0336< 0.0517 < 0.0752	0.2364< 0.6459 < 1.3530	<<0.001***
atpB	0.0589< 0.0782 < 0.1012	0.0505< 0.0684< 0.0902	0.5487< 1.3511 < 2.6995	<<0.001***	0.1106< 0.1612 < 0.2252	0.0288< 0.0452 < 0.0669	<<0.001***	0.0899< 0.1381 < 0.2013	0.0377< 0.0562< 0.0800	0.007**	0.0846< 0.1301 < 0.1895	0.0288< 0.0452 < 0.0668	0.4015< 1.0073 < 2.0214	<<0.001***
atpF	0.1908< 0.2571 < 0.3372	0.1876< 0.2548< 0.3366	0.0388< 0.3118< 0.9992	0.856	0.3108< 0.4620 < 0.6573	0.0975< 0.1573 < 0.2377	0.007**	0.3284< 0.4979 < 0.7186	0.1034< 0.1632 < 0.2421	0.007**	0.3251< 0.4926 < 0.7108	0.0978< 0.1577 < 0.2381	0.0254< 0.2651 < 0.8650	0.024*
atpI	0.0387< 0.0622 < 0.0936	0.0388< 0.0622 < 0.0936	NA	1	0.0418< 0.0836 < 0.1469	0.0249< 0.0497 < 0.0872	0.304	0.0418< 0.0835 < 0.1469	0.0249< 0.0497 < 0.0873	0.304	0.0418< 0.0835 < 0.1469	0.0249< 0.0497 < 0.0872	NA	0.589
cemA	0.3029< 0.3749 < 0.4575	0.2956< 0.3669 < 0.4490	0.3793< 5.9533 <19.9627	0.201	0.4769< 0.6621 < 0.8927	0.2093< 0.2788 < 0.3626	0.014*	0.4581< 0.6442 < 0.8754	0.2161< 0.2870 < 0.3720	0.024*	0.4945< 0.6855 < 0.9212	0.2106< 0.2807 < 0.3654	NA	0.217
ndhB	0.2298< 0.3363 < 0.4719	0.2032< 0.3040 < 0.4342	1.4923< 6.1204 <15.9617	0.047*	0.4045< 0.6414 < 0.9577	0.0757< 0.1578 < 0.2843	0.008**	0.3379< 0.5594 < 0.8617	0.1112< 0.2089 < 0.3519	0.063	0.3392< 0.5603 < 0.8626	0.0761< 0.1585 < 0.2860	1.4625< 6.0869 <15.9171	0.010*
ndhE	0.1410< 0.1945 < 0.2604	0.1291< 0.1807 < 0.2448	0.2726<2.5480 <7.7517	0.233	0.2267< 0.3294 < 0.4610	0.0428< 0.0855 < 0.1505	0.004**	0.2246< 0.3335 < 0.4742	0.0495< 0.0958 < 0.1640	0.009**	0.2099< 0.3121 < 0.4447	0.0428< 0.0856 < 0.1506	0.0429< 0.7893 <2.5430	0.014*
psaA	0.0234< 0.0350 < 0.0500	0.0226< 0.0341 < 0.0490	0.0048< 0.0980 < 0.4347	0.433	0.0257< 0.0495 < 0.0849	0.0171< 0.0291 < 0.0458	0.212	0.0238< 0.0476< 0.0836	0.0181< 0.0303 < 0.0471	0.300	0.0237< 0.0473< 0.0830	0.0170< 0.0290 < 0.0458	0.0043< 0.0965 < 0.4306	0.398
psaB	0.0257< 0.0369 < 0.0510	0.0252< 0.0365 < 0.0506	0.0028< 0.0596 < 0.2647	0.694	0.0416< 0.0660 < 0.0985	0.0111< 0.0208 < 0.0349	0.002**	0.0418< 0.0672 < 0.1013	0.0119< 0.0217 < 0.0359	0.003**	0.0416< 0.0668 < 0.1006	0.0111< 0.0208 < 0.0349	0.0017< 0.0543 < 0.2436	0.010*
psaI	0.0631< 0.1306 < 0.2409	0.0622< 0.1308< 0.2373	0.0000< 0.0046<1 0000.000 0	0.760	0.0373< 0.1514< 0.3973	0.0481< 0.1219< 0.2506	0.668	0.0376< 0.1528< 0.4009	0.0478< 0.1210< 0.2490	0.670	0.0376< 0.1523< 0.4007	0.0479< 0.1217< 0.2497	NA	0.904
psbA	0.0032< 0.0104 < 0.0241	0.0032< 0.0102 < 0.0241	NA	1	0.0023< 0.0140 < 0.0437	0.0014< 0.0082 < 0.0253	0.596	0.0024< 0.0142 < 0.0437	0.0014< 0.0082 < 0.0253	0.596	0.0023< 0.0141 < 0.0437	0.0014< 0.0082 < 0.0253	NA	0.869
psbC	0.0156< 0.0285 < 0.0469	0.0158< 0.0288 < 0.0475	0.0000< 0.0000 < 0.3739	0.603	0.0146< 0.0367 < 0.0744	0.0102< 0.0239 < 0.0462	0.466	0.0150< 0.0378 < 0.0767	0.0101< 0.0234 < 0.0454	0.419	0.0151< 0.0380 < 0.0770	0.0103< 0.0238 < 0.0462	0.0000< 0.0000 < 0.3623	0.642
psbH	0.0906< 0.1497 < 0.2308	0.0906< 0.1499 < 0.2309	NA	1	0.1249< 0.2743< 0.5147	0.0541< 0.1084< 0.1907	0.146	0.1249< 0.2745< 0.5149	0.0541< 0.1080< 0.1905	0.146	0.1249< 0.2745< 0.5145	0.0541< 0.1084< 0.1906	0.0000< 0.0379 <10000.	0.348
psbZ	0.0406< 0.0947 < 0.1838	0.0466< 0.1087 < 0.2114	0.0000< 0.0000 < 0.2226	0.239	0.0483< 0.1354 < 0.2931	0.0090< 0.0540 < 0.1673	0.324	0.0650< 0.1824 < 0.3959	0.0072< 0.0435 < 0.1350	0.134	0.0663< 0.1858 < 0.4036	0.0090< 0.0541 < 0.1685	0.0000< 0.0000 < 0.2090	0.217
rbcL	0.1113< 0.1424 < 0.1788	0.1144< 0.1466 < 0.1847	0.0007< 0.0491 < 0.2226	0.292	0.1030< 0.1578 < 0.2289	0.0989< 0.1350 < 0.1791	0.626	0.1119< 0.1732 < 0.2538	0.0957< 0.1302 < 0.1721	0.385	0.1132< 0.1752 < 0.2567	0.0989< 0.1351 < 0.1792	0.0003< 0.0480 < 0.2191	0.421
rpoA	0.2387< 0.2841 < 0.3351	0.2315< 0.2764 < 0.3270	3.5974< 18.9552 <48.5934	0.073	0.4492< 0.5724 < 0.7169	0.1390< 0.1797 < 0.2278	<<0.001***	0.4469< 0.5738 < 0.7231	0.1427< 0.1843 < 0.2332	<<0.001***	0.4270< 0.5493 < 0.6937	0.1391< 0.1798 < 0.2279	0.2715< 3.7569 <10.4348	<<0.001***
rpoB	0.1130< 0.1310 < 0.1508	0.1114< 0.1295 < 0.1494	0.0657< 0.2165 < 0.4884	0.486	0.1614< 0.1979 < 0.2395	0.0753< 0.0938 < 0.1151	<<0.001***	0.1635< 0.2018 < 0.2457	0.0770< 0.0955 < 0.1168	<<0.001***	0.1616< 0.1994 < 0.2429	0.0753< 0.0938 < 0.1151	0.0457< 0.1733 < 0.4014	<<0.001***
rpoC1	0.1399< 0.1669 < 0.1971	0.1403< 0.1679 < 0.1989	0.0402< 0.1431 < 0.3378	0.812	0.2503< 0.3177 < 0.3960	0.0791< 0.1035 < 0.1326	<<0.001***	0.2739< 0.3506 < 0.4407	0.0805< 0.1047 < 0.1331	<<0.001***	0.2732< 0.3497 < 0.4395	0.0793< 0.1038 < 0.1329	0.0292< 0.1218 < 0.2983	<<0.001***
rpoC2	0.2731< 0.2990 < 0.3266	0.2746< 0.3010 < 0.3290	0.0670< 0.2169 < 0.4587	0.612	0.3474< 0.3999 < 0.4576	0.2232< 0.2518 < 0.2826	<<0.001***	0.3581< 0.4136 < 0.4747	0.2226< 0.2508 < 0.2813	<<0.001***	0.3599< 0.4158 < 0.4770	0.2238< 0.2524 < 0.2834	0.0528< 0.1872 < 0.4042	0.001**

and *ndhB* compared to autotrophs (Table V-A). However, the most severe relaxation seemed to have occurred along the branch leading to the *Striga*/*Schwalbea*-clade showing an at least fivefold increased ω compared to that of hemiparasites. For the genes *atpF, cemA* and *ndhE*, a two-rate model fits to explain changes of ω between semi-heterotrophs and autotrophs, clearly separating *Striga* and *Schwalbea* from autotrophs. Such clear separation of the two hemiparasites from autotrophs also account for the photosystem gene *psaB* whose dN/dS-ratio is estimated fourfold higher than that of autotrophs and the subtree root. Nevertheless, ω below 0.1 indicates that hemiparasite *psaB* continuously evolves under strong purifying selection. Distinguishing changes in selection pressures in *rpo* genes of hemiparasites are significant with $p << 0.001$ for models assuming either two or three different ω. While the dN/dS-ratio of *rpoA* and *B* significantly differ between hemiparasites including their root and autotrophs, *Striga* and *Schwalbea* clearly depart from ω of autotrophs incl. subtree root in the genes *rpoC1* and *C2*. Although all changes are indicative of relaxed purifying selection in hemiparasites, greatest variation of ω occurs in *rpoA.*

2.6. Retained plastid encoded thylakoid ATP-synthase genes evolve at significantly lowered non-synonymous substitution rates in a subset of holoparasitic broomrapes.

We recently showed that a subset of broomrape holoparasites retain *atp*-genes (see Chapter IV for details) which opens up the possibility of an additional function of the plastid ATP-Synthase function beyond photosynthesis in these taxa. If functional constraints exist, we would expect to find several of the *atp* genes to evolve under purifying selection. Moreover, if *atp* gene function was essential for a certain period of time during holoparasitism, we would likely encounter great differences concerning the evolution of plastid-encoded ribosomal protein subunits in *Atp*-retaining taxa compared to others. Specifically, we would assume those to evolve more stringently as their ribosomal protein function is needed to guarantee correct translation of ATP-synthase subunits. Unexpectedly, relative reduction in synonymous and, particularly, non-synonymous rates are significantly lower in the retained ATP-synthase gene of *Phelipanche* compared to *Lindenbergia*. Similarly, the dN- and dS-rates reduction accounts for many genes of housekeeping functions (Fig V-6). The *Orobanche*-clade shows a comparable evolutionary pattern regarding rate changes like the *Phelipanche*-clade. However, rate changes are very complex across the different *atp*-subunits. Compared to *Lindenbergia*, *atpA, B, I* evolve at notably elevated synonymous rates in several yet not all holoparasites. Acceleration of dS is insignificant for *atpF* and *atpH; atpE*-rate increases in *P. purpurea* and *O. cumana*. Non-synonymous rates increase in *atpB* with high significance in *Phelipanche* species. The

Table V-B **Changes of selection in plastid *atp* genes of holoparasites.** ATP-synthase genes were analyzed regarding changes of selection between autotrophs and holoparasites. Clade-specific ω-values from 5 different models (mid-lines without "<") are provided with their respective lower ("*nn* <") and upper ("< *nn*") confidence levels including their respective p-value from LRT against the null model assuming a global ω. The best fit-model according to AIC-ranks is highlighted in blue. Clade A corresponds to the holoparasite subtree, clade B corresponds to autotrophic rest of the tree. Branch (T) denotes the root of the holoparasite clade. Asterisks indicate the significance level (<0.05*, <0.01**, <0.001***).

	Single rate model	2 selective paths									3 selective paths			
		Branch (T) vs. both clades			Clade A/branch vs. clade B			Clade A vs. clade B /branch			Clade A, Clade B, branch			
Gene	ω_{global}	ω_{A+B}	ω_T	p-value	ω_{A+T}	ω_B	p-value	ω_A	ω_{B+T}	p-value	ω_A	ω_B	ω_T	p-value
atpA	0.1168<	0.1168<	0.1168<	0.997	0.1602<	0.0703<	<<0.001***	0.1602<	0.0703<	<<0.001***	0.1608<	0.0703<	0.0000<	<<0.001***
	0.1363	0.1362 <	0.1362		0.1949	0.0910		0.1949	0.0910<		0.1955	0.0910	0.0000	
	< 0.1579	0.1578	< 0.1578		< 0.2342	< 0.1154		< 0.2342	0.1154		< 0.2351	< 0.1155	<1.3502	
atpB	0.1232<	0.1248<	0.0000<	0.174	0.1531<	0.0639<	<<0.001***	0.1531<	0.0636<	<<0.001***	0.1549<	0.0640<	0.0000<	<<0.001***
	0.1403	0.1421<	0.0000<		0.1777	0.0846		0.1777	0.0841		0.1798	0.0847	0.0000	
	< 0.1590	0.1610	0.4374		< 0.2048	< 0.1093		< 0.2048	< 0.1086		< 0.2072	< 0.1094	<0.3512	
atpE	0.2618<	0.2572<	0.0759<	0.269	0.3678<	0.1036<	0.001**	0.3628<	0.1102<	0.002**	0.3586<	0.1035<	0.0502<	0.003**
	0.3217	0.3165<	3.1235		0.4663	0.1609		0.4615	0.1694		0.4559	0.1607<	3.1622	
	<0.3900	0.3849	<10.8216		< 0.5809	< 0.2365		< 0.5773	< 0.2465		< 0.5704	0.2363	<11.0097	
atpF	0.4078<	0.4094<	0.0000<	0.770	0.6177<	0.1937<	<<0.001***	0.6293<	0.1971<	<<0.001***	0.6286<	0.1944<	0.0000<	<<0.001***
	0.4777	0.4803<	0.3342		0.7470	0.2609		0.7627	0.2641		0.7620	0.2619<	0.3455	
	< 0.5554	0.5590	< 1.2664		< 0.8942	< 0.3423		<0.9145	<0.3448		< 0.9135	0.3434	< 1.2674	
atpH	0.0202<	0.0174<	0.0000<	0.508	0.0343<	0.0006<	0.028*	0.0363<	0.0006<	0.027*	0.0369<	0.0006<	0.0000<	0.0849
	0.0442	0.0407<	0.1436		0.0800	0.0107		0.0854	0.0105		0.0861	0.0106	0.0000	
	< 0.0823	0.0785	< 0.7041		< 0.1552	< 0.0468		< 0.1664	< 0.0460		< 0.1670	< 0.0468	< 0.5234	
atpI	0.1275<	0.1298<	0.0000<	0.126	0.1966<	0.0427<	<<0.001***	0.2038<	0.0412<	<<0.001***	0.2046<	0.0427<	0.0000<	<<0.001***
	0.1565	0.1591<	0.0000		0.2479	0.0677		0.2569	0.0654		0.2579	0.0676	0.0000	
	< 0.1896	0.1929	< 0.1811		< 0.3076	< 0.1010		< 0.3188	< 0.0976		< 0.3200	< 0.1010	< 0.1638	

remainder *atp* genes evolve with smaller non-synonymous relative rates in *P. ramosa* and *P. lavandulacea*, whereas *P. purpurea* exhibits slightly higher dN values, although this is only significant in case of the *atpA* gene. *Orobanche* species either share a similar relative non-synonymous rate as *Lindenbergia* or exhibits slightly though not significantly lower rates. Mutational rates are slightly higher in *P. purpurea* than in *P. ramosa* and *P. lavandulacea*. Decreasing rates of particularly non-synonymous rates as well as similar levels of accelerations in dS and dN could constitute a stabilizing factor strongly suggesting selection on protein level in *atp*-genes of at least some holoparasites (e.g. *Phelipanche*). As described above (section V-2.5), we evaluated selection changes between the holoparasites subtree and autotrophs. All *atp*-genes of holoparasites significantly differ in ω from autotrophs (Table V-B, Supplemental material: Table SV-C). Overlapping ranges of ω between autotrophs and holoparasites are estimated for *atpH*, whereas estimates for the remainder genes are clearly non-overlapping. The ratio of dN/dS of *atpA,F,H,I* in *Orobanche* and *Phelipanche* species are clearly separate from autotrophs including the holoparasite root. In contrast, estimates of ω shared between holoparasites and the holoparasite root distinct from that of autotrophs can be assumed for *atpB* and *atpE*. With the exception of *atpF*, estimate of ω point in all cases towards purifying selection (Table V-B), although relaxed compared to autotrophs. Nevertheless these results suggest that selection of *atp*-genes in holoparasites occurs on protein level.

3. DISCUSSION

We analyzed the evolutionary patterns of substitution rate variation and its effects upon selection pressures of all plastid genes between autotrophic and photosynthetic and non-photosynthetic heterotrophic broomrape species. We could clearly demonstrate that parasitic broomrape lineages exhibit a great diversity of relative nucleotide substitution rates that accompanies the transition from autotrophy to a hemiparasitic way of life. Plastid genes in the obligate photosynthetic heterotrophs *Schwalbea* and *Striga* evolve at significantly elevated mutational rates compared to the autotrophic sister *Lindenbergia.* Rate acceleration affects not only genes involved in photosynthesis or photosynthesis-coupled pathways, but also shapes the evolution of the plastid genetic apparatus. Overall substitution rates are on average remarkable higher in *Striga* compared to *Schwalbea*. It has been suggested earlier, based upon single-gene analyses that plastid rate acceleration also circumscribes several more hemiparasitic species that are closely related to *Striga* (13, 14). Great differences between relative substitutional changes exist in several autotrophic lineages with highly reconfigured genomes compared to other angiosperms (6, 8, 10). The high degree of structural reconfiguration of the plastid genome in *Striga* (36; S. Wicke et al., unpublished data) may likewise correlate with the relatively higher rates in *Striga* compared to *Schwalbea*, too. Strikingly and evident in both taxa, non-synonymous rates are elevated suggesting a notable relaxation of functional constraints in the plastid genome. The transition to a heterotrophic lifestyle apparently allows relaxation of purifying selection in photosynthesis genes and effects genes of housekeeping functions alike. We observe relaxed purifying selection in photosystem genes as well as in its assembly factors. Pseudogenization and gene loss of several plastid *Ndh*-complex subunits has already started in both hemiparasites. Our analyses imply that several more genes are in the process of being functionally lost; particularly subunits for the plastid encoded polymerase are candidates for genes that gradually lose functionality. All *rpo* genes evolve under relaxed purifying selection on *Schwalbea* and *Striga*. They also exhibit several large indels and premature stop codons that put their function into question (see chapter IV). Relaxed purifying selection in plastid encoded-polymerase subunits suggests that significant changes of selection might have also occurred in translation-relevant elements (e.g. *rpl*/*rps*-genes), especially since some of these exhibit highly significant increases in nonsynonymous substitutions (Fig. V-5).

So far, no clear evolutionary pattern explaining the extreme difference among parasitic Orobanchaceae could be detected based upon single-gene analyses (12, 21, 27, 28). Recently, we showed that plastid gene retention in holoparasite plastomes is mainly influenced by conserved neighboring elements and the operon arrangement of plastid genes (Chapter IV). A similar effect seems to apply also for rate variation among a subset

of genes in both hemiparasites and holoparasites, in particular regarding the evolution of mutational rates. We observe similar levels of rate elevation or reduction among genes that are encoded within one transcription unit (e.g. *psaA/psaB; psbC/D, rpoB/C1/C2, rps7/12, rps2/atpA/F/H/I*). Transcription and repair systems may be influencing factors concerting the evolution of genes within one transcription unit. A close interrelation between polymerases and transcribed genes has been suggested earlier for the nuclear genome implying strong co-evolutionary patterns (41). In consequence, we may assume that those aspects also act on plastid gene evolution. Relaxed purifying selection on the plastid-encoded polymerase (PEP) transcribing the majority of photosynthesis elements might eventually lead to elevated rates of nucleotide substitutions and could, subsequently influence selectional patterns in photosynthesis-relevant genes – or vice versa.

Rate variation between different genes in holoparasites and among different lineages of holoparasites is extremely complex in holo-heterotrophs. Obviously lineage-effects strongly influence mutational rates in the plastid genome of broomrape species. In contrast, generation time effects are rather unsuitable to explain rate variation among holo-heterotrophs, or at best represent a negative correlation. *Epifagus, Conopholis, Cistanche,* and *Boulardia* parasitize perennials. Thus, they exhibit a rather perennial lifestyle as well and we would assume their rates to be slower than parasites with prominently more annual hosts. However, broomrape perennials investigated here exhibit higher mutational rates compared to *Orobanche* and *Phelipanche* including species parasitizing annual hosts, thus being at least facultatively annuals themselves. An interrelation of molecular machineries involved in the maintenance of plastid genomes (e.g. DNA repair systems) encoded by nuclear elements may contribute substantially to the rate evolution, as known with respect to the structural evolution of the plastid chromosome (42–44). In contrast to the current observations derived from a selected amount of angiosperm plastomes (6, 4, 8, 10) and photosynthetic heterotrophs, relative substitution rates do not significantly accelerate in holoparasites with highly reconfigured plastomes, but decrease compared to relatives with conservative plastome structures. We could clearly show that holoparasites with atypical chromosomal structures evolve at either similar mutational rates or show significant trends towards lower relative changes. This phenomenon applies not only to housekeeping genes, but also to remnant genes of the photosynthetic apparatus. The rate reduction in both gene classes is particularly unexpected since the same genes exhibit elevated rates and relaxed purifying selection in the hemiparasites *Striga* and *Schwalbea*.

Reasons for rate reduction in holoparasites after phases of rate acceleration are unknown so far. The eventual deletion of all dispensable regions and, thus, deletion of freely evolving segments may contribute a "stabilizing" factor in that it reconstitutes the compactness of the plastid chromosome with a well-balanced proportion of regions

sharing similar evolutionary patterns, especially regarding selection. Although not tested herein, lineages with a proportion of coding to non-coding plastid regions similar to that of autotrophs (see Chapter IV) appear to evolve with lower rates than those holoparasites departing more strongly from autotrophs. The arrangement of plastid genes in polycistronic transcription units may notably affect and influence the degree of substitutional rate changes among genes of the same functional complex ("operon-effect"). The vicinity of a gene to conserved neighboring genes or elements ("neighboring-gene effect") may also play a role in shaping rate evolution and selectional changes among plastid genes. For instance, differences of substitutional rates and selection of *rpo* genes of hemiparasites may be due to the different localizations of the four subunits within the plastid genome. The gene *rpo*A is located nearly 50 kb afar from *rpoB*/*C1*/*C2.* The latter are encoded together within the same transcription unit and share similar patterns of rate elevation and overlapping ranges of ω (section V-2.2. and V-2.4.). A similar observation applies for *atp* genes of holoparasites. *AtpB* and *atpE* are encoded within one operon, whereas the remainder *atp* genes are located in a multi-functional operon together with the ribosomal protein genes *rps2*, which is an essential gene (45) that also flanks the *rpoB*/*C1*/*C2* operon (Chapter IV). Follow-up analyses will be required to evaluate aspects of operon-specific effects of rate elevations and their impact on selectional changes.

Operon-effect and neighboring-gene effect are not mutually exclusive. Nevertheless, they appear to only insufficiently explain the long retention and continued evolution of holoparasite *atp* genes. We demonstrated here that retained plastid thylakoid ATP-synthase genes evolve with significantly reduced relative non-synonymous substitution rates in species of *Orobanche* and *Phelipanche*. Even more, we could show that despite a relaxation compared to autotrophs, holoparasite *atp* genes still underlie purifying selection, which corroborates our hypothesis that ATP-synthase subunits may be potentially functional (Chapter IV). Thus, another reasonable hypothesis for gene retention and "secondary" rate reduction in *atp*- and housekeeping genes of holoparasites could be based on the retention of gene function after the loss of photosynthesis. In autotrophs, the thylakoid ATP-Synthase complex forms ATP using the proton gradient generated by the photosynthetic electron flux (see Chapter II, IV see for more details). In some holoparasites, however, the complex might still be essential decoupled from photosynthesis for ATP-generation or has been for a longer period of time during the evolution of *Orobanche* and *Phelipanche.* If *atp* genes were indeed required in some of the holoparasite lineages, this may also explain rate reduction in plastid ribosomal elements (*rpl*/*rps-genes*) required to efficiently translate ATP-synthase subunits. In consequence to our findings here, we performed a BLAST-search against transcriptome libraries from *Phelipanche aegyptiaca* (http://ppgp.huck.psu.edu) to check for expressed nuclear encoded plastid ATP-synthase subunits. Although validation is still underway, we detected several

unigene fragments with high similarity to both the nuclear encoded gamma and delta subunits of the plastid ATP-synthase complex. Wickett et al. recently demonstrated that the non-photosynthetic *P. aegyptiaeca* holds upright the import of some photosynthesis-relevant elements (e.g. enzymes for chlorophyll synthesis) (46). It is thus likely that ATP-generation by the thylakoid ATP-synthase is maintained as well. A thorough experimental analyses will be required to evaluate whether expression of both nuclear and plastid *atp*-subunits in members of *Phelipanche* and *Orobanche* species also results in putatively functional protein complexes.

4. MATERIAL AND METHODS

4.1. Taxon sampling and plastome sequencing.

We reconstructed the genomes of four parasitic species (*Striga hermonthica, Orobanche cumana, O. pancicii, Phelipanche lavandulacea;* Table SV-C), in addition to those sequenced earlier (22; Chapter III). The dataset eventually comprised 15 Orobanchaceae plus three outgroup taxa: *Nicotiana tabacum* (Z000401; 47), *Mimulus guttatus* (Mimulus Genome Project, DoE Joint Genome Institute) and *Anthirrhinum majus* (GQ996966-GQ997048; 48). (Plastome sequencing has been performed via two different approaches using fosmid libraries (49), and shotgun-pyrosequencing from total genomic DNA. A detailed description of the sequencing procedures, assembly and sequence finishing as well as corresponding voucher information is provided in Chapter IV. Annotation of protein coding genes has been carried out using DOGMA (Wyman et al. 2007) with manual modifications.

Table SV-C **Plant material used for plastome sequencing**. Information of the source of plant material and the respective voucher information summarized for all herein newly sequenced Orobanchaceae species sorted in alphabetical order. Information on the applied sequencing method per taxon is also provided. [Abbreviations: WGSP – whole genome shotgun pyrosequencing, FSS – fosmid clone shotgun Sanger-sequencing, FPS – fosmid-clone pyrosequencing]

Taxon name	Source/ Voucher	Sequencing method
Orobanche cumana	Cultivated on *Vicia faba* at at the Botanical Garden Bonn; voucher deposited as S. Wicke #OC41 at the Bonn University Herbarium.	fosmid libraries: FSS; WGSP
Orobanche pancicii	Collected in Styria, Austria, where it was parasitizing *Knautia drymeia*, voucher deposited as G. Schneeweiss 42, Vienna University Herbarium	WGSP
Phelipanche lavandulacea	wild collection, parasitizing *Bituminaria bituminosa*, voucher deposited as Schönswetter & Tribsch, 12761 at the Bonn University Herbarium.	WGSP
Striga hermonthica	DNA received from Y. Zhang/W. dePamphilis (PennState University); voucher deposited in the private herbarium of C.W. dePamphilis.	fosmid libraries : FSS and FSP; WGSP

4.2. Tree reconstruction.

A concatenated dataset of 77 plastid protein coding genes (see Supplemental material: Table SV-A for details) have been used to reconstruct the phylogenetic relationships of Orobanchaceae to provide a sound basis for hypothesis testing. The two largest plastid genes y*cf1* and y*cf2* have been excluded because of their susceptibility to sequencing and assembly errors. Absence of a gene in parasitic lineages was treated as indel. The dataset has been aligned using MAFFT (50, 51) under the iterative method FFT-NS-I; the genes *accD* and *clpP* were aligned in E-INS-I mode in order to allow for longer gaps between stretches of conserved elements. Remaining ambiguous alignment positions have been manually refined in PhyDE (www.phyde.de). Both the original MAFFT output file and the refined alignments are available from the first author upon request.

Using PAUP 4.0b, a maximum likelihood (ML) tree was reconstructed using the GTR+ Γ+I model selected by the AIC in ModelTest 3.7 (52) we used four rate categories, and the gamma shape parameters, proportion of invariable sites, nucleotide frequencies and transition rates of the GTR-model were optimized via ML. 500 bootstrap replicates were run with the same settings. Two runs of each one million generations was run with eight chains each in MrBayes. Chain temperature was set to 0.2, and each chain was sampled every 1000th generations. 10% of trees were discarded as burn-in fraction. Treegraph 2 (53) was employed to visualize the results.

4.3. Analysis of mutational rates and hypothesis testing.

Mutational rates were analyzed using HyPhy (54) based upon the ML-tree. Using custom HyPhy batch files, relative nucleotide substitution rates were estimated for every plastid gene between *Lindenbergia* and two lamiid outgroup taxa (*Anthirrhinum majus, Mimulus guttatus*) using *Nicotiana* as outgroup. In order to evaluate the significance of rate differences between two species, the log-likelihoods of an unconstrained (i.e. allowing individual rates per taxon) and constrained model (i.e. rates on branches of interest were set equal) were compared in a likelihood ratio tests. For every gene, an individual likelihood function was build and subsequently optimized for both the constrained and unconstrained model, using the ML topology and a GTR+Γ substitution matrix. For evaluation of mutational rate changes between photosynthetic and holo-heterotrophic Orobanchaceae, relative substitution rate tests have been performed as described above for *Lindenbergia* and all parasitic taxa for all genes retained in parasite plastomes. Similarly, substitution rates were analyzed between pairs of hemiparasites (*Schwalbea, Striga*) and holoparasites (nine taxa).

Selection pressures were examined via the ratio ω of nonsynonymous (dN) to synonymous (dS) substitution rates. Normally, ω is low in genes under purifying selection (ω << 1), while higher values of ω (ω >> 1) indicate relaxed evolutionary pressure. The significance of differences in dS, dN and ω, respectively, was addressed via LRT as described above, using MG94 as codon model. Changes of ω were evaluated similarly by comparing constrained (ω set equal for the two taxa) and unconstrained models in likelihood ratio test. Furthermore, we tested genewise selectional changes between different clades using the batch script "SelectionLRT.bf" provided in the HyPhy package (40). In order to test differences between autotrophs (*Nicotiana, Anthirrhinum, Mimulus*, and *Lindenbergia*) and hemiparasites (*Striga, Schwalbea*), we excluded the holoparasites and used the following tree topology: (*Nicotiana*, (*Anthirrhinum*, *Mimulus*), (*Lindenbergia*, (*Striga*, *Schwalbea*))). For LRTs between photosynthetic plants and holoparasites, we excluded holoparasites lacking the complete set of *atp* genes (i.e. *Epifagus, Conopholis, Cistanche, Boulardia, O. gracilis*). The tree used was as follows: (*Nicotiana*, (*Mimulus*, *Anthirrhinum*), (*Lindenbergia*, ((*Striga*, *Schwalbea*), ((*O. cumana*, (*O. crenata*, *O. pancicii*)), (*Myzorrhiza*, (*P. purpurea*, (*P. ramosa*, *P. lavandulacea*)))))))).

5. ACKNOWLEDGMENTS

We would like to thank Monika Ballmann and Karola Maul (both University of Bonn), Lena Landherr-Sheaffer and Norman Wickett (both PennState University), Susanne Sangenstedt and Bastian Schäferhoff (University of Münster) for excellent technical and bioinformatics assistance. Sincere thanks is due to Klaus Bahr, Wolfram Lobin (both Botanical Garden Bonn), Barbara Ditsch (Botanical Garden Dresden), Mats Hjertson (Uppsala University), Kay Kirkman (Joseph W. Jones Ecological Research Center) for cultivating and/or providing fresh material. This study received financial support from the Austrian Science Fund (FWF grant 19404 to G.M.S), the German Science Foundation (DFG grant MU2875/2, to K.F.M., QU153/2 to D.Q. and SRXX/X to S.S.R.), and the US National Science Foundation (N.S.F. grants DEB-0120709 and DBI-0701748 to C.W.D.). Financial support to S.W. from the University of Vienna (Austria) and the Botanical Society of America is gratefully acknowledged.

6. AUTHORS' CONTRIBUTIONS

S.W contributed to the conceptual layout of this study, carried out all wet and dry lab work for plastid genome sequencing, conceived of and performed hypothesis tests, analyzed the results and wrote the manuscript. K.F.M., D.Q., G.M.S, C.W.D. contributed to the conceptual layout of this study, critically discussed hypothesis tests and results, and critically revised the manuscript, Y.Z. and B.S. contributed to wet and dry lab works for *Striga, Conopholis, O. cumana*. S.S.R. critically revised the manuscript.

This chapter will be published in a modified version as a research article in a peer-reviewed journal. The tentative author list and title are as follows:

> *Wicke S, Müller KF, Quandt D, dePamphilis CW, Zhang Y, Bellot S, Renner SS, and Schneeweiss GM. Genome-wide analyses of plastid mutation rates uncover extreme relaxation of selective constraints in hemiparasites and reveal purifying selection in ATP-synthase genes of holoparasitic plants.*

7. REFERENCES

1. Wolfe KH, Li WH, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc Natl Acad Sci USA* 84:9054 - 9058.
2. Gaut BS, Muse SV, Clark WD, Clegg MT (1992) Relative rates of nucleotide substitution at the rbcL locus of monocotyledonous plants. *J Mol Evol* 35:292-303.
3. Wolf PG et al. (2010) The evolution of chloroplast genes and genomes in ferns. *Plant Mol Biol*:1-11.
4. Perry AS, Wolfe KH (2002) Nucleotide substitution rates in legume chloroplast DNA depend on the presence of the inverted repeat. *J Mol Evol* 55:501-508.
5. Schäferhoff B (2011) Carnivory in Lamiales: Phylogeny, taxonomy, and chloroplast genome evolution. Dissertation. Westfälische-Wilhems-Universität Münster.
6. Jansen RK et al. (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc Natl Acad Sci USA* 104:19369 - 19374.
7. Chris Blazier J, Guisinger-Bellian MM, Jansen RK (2011) Recent loss of plastid-encoded *ndh* genes within *Erodium* (Geraniaceae). *Plant Mol Biol* 76:1-10.
8. Guisinger MM, Kuehl JV, Boore JL, Jansen RK (2008) Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. *Proc Natl Acad Sci USA* 105:18424-18429.
9. Magee AM et al. (2010) Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Res* 20:1700-1710.
10. Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D (2011) The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol Biol* 76:273-297.
11. Müller K et al. (2004) Evolution of carnivory in Lentibulariaceae and the Lamiales. *Plant Biol (Stuttg)* 6:477-490.
12. Wolfe KH, Morden CW, Ems SC, Palmer JD (1992) Rapid evolution of the plastid translational apparatus in a nonphotosynthetic plant: loss or accelerated sequence evolution of tRNA and ribosomal protein genes. *J Mol Evol* 35:304 - 317.
13. dePamphilis CW, Young ND, Wolfe AD (1997) Evolution of plastid gene *rps2* in a lineage of hemiparasitic and holoparasitic plants: Many losses of photosynthesis and complex patterns of rate variation. *Proc Natl Acad Sci USA* 94:7367-7372.

14. Young ND, dePamphilis CW (2005) Rate variation in parasitic plants: correlated and uncorrelated patterns among plastid genes of different function. *BMC Evol Biol* 5:16.

15. Logacheva MD, Schelkunov MI, Penin AA (2011) Sequencing and analysis of plastid genome in mycoheterotrophic orchid *Neottia nidus-avis*. *Genome Biol Evol* 3:1296-1303.

16. Muse SV (2000) Examining rates and patterns of nucleotide substitution in plants. *Plant Molecular Biology* 42:25-43.

17. Albert VA, Jobson RW, Michael TP, Taylor DJ (2010) The carnivorous bladderwort (*Utricularia*, Lentibulariaceae): a system inflates. *J Exp Bot* 61:5 -9.

18. Leebens-Mack JH, dePamphilis CW (2002) Power analysis of tests for loss of selective constraint in cave crayfish and nonphotosynthetic plant lineages. *Mol Biol Evol* 19:1292 - 1302.

19. Muse S, Gaut B (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* 11:715-724.

20. Kosakovsky-Pond SL, Frost SDW (2005) A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Mol Biol Evol* 22:478 -485.

21. Wolfe AD, dePamphilis CW (1998) The effect of relaxed functional constraints on the photosynthetic gene *rbcL* in photosynthetic and nonphotosynthetic parasitic plants. *Mol Biol Evol* 15:1243 - 1258.

22. Wolfe KH, Morden CW, Palmer JD (1992) Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proc Natl Acad Sci USA* 89:10648-10652.

23. Haberhausen G, Zetsche K (1994) Functional loss of *ndh* genes in an otherwise relatively unaltered plastid genome of the holoparasitic flowering plant *Cuscuta reflexa*. *Plant Mol Biol* 24:217 - 222.

24. Nickrent D, García M (2009) On the brink of holoparasitism: Plastome evolution in Dwarf Mistletoes (*Arceuthobium*, Viscaceae). *J Mol Evol* 68:603-615.

25. Nickrent DL, Duff RJ, Konings DAM (1997) Structural analyses of plastid-derived 16S rRNAs in holoparasitic angiosperms. *Plant Molecular Biology* 34:731 - 743.

26. Delavault PM, Sakanyan V, Thalouarn P (1995) Divergent evolution of two plastid genes, *rbcL* and *atpB*, in a non-photosynthetic parasitic plant. *Plant Mol Biol* 29:1071 - 1079.

27. Young ND, dePamphilis CW (2000) Purifying selection detected in the plastid gene matK and flanking ribozyme regions within a group II intron of nonphotosynthetic plants. *Mol Biol Evol* 17:1933 -1941.

28. Randle CP, Wolfe AD (2005) The evolution and expression of RBCL in holoparasitic sister-genera *Harveya* and *Hyobanche* (Orobanchaceae). *Am J Bot* 92:1575-1585.

29. Guisinger MM, Kuehl JV, Boore JL, Jansen RK (2011) Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. *Mol Biol Evol* 28:583-600.

30. Knauf U, Hachtel W (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.

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

32. McNeal JR, Kuehl J, Boore J, de Pamphilis C (2007) Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus *Cuscuta*. *BMC Plant Biol* 7:57.

33. Wickett NJ et al. (2008) Functional Gene losses occur with minimal size reduction in the plastid genome of the parasitic liverwort *Aneura mirabilis*. *Mol Biol Evol* 25:393-401.

34. Westwood JH, Yoder JI, Timko MP, dePamphilis CW (2010) The evolution of parasitism in plants. *Trends Plant Sci* 15:227-235.

35. Bennett JR, Mathews S (2006) Phylogeny of the parasitic plant family Orobanchaceae inferred from phytochrome A. *Am J Bot* 93:1039-1051.

36. Downie SR, Palmer JD (1992) Restriction site mapping of the chloroplast DNA inverted repeat - a molecular phylogeny of the Asteridae. *Ann Mo Bot Gard* 79:266-283.

37. Schneeweiss GM, Colwell A, Park J-M, Jang C-G, Stuessy TF (2004) Phylogeny of holoparasitic Orobanche (Orobanchaceae) inferred from nuclear ITS sequences. *Mol Phylogenet Evol* 30:465-478.

38. Park J-M, Manen J-F, Colwell A, Schneeweiss GM (2008) A plastid gene phylogeny of the non-photosynthetic parasitic *Orobanche* (Orobanchaceae) and related genera. *J Plant Res* 121:365-376.

39. Frost SDW et al. (2005) Characterization of human immunodeficiency virus type 1 (HIV-1) envelope variation and neutralizing antibody responses during transmission of HIV-1 subtype B. *J Virol* 79:6523 -6527.

40. Kosakovsky-Pond SL, Poon AFY, Frost SDW (2007) in *The phylogenetic handbook: A practical approach to DNA and protein phylogeny*, pp 419-450.

41. Carter R, Drouin G (2009) The evolutionary rates of Eukaryotic RNA polymerases and of their transcription factors are affected by the level of concerted evolution of the genes they transcribe. *Mol Biol Evol* 26:2515 -2520.

42. Day A, Madesis P (2007) in *Cell and Molecular Biology of Plastids*, Topics in Current Genetics. (Springer, Berlin / Heidelberg), pp 65-119. Available at: http://dx.doi.org/10.1007/4735_2007_0231.

43. Gray B, Ahner B, Hanson M (2009) Extensive homologous recombination between introduced and native regulatory plastid DNA elements in transplastomic plants. *Transgenic Res* 18:559-572.

44. Maréchal A, Brisson N (2010) Recombination and the maintenance of plant organelle genome stability. *New Phytol* 186:299-317.

45. Fleischmann TT et al. (2011) Nonessential plastid-encoded ribosomal proteins in tobacco: A developmental role for plastid translation and implications for reductive genome evolution. *Plant Cell*. Doi: 10.1105/tpc.111.088906.

46. Wickett NJ et al. (2011) Transcriptomes of the parasitic plant family Orobanchaceae reveal surprising conservation of chlorophyll synthesis. *Curr Biol* 21:2098-2104.

47. Shinozaki K et al. (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J* 5:2043 - 2049.

48. Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE (2010) Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc Natl Acad Sci USA* 107:4623-4628.

49. McNeal JR et al. (2006) Using partial genomic fosmid libraries for sequencing complete organellar genomes. *Biotechniques* 41:69-73.

50. Katoh K, Misawa K, Kuma K-ichi, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucl Acids Res* 30:3059-3066.

51. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33:511-518.

52. Posada D, Crandall KA (2001) Modeltest v3.06: testing the model of DNA substitution. . *Bioinformatics* 14:817-818.

53. Stöver B, Müller KF (2010) TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics* 11:7.

54. Kosakovsky-Pond SL, Frost SDW, Muse SV (2004) HyPhy: hypothesis testing using phylogenies. *Bioinformatics*:bti079.

8. SUPPLEMENTAL MATERIAL

Table SV-A **Overview of the gene content of 15 autotrophic and heterotrophic Orobanchaceae as well as two autotrophic Scophulariacean taxa and *Nicotiana*.** Presence ("x") or absence ("lost") of 77 plastid protein-coding genes is summarized for 18 investigated photosynthetic and nonphotosynthetic taxa. Cases in which only partial genes fragments were used in the analysis are indicated by either "5'" – or "3'"referring to the fragment/exon included. The total number of genes per species is summarized below as well as those genes that exhibit severe reading frame distortions by indels or premature stop codons ("Put. Ψ"). The table continues the next page.

Abbreviations: *Nic – Nicotiana tabacum, Ant – Anthirrhinum majus, Mim – Mimulus guttatus, Lin – Lindenbergia philippensis, Str – Striga hermonthica, Sch – Schwalbea americana, Epi – Epifagus virginiana, Con – Conopholis americana, Cis – Cistanche phelypaea, Bou – Boulardia macrolepis, Ocu – Orobanche cumana, Ogr – O. gracilis, Ocr – O. crenata, Opa – O. pancicii, Myz – Myzorrhiza californica, Ppu – Phelipanche purpurea, Pra – P. ramosa, Pla – P. lavandulacea;* NA – not available.

Gene ID	*Nic*	*Ant*	*Mim*	*Lin*	*Str*	*Sch*	*Epi*	*Con*	*Cis*	*Bou*	*Ocu*	*Ogr*	*Ocr*	*Opa*	*Myz*	*Ppu*	*Pra*	*Pla*
accD	x	x	x	x	lost	lost	x	x	lost	x	x	x	x	lost	x	lost	lost	lost
atpA	x	x	x	x	x	x	lost	lost	lost	lost	x	lost	x	x	x	x	x	x
atpB	x	x	x	x	x	x	lost	lost	lost	lost	x	lost	x	x	x	x	x	x
atpE	x	x	x	x	x	x	lost	lost	lost	lost	x	lost	x	x	x	x	x	x
atpF	x	x	x	x	x	x	lost	lost	lost	lost	x	x	x	x	x	x	x	x
atpH	x	x	x	x	x	x	lost	lost	lost	lost	x	lost	x	x	x	x	x	x
atpI	x	x	x	x	x	x	lost	lost	lost	lost	x	lost	x	x	x	x	x	x
ccsA	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
cemA	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
clpP	x	x	x	x	x	x	x	x	lost	x	lost	x	x	lost	x	3'	lost	x
infA	lost	x	x	x	x	x	lost	lost	x	x	x	lost	lost	lost	x	x	x	x
matK	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
ndhA	x	x	x	x	5'	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
ndhB	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
ndhC	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
ndhD	x	x	x	x	lost	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
ndhE	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
ndhF	x	x	x	x	5'-	5'	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
ndhG	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
ndhH	x	x	x	x	lost	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
ndhI	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
ndhJ	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
ndhK	x	x	x	x	lost	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
petA	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
petB	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
petD	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
petG	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
petL	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
petN	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psaA	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psaB	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psaC	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psaI	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psaJ	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psbA	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psbB	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psbC	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psbD	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psbE	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psbF	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psbH	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost

Gene ID	*Nic*	*Ant*	*Mim*	*Lin*	*Str*	*Sch*	*Epi*	*Con*	*Cis*	*Bou*	*Ocu*	*Ogr*	*Ocr*	*Opa*	*Myz*	*Ppu*	*Pra*	*Pla*
psbI	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psbJ	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psbK	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psbL	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psbM	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psbN	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psbT	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
psbZ	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
rbcL	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
rpl14	x	x	x	x	x	x	lost	x	x	x	x	x	x	x	x	x	x	x
rpl16	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rpl2	x	x	3'	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rpl20	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rpl22	x	x	x	x	x	x	lost	x	x	x	x	x	x	x	x	x	x	x
rpl23	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	x	lost	lost	lost
rpl32	x	x	x	x	x	x	lost	lost	x	x	x	x	x	x	x	x	x	x
rpl33	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rpl36	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rpoA	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
rpoB	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
rpoC1	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
rpoC2	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
rps11	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rps12	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rps14	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rps15	x	x	x	x	x	x	lost	lost	x	x	x	x	x	x	x	x	x	x
rps16	x	x	3'	x	x	x	lost	lost	x	lost	lost	lost	x	lost	x	lost	lost	lost
rps18	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rps19	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rps2	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rps3	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	3′	x	x
rps4	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rps7	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rps8	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
ycf3	x	x	NA	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
ycf4	x	x	x	x	x	x	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost	lost
No. genes	**76**	**77**	**76**	**77**	**70**	**70**	**20**	**21**	**25**	**24**	**32**	**25**	**33**	**30**	**35**	**32**	**31**	**32**
Put. Ψ					*rpoA-C, rps18*				*rpl22*	*rpl22*			*atpF, rps19*			*rps3, rps15*	*atpB, rps3*	*rps3*

Table SV-B **Results of relative rate tests between *Anthirrhinum majus* and *Lindenbergia*, and *Mimulus guttatus* and *Lindenbergia*, respectively.** LRT-results of relative nucleotide substitution rate estimates of *Lindenbergia* versus two outgroup taxa is summarized below for 77 plastid genes. Per test, the relative rate (μ) is provided per taxon as well as the LRT statistics including the p-value. Asterisks mark the significance level (<0.05*, <0.01**, <0.001***). The table continues the next page.

	Lindenbergia vs. Anthirrhinum				*Lindenbergia vs. Mimulus*			
Gene	***Anthirrhinum* μ**	***Lindenbergia* μ**	**LRT**	**p-value**	***Mimulus* μ**	***Lindenbergia* μ**	**LRT**	**p-value**
accD	0.000	0.047	0.021	<<0.001***	0.000	0.047	0.021	<<0.001***
atpA	0.054	0.015	0.034	<<0.001***	0.057	0.015	0.035	<<0.001***
atpB	0.079	0.038	0.058	0.005**	0.046	0.038	0.042	0.048
atpE	0.047	0.019	0.033	0.191	0.042	0.019	0.031	0.259
atpF	0.075	0.040	0.058	0.152	0.048	0.040	0.044	0.721
atpH	0.000	0.049	0.049	0.597	0.149	0.049	0.074	0.244
atpI	0.041	0.049	0.045	0.662	0.045	0.049	0.047	0.836
ccsA	0.097	0.081	0.089	0.495	0.098	0.081	0.089	0.485
cemA	0.055	0.029	0.042	0.110	0.044	0.029	0.037	0.340
clpP	0.061	0.011	0.036	0.002**	0.022	0.011	0.017	0.333
infA	0.049	0.059	0.054	0.766	0.039	0.059	0.049	0.528
matK	0.126	0.083	0.104	0.026*	0.139	0.083	0.110	0.005**
ndhA	0.089	0.064	0.078	0.274	0.073	0.064	0.069	0.690
ndhB	0.003	0.001	0.002	0.549	0.008	0.001	0.005	0.048*
ndhC	0.063	0.047	0.055	0.607	0.057	0.047	0.053	0.683
ndhD	0.115	0.061	0.089	0.002**	0.064	0.061	0.063	0.817
ndhE	0.038	0.030	0.034	0.786	0.066	0.030	0.050	0.257
ndhF	0.132	0.121	0.127	0.543	0.110	0.121	0.110	1.000
ndhG	0.107	0.032	0.085	0.184	0.062	0.033	0.061	0.990
ndhH	0.084	0.047	0.065	0.041*	0.076	0.047	0.061	0.103
ndhI	0.075	0.041	0.058	0.204	0.082	0.041	0.062	0.135
ndhJ	0.037	0.035	0.036	0.912	0.036	0.035	0.035	0.975
ndhK	0.105	0.042	0.072	0.008**	0.044	0.042	0.043	0.920
petA	0.058	0.096	0.077	0.084	0.069	0.096	0.082	0.223
petB	0.048	0.048	0.048	0.982	0.043	0.048	0.045	0.824
petD	0.039	0.031	0.035	0.688	0.036	0.031	0.034	0.815
petG	0.028	0.087	0.056	0.297	0.028	0.085	0.057	0.300
petL	0.195	0.095	0.145	0.403	0.000	0.096	0.047	0.095
petN	0.030	0.015	0.023	0.554	0.000	0.015	0.008	0.237
psaA	0.043	0.028	0.036	0.118	0.039	0.028	0.034	0.262
psaB	0.036	0.023	0.030	0.131	0.029	0.023	0.026	0.505
psaC	0.051	0.067	0.059	0.678	0.058	0.067	0.062	0.805
psaI	0.004	0.004	0.004	1.000	0.016	0.004	0.008	0.181
psaJ	0.043	0.065	0.054	0.648	0.000	0.065	0.024	0.051
psbA	0.017	0.009	0.013	0.186	0.019	0.009	0.014	0.145
psbB	0.069	0.049	0.059	0.167	0.050	0.049	0.050	0.925
psbC	0.022	0.017	0.020	0.489	0.021	0.017	0.019	0.595
psbD	0.039	0.024	0.031	0.227	0.015	0.024	0.019	0.405
psbE	0.007	0.003	0.004	0.602	0.003	0.003	0.003	1.000
psbF	0.027	0.000	0.013	0.238	0.054	0.000	0.027	0.095
psbH	0.132	0.096	0.115	0.588	0.096	0.097	0.096	0.999
psbI	0.074	0.000	0.036	0.040*	0.049	0.000	0.024	0.094
psbJ	0.033	0.033	0.033	0.994	0.102	0.033	0.067	0.301
psbK	0.069	0.060	0.064	0.848	0.000	0.060	0.029	0.039*
psbL	0.046	0.023	0.034	0.556	0.000	0.022	0.011	0.238
psbM	0.009	0.000	0.009	1.000	0.027	0.000	0.017	1.000
psbN	0.017	0.016	0.016	0.958	0.017	0.016	0.016	0.977
psbT	0.069	0.068	0.068	0.986	0.134	0.068	0.098	0.438
psbZ	0.018	0.036	0.027	0.559	0.036	0.036	0.036	0.991
rbcL	0.045	0.018	0.032	0.006**	0.021	0.018	0.019	0.733
rpl14	0.074	0.020	0.047	0.027*	0.033	0.020	0.027	0.475
rpl16	0.070	0.055	0.062	0.587	0.036	0.055	0.046	0.407

	Lindenbergia vs. Anthirrhinum				*Lindenbergia vs. Mimulus*			
Gene	***Anthirrhinum* μ**	***Lindenbergia* μ**	**LRT**	**p-value**	***Mimulus* μ**	***Lindenbergia* μ**	**LRT**	**p-value**
rpl2	0.009	0.012	0.010	0.712	0.011	0.012	0.012	0.970
rpl20	0.078	0.054	0.066	0.411	0.031	0.054	0.043	0.321
rpl22	0.213	0.097	0.158	0.016*	0.114	0.097	0.106	0.674
rpl23	0.000	0.005	0.003	0.241	0.010	0.005	0.008	0.555
rpl32	0.159	0.087	0.125	0.302	0.082	0.086	0.084	0.938
rpl33	0.064	0.000	0.032	0.018*	0.031	0.000	0.016	0.096
rpl36	0.066	0.000	0.033	0.040*	0.000	0.000	0.000	1.000
rpoA	0.068	0.054	0.061	0.382	0.069	0.054	0.061	0.365
rpoB	0.060	0.048	0.054	0.215	0.038	0.048	0.043	0.292
rpoC1	0.065	0.047	0.056	0.149	0.055	0.047	0.051	0.487
rpoC2	0.069	0.044	0.057	0.002**	0.060	0.044	0.052	0.053
rps11	0.135	0.093	0.113	0.310	0.076	0.093	0.085	0.632
rps12	0.005	0.005	0.005	0.978	0.020	0.005	0.013	0.163
rps14	0.010	0.044	0.028	0.130	0.061	0.044	0.053	0.592
rps15	0.216	0.116	0.164	0.123	0.116	0.116	0.116	0.986
rps16	0.034	0.030	0.032	0.876	0.083	0.030	0.052	0.141
rps18	0.000	0.058	0.032	0.008**	0.051	0.058	0.055	0.820
rps19	0.009	0.037	0.023	0.169	0.074	0.037	0.055	0.239
rps2	0.043	0.016	0.029	0.055	0.050	0.016	0.032	0.023*
rps3	0.059	0.019	0.039	0.020*	0.090	0.019	0.054	<<0.001***
rps4	0.083	0.045	0.064	0.120	0.046	0.045	0.046	0.954
rps7	0.005	0.005	0.005	0.994	0.000	0.005	0.003	0.239
rps8	0.053	0.023	0.038	0.157	0.036	0.023	0.030	0.515
ycf3	0.002	0.029	0.026	0.760	NA	NA	NA	NA
ycf4	0.047	0.035	0.041	0.522	0.013	0.035	0.024	0.118

Table SV-C **AIC ranks for model selection from analysis of selectional changes in selected plastid genes of hemiparasites and photosynthetic relatives.** AIC results from model tests are summarized below for the null model as well as three different two-rate models and one three-rate model. An asterisk marks the AIC-preferred model.

		AIC			
Gene ID	**Null model**	**Branch vs. clade A/B**	**Clade A/branch vs. clade B**	**Clade A vs. clade B/branch**	**Clade A, Clade B, branch**
atpA	6,501.43963	6,496.84706	6,490.08170	6,495.73469	6,489.02425*
atpB	6,385.13133	6,375.5291	6,371.44559	6,379.75655	6,367.92198*
atpF	2,594.86412	2,596.83152	2,589.65187	2,589.58009*	2,591.36673
atpI	2,973.87541*	2,975.87565	2,974.81835	2,974.81807	2,976.81796
cemA	3,405.10859	3,405.47553	3,401.11206*	3,402.05620	3,406.05230
ndhB	4,878.66534	4,876.74834	4,873.73095	4,877.21725	4,873.44697*
ndhE	1,602.54484	1,603.12339	1,596.32234*	1,597.80623	1,598.07437
psaA	8,583.48982*	8,584.87749	8,583.93549	8,584.41617	8,585.65088
psaB	8,415.08398	8,416.92924	8,407.84134*	8,408.33339	8,409.81163
psaI	465.91260*	467.81974	467.72875	467.73169	469.71195
psbA	3,763.81960*	3,765.82073	3,765.53924	3,765.53988	3,767.53878
psbC	5,164.75841*	5,166.48797	5,166.22615	5,166.10626	5,167.87238
psbH	1,045.33393	1,047.33396	1,045.22088	1,045.22083*	1,047.22085
psbZ	741.08152	741.69524	742.11071	740.83301*	742.03006
rbcL	5,977.70281*	5,978.59471	5,979.46489	5,978.94871	5,979.97022
rpoA	5,190.97589	5,189.76938	5,173.80485*	5,175.54759	5,174.68319
rpoB	14,103.34471	14,104.85971	14,088.65856*	14,089.32313	14,090.62086
rpoC1	9,129.19808	9,131.14184	9,106.97723	9104.58588*	9,106.53185
rpoC2	20,852.86209	20,854.60446	20,843.36754	20,842.05352*	20,843.82140

Table SV-D **AIC ranks for model selection from analysis of selectional changes in six plastid *atp* genes of holoparasites and photosynthetic plants.** AIC results from model tests are summarized below for the null model as well as three different two-rate models and one three-rate model. An asterisk marks the AIC-preferred model.

		AIC			
Gene ID	**Null model**	**Branch vs. clade A/B**	**Clade A/branch vs. clade B**	**Clade A vs. clade B /branch**	**Clade A, Clade B, branch**
atpA	9,214.92208	9,216.92206	9,201.37092	9,201.37086*	9,203.29706
atpB	10,767.72435	10,767.87987	10,751.18543*	10,752.11614	10,751.83252
atpE	2,909.63564	2,910.41609	2,900.92489*	2,902.05756	2,902.13225
atpF	4,316.90130	4,318.81600	4,302.28651	4,301.92039*	4,303.87154
atpH	1,267.25863	1,268.82206	1,264.43279	1,264.35333*	1,266.32535
atpI	4,704.52667	4,704.19500	4,684.76508	4,682.52079*	4,683.17921

– CHAPTER VI –

SUMMARY, CONCLUSIONS AND RESEARCH PROSPECTS

CONTENTS.

This chapter contains approximately 2,300 words.

1. SUMMARY AND CONCLUSIONS OF THIS WORK

The presented work aimed to illuminate evolutionary patterns of plastid genome reduction under relaxed selective pressure. To this end, the complete plastid chromosome sequences of 14 photosynthetic and non-photosynthetic species of the broomrape family (Orobanchaceae) were sequenced using both a fosmid-libraries approach as well as whole-genome shotgun pyrosequencing (WGSP). A detailed analysis of WGSP for the reconstruction of plastid chromosomes from plastid-unenriched DNA based upon simulated and empirical 454 datasets demonstrated that the proper choice of locus-specific amounts of short-read data reduces computational efforts and, more importantly, contributes substantially to ease and quality of the assembly process. By using a simple exponential decay model, this work suggests for the first time a method for *a priori* approximation of the optimal sequence pool size for the reconstruction of plastid genomes from WGSP datasets. Analysis of plastome structure, nucleotide substitution rates and selectional constraints among photosynthetic and holoparasitic broomrapes provides new insights into process of reductive plastome evolution showing that pseudogenization and the deletion of plastomic segments take alternative paths under relaxed selective constraints. The results obtained in the course of this work imply that not only the loss of photosynthesis initiates the process of dramatic plastome reduction. Notable changes in plastid chromosome structure, gene content and substitutional rates already accompany (or set in shortly after) the transition to a parasitic way of life in Orobanchaceae. Although causality remains unclear, the eventual loss of photosynthesis seeds excessive non-functionalization of plastid genes due to pseudogenization or deletion. Convergent gene losses do exist between Orobanchaceae and other lineages of non-photosynthetic plants [1–10]. This study shows for the first time that gene deletion does not occur free of constraints after the loss of selective pressures. The observations and results of this study are well-suitable to address the questions asked by dePamphilis et al. (1997, p7367):

1.1. "… is there some predictable order to the genetic changes associated with the loss of photosynthesis? […]"

In light of the results of the present study, gene deletion after the loss of photosynthesis is a highly complex process that does not follow a strict order. Results of the present thesis, however, indicate that there are certain trends in plastomic changes:

- Analyses of plastomes from hemiparasitic plants and ancestral genome reconstruction (ASR) suggest that, within Orobanchaceae, the thylakoid NAD(P)H-dehydrogenase complex encoded by *ndhA-K* probably displays the earliest functional loss. *Ndh*-gene loss was observed several times independently in plant lineages with a certain degree of heterotrophy [4,5,8,11–15], but occurs also in

autotrophs such as Pinaceae [16] and Geraniaceae [17]. Reverse genetic approaches revealed that functional plastid *ndh* subunits are not essential for plant survival under non-stress conditions [18]. Thus, their loss may be only of weak effect in heterotrophs relying on a host plant for nutritional supplies.

- The present work uncovered relaxed purifying selection in all genes for the plastid-encoded polymerase complex (PEP, encoded by *rpo* genes). Although other photosynthesis-related genes also evolve at elevated rates in *Striga* and *Schwalbea*, no other complex-wide relaxation of purifying selection was detected. Reconstruction of ancestral gene contents additionally suggests that the last common ancestor of the holoparasitic broomrape clade had probably already lost PEP function completely. Early pseudogenization of *rpo* genes was also reported in *Cuscuta* [5,4,19,20]. Similar to *ndh* genes, malfunction or loss of PEP can be compensated in that a nuclear-encoded polymerase takes over the transcription of plastid genes from alternative promoters [19,21–23]. In autotrophs, PEP transcription provides the necessary abundance of transcripts for efficient function of the photosynthesis light reaction complexes [24]. Thus, their function may be dispensable in an obligate heterotrophic lifestyle.
- This work demonstrated that deletion of genes for photosystems (*psa*, *psb* genes) and electron transport (*pet*, *atp* genes) strongly depends on the vicinity to conserved (housekeeping) elements. For the most recent common ancestor of broomrapes, ASR infers only five subunits as deleted from the plastome, but suggests pseudogenization of the majority of *psa*/*psb* and *pet* genes. In contrast, *atp* genes survived as potentially functional genes in several representatives. Similarly, *rbcL* function appears to be lost rather late during the evolution of holoparasites. Possibly, a putative alternative role decoupled from photosynthesis may attribute to *rbcL* retention [5,25–27], and may similarly preserve *atp*-gene functionality. Besides this, arrangement in multi-functional operons appears to protect photosynthesis elements from an early deletion. Thus, the survival probability of dispensable plastomic segments seems to decrease with an increasing distance to essential neighboring elements. This work suggests that neighboring genes and the "operon-effect" also influence the evolution of substitution rates and account for changes in selectional pressures upon subunits of functional complexes.
- The loss of plastid housekeeping gene from the plastomes of broomrape holoparasites (and also other non-photosynthetic plants) appears to not follow evident patterns. Except for *rps16* and *rpl23*, both of which are lost frequently in angiosperms [28–30], the kind of *rpl*/*rps*-gene and tRNA-losses is lineage-specific in Orobanchaceae and other non-photosynthetic plants affecting both essential and non-essential ribosomal protein genes and tRNAs [3–5,7,9,10,31]. The data obtained during this work may, however, shed light on factors influencing the longevity of

genes for the genetic apparatus. Functionality of photosynthesis elements *beyond* photosynthesis may be a major determining factor. Lineages retaining *atp* genes show on average less reduction in housekeeping genes (incl. tRNAs) than those without ATP-synthase genes. Ribosome function appears to be maintained in lineages with a comparably small number of preserved plastid-encoded ribosomal protein subunits. Thus, the rate of functional gene transfers from the plastid to the nucleus may be another important factor shaping the tempo and series of non-functionalization among plastid housekeeping elements from holoparasite plastomes. In the course of this work, experimental evidence was gathered showing that some holoparasites harbor apparently a great abundance of nuclear-encoded copies of plastid origin which may be indicative of increased rates of genes transfer under relaxed selective pressure.

1.2. *"… Have the profound alterations to the plastid genome seen in Epifagus and Conopholis all followed the loss of photosynthesis, or might some have begun before its loss? […]"*

This work showed that profound alterations to the plastid genome already accompany the transition to an obligate heterotrophic lifestyle in Orobanchaceae. Compared to closely related non-parasites, the hemiparasites *Schwalbea* and *Striga* exhibit notable changes in gene synteny including large-scale rearrangements as well as local gene losses. Even more, elements for both photosynthesis and housekeeping elements evolve at high rates with severely skewed ratios of non-synonymous to synonymous substitution rates departing from purifying selection in the entire plastid polymerase-complex. Relaxed purifying selection also prevails in some genes for photosystems and electron transport. It may thus be conceivable that the transition to obligate hemiparasitism initializes possibly irreversible alterations to the coding capacity of plastid chromosomes and afflicts distinct functional complexes. However, it remains unclear whether changes observed in hemiparasite plastomes eventually contribute to the loss of photosynthesis, or if the transition to holoparasitism occurs before non-functionalization of photosynthesis-related elements. Plastid genome structure of *Myzorrhiza* and the suspected coding capacity of other holoparasitic species [21,25,26,32,33] that evolved independently within Orobanchaceae imply that (at least) some lineages might have preserved the ability of photosynthesis for some time after the transition to a holo-heterotrophic lifestyle. Comparably little plastid genome reduction and/or retention of photosynthesis elements has also been described for lineages of *Cuscuta* [4,5] and a non-photosynthetic orchid [10].

1.3. "… Are the similarities between the plastid genomes of […] [Epifagus and Conopholis] due to convergence or to shared ancestry? […]"

Reductive evolution takes alternative paths after the loss of photosynthesis in the different clades of broomrape holoparasites. The similarities of *Epifagus* and *Conopholis* regarding their plastid genomes are in fact due to shared ancestry, as these genera constitute sister groups. With few exceptions only, the content of unique genes is very similar between the two species distinguishing them from other farther related Orobanchaceae species. In terms of coding capacity, *Epifagus* and *Conopholis* possess the most reduced plastid chromosomes among currently investigated Orobanchaceae. Due the loss of one entire IR segment, *Conopholis* retains the smallest plastid chromosomes among Orobanchaceae and across all land plants examined so far. Despite such large scale chromosomal modification between *Epifagus* and *Conopholis*, gene synteny as well as nucleotide substitution rates are most similar between the two parasites. Even more, *Epifagus* and *Conopholis* show a minor codon bias related to the nature of tRNA isoacceptors lost from their plastomes [34–36] that is not shared with other broomrapes. Nevertheless, convergent patterns of gene losses and other significant genetic changes do occur across different non-photosynthetic heterotrophs. Concerning the coding capacity, *Boulardia* and *Cistanche* exhibit similar degrees of plastid genome reduction, although both have lost notably fewer tRNA and ribosomal protein genes. Outside Orobanchaceae, the plastid chromosomes of non-photosynthetic heterotrophs converge to a coding capacity of the *Epifagus*/*Conopholis* clade [4,5,9].

1.4. "… How rapidly do the structural and other changes to the plastid genome come about in parasitic plants? […]"

The question regarding the tempo with which structural and other changes take place in heterotrophic plants is probably the most difficult one to answer. An absolute time estimate is currently impossible as fossilized Orobanchaceae are lacking impeding the reliable calibration of the trees. Moreover, the tempo of genetic and genomic changes appears to be highly lineage specific within Orobanchaceae, and may not simply be estimated from the absolute time since loss of photosynthesis, nor from the time since the transition to a heterotrophic lifestyle. Although comparative plastid genome analyses among closely related species provides invaluable insights into the patterns and series of gene loss and pseudogenization, it cannot answer questions related to the timing of single events. Given that obligate hemiparasites already exhibit remarkable changes in their plastomes, it can be speculated that – on an evolutionary time-scale – changes appear to occur very rapidly after the relaxation of selective pressures on the photosynthesis apparatus due to an obligate heterotrophic lifestyle. This study uncovered possible elements influencing the series of pseudogenization and gene losses under relaxed

evolutionary constraints (see section VI-1.1.). It revealed that reduced selective pressure relaxes the structural maintenance of the plastid chromosomal structure leading to numerous lineage specific changes of gene order, an increasing amount of plastid repetitive elements as well as a significant nucleotide compositional drift towards AT-richness. However, it also demonstrated that reductive evolution proceeds with remarkable differences among even very closely related species, although the underlying reasons remain elusive. In advance to answering question concerning the timing of genetic changes, causalities require clarification to at least some degree: Is gene loss and relaxation of structural maintenance of the plastid chromosome in hemiparasites cause or consequence of obligate hemiparasitism? Comparative analyses of a several more plastomes of close related Orobanchaceae heterotrophs might corroborate herein described evolutionary patterns that allow concluding the relative timing of events and causal relations. On top, it needs to be clarified to what extent non-photosynthetic pathways in plastids influence reductive plastome evolution between lineages of broomrape hemi- and holoparasites.

2. OUTLOOK

20 years after sequencing of the first plastid genome of the parasitic plant *Epifagus* [1,37], this work significantly increases our understanding of the series of structural and genetic changes of reductive genome evolution under relaxed selective constraints. The results of this study raise numerable new questions and hypotheses for future research projects concerning plastid genetics in both autotrophs and heterotrophs. One of those prospective projects concerns the issue whether the evolutionary patterns in reductive plastome evolution observed here are specific to the broomrape family, or if they confer to other heterotrophic plants as well. Specifically, the effect of operons and neighboring genes on the tempo of pseudogenization of dispensable genic fragments needs further attention in an extended set of taxa of Orobanchaceae, and other parasitic plant families. Frequently occurring genomic rearrangements in hemi- and holoparasites suggest relaxation of maintenance and repair of plastid DNA mediated by a number of nuclear-encoded elements [38–42]. Although the plastome is still required functionally for encoding housekeeping elements (at least as long as one gene functioning beyond the primary genetic apparatus needs to be translated), relaxation of selective pressure on photosynthesis genes due to heterotrophy may relax purifying selection on repair proteins as well. Higher mutation rates in those genes might lead to a greater abundance of improper and illegitimate recombination in plastid chromosomes of broomrape hemi- and holoparasites. Insertions and/or deletions may eventually afflict plastomic loci encoding photosynthesis-relevant elements. Eventual protein malfunction or loss of one subunit

probably reduces the functionality of an entire photosynthesis-related complex. Consequently, this scenarios would increase the dependency of the parasite on its host plant. Alternatively, early dark vegetative phases of root parasites may allow for the relaxation of elements such as PEP that guarantee high abundance of elements for light harvesting and electron transport. A thorough comparative genomic and transcriptomic analyses of photosynthesis and related elements between various hemiparasitic transition forms of root and shoot parasites may be a suitable approach to evaluate causalities of genetic changes under relaxed selective constraints. Another evidently open question concerns the survival of plastid *atp* genes in some holoparasite lineages: Do ATP-synthase genes function decoupled of photosynthesis? And if so, why are they preserved only in a subset of non-photosynthetic plants? Studies focusing on the evolution of lineage specific physiological pathways and the role of the plastid in parasitic plants may likely shed light on the different paths of reductive evolution among different lineages of heterotrophic plants within and outside the Orobanchaceae family. Are energetic and nutritional causes the reason for the strong nucleotide compositional bias in broomrape plastomes? Is the compositional bias related to structural changes including the deletion of plastomic segments, or rather a consequence thereof? What is the role of intracellular gene transfers in shaping the plastid chromosome structure of parasitic plants?

Due to the great diversity and wide range of trophic forms, the broomrape family has already qualified as an ideal group for studying evolutionary questions regarding plastome reduction and functional gene transfer. For the same reasons, representatives of Orobanchaceae will also be highly suitable for studying large-scale evolutionary changes of photosynthesis and photosynthesis-related elements of the nuclear genome, clarifying the role of the plastid in non-photosynthetic plants, and addressing questions regarding the impact of trophic and life form changes on the genome.

3. REFERENCES

1. Wolfe KH, Morden CW, Palmer JD (1992) Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc Natl Acad Sci USA 89: 10648–10652.

2. Knauf U, Hachtel W (2002) The genes encoding subunits of ATP synthase are conserved in the reduced plastid genome of the heterotrophic alga *Prototheca wickerhamii.* Mol Genet Genom 267: 492–497.

3. de Koning AP, Keeling PJ (2006) The complete plastid genome sequence of the parasitic green alga *Helicosporidium* sp. is highly reduced and structured. BMC Biology 4: 12.

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

5. McNeal JR, Kuehl J, Boore J, de Pamphilis C (2007) Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus *Cuscuta*. BMC Plant Biol 7: 57.

6. Wickett NJ, Zhang Y, Hansen SK, Roper JM, Kuehl JV, et al. (2008) Functional Gene losses occur with minimal size reduction in the plastid genome of the parasitic liverwort *Aneura mirabilis*. Mol Biol Evol 25: 393–401.

7. Nickrent D, García M (2009) On the brink of holoparasitism: Plastome evolution in Dwarf Mistletoes (*Arceuthobium*, Viscaceae). J Mol Evol 68: 603–615.

8. Maul K (2011) Assessing patterns of chloroplast genome reduction in hemi- and holoparasitic plants [Diploma thesis]. Universität Bonn.

9. Delannoy E, Fujii S, des Francs CC, Brundrett M, Small I (2011) Rampant gene loss in the underground orchid *Rhizanthella gardneri* highlights evolutionary constraints on plastid genomes. Mol Biol Evol 28: 2077–2086.

10. Logacheva MD, Schelkunov MI, Penin AA (2011) Sequencing and analysis of plastid genome in mycoheterotrophic orchid *Neottia nidus-avis*. Genome Biol Evol 3: 1296–1303.

11. Haberhausen G, Zetsche K (1994) Functional loss of *ndh* genes in an otherwise relatively unaltered plastid genome of the holoparasitic flowering plant *Cuscuta reflexa*. Plant Mol Biol 24: 217–222.

12. Chang CC, Lin HC, Lin IP, Chow TY, Chen HH, et al. (2006) The chloroplast genome of Phalaenopsis aphrodite (Orchidaceae): Comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. Mol Biol Evol 23: 279–291.

13. Wickett NJ, Fan Y, Lewis P, Goffinet B (2008) Distribution and evolution of pseudogenes, gene losses, and a gene rearrangement in the plastid genome of the nonphotosynthetic liverwort, *Aneura mirabilis* (Metzgeriales, Jungermanniopsida). J Mol Evol 67: 111–122.

14. Wu F-H, Chan M-T, Liao D-C, Hsu C-T, Lee Y-W, et al. (2010) Complete chloroplast genome of *Oncidium* Gower Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. BMC Plant Biol 10: 68.

15. Schäferhoff B (2011) Carnivory in Lamiales: Phylogeny, taxonomy, and chloroplast genome evolution [Dissertation]. Westfälische Wilhelms-Universität Münster.

16. Wakasugi T, Tsudzuki J, Ito S, Nakashima K, Tsudzuki T, et al. (1994) Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the Black Pine *Pinus thunbergii*. Proc Natl Acad Sci USA 91: 9794–9798.

17. Chris Blazier J, Guisinger-Bellian MM, Jansen RK (2011) Recent loss of plastid-encoded *ndh* genes within *Erodium* (Geraniaceae). Plant Mol Biol 76: 1–10..

18. Peltier G, Cournac L (2002) Chlororespiration. Ann Rev Plant Biol 53: 523–550.

19. Krause K, Berg S, Krupinska K (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..

20. Revill MJW, Stanley S, Hibberd JM (2005) Plastid genome structure and loss of photosynthetic ability in the parasitic genus *Cuscuta*. J Exp Bot 56: 2477–2486.

21. Lusson NA, Delavault PM, Thalouarn P (1998) The *rbcL* gene from the non-photosynthetic parasite *Lathraea clandestina* is not transcribed by a plastid-encoded RNA polymerase. Curr Genet 34: 212–215.

22. Berg S, Krause K, Krupinska K (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.

23. Legen J, Kemp S, Krause K, Profanter B, Herrmann RG, et al. (2002) Comparative analysis of plastid transcription profiles of entire plastid chromosomes from tobacco attributed to wild-type and PEP-deficient transcription machineries. Plant J 31: 171–188.

24. Hajdukiewicz PTJ, Allison LA, Maliga P (1997) The two RNA polymerases encoded by the nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids. EMBO J 16: 4041–4048.

25. Wolfe AD, dePamphilis CW (1997) Alternate paths of evolution for the photosynthetic gene *rbcL* in four nonphotosynthetic species of *Orobanche*. Plant Mol Biol 33: 965–977.

26. Randle CP, Wolfe AD (2005) The evolution and expression of RBCL in holoparasitic sister-genera *Harveya* and *Hyobanche* (Orobanchaceae). Am J Bot 92: 1575–1585.

27. Schwender J, Goffman F, Ohlrogge JB, Shachar-Hill Y (2004) Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds. Nature 432: 779–782.

28. Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, et al. (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA 104: 19369–19374.

29. Magee AM, Aspinall S, Rice DW, Cusack BP, Sémon M, et al. (2010) Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Res 20: 1700–1710.

30. Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D (2011) The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol 76: 273–297.

31. Fleischmann TT, Scharff LB, Alkatib S, Hasdorf S, Schöttler MA, et al. (2011) Nonessential plastid-encoded ribosomal proteins in tobacco: A developmental role for plastid translation and implications for reductive genome evolution. Plant Cell. Doi: 10.1105/tpc.111.088906.

32. Delavault PM, Sakanyan V, Thalouarn P (1995) Divergent evolution of two plastid genes, *rbcL* and *atpB*, in a non-photosynthetic parasitic plant. Plant Mol Biol 29: 1071–1079.

33. Delavault PM, Russo NM, Lusson NA, Thalouarn P (1996) Organization of the reduced plastid genome of *Lathraea clandestina*, an achlorophyllous parasitic plant. Physiol Plant 96: 674–682.

34. Lohan AJ, Wolfe KH (1998) A subset of conserved tRNA genes in plastid DNA of nongreen plants. Genetics 150: 425–433.

35. Morden CW, Wolfe KH, dePamphilis CW, Palmer JD (1991) Plastid translation and transcription genes in a nonphotosynthetic plant - Intact, missing and pseudogenes. EMBO J 10: 3281–3288.

36. Wolfe KH, Morden CW, Ems SC, Palmer JD (1992) Rapid evolution of the plastid translational apparatus in a nonphotosynthetic plant: loss or accelerated sequence evolution of tRNA and ribosomal protein genes. J Mol Evol 35: 304–317.

37. dePamphilis CW, Palmer JD (1990) Loss of photosynthetic and chlororespiratory genes from the plastid genome of a parasitic flowering plant. Nature 348: 337–339.

38. Maréchal A, Parent J-S, Véronneau-Lafortune F, Joyeux A, Lang BF, et al. (2009) Whirly proteins maintain plastid genome stability in Arabidopsis. Proc Natl Acad Sci USA 106: 14693–14698.

39. Rowan BA, Oldenburg DJ, Bendich AJ (2010) *RecA* maintains the integrity of chloroplast DNA molecules in Arabidopsis. J Exp Bot 61: 2575–2588.

40. Maréchal A, Brisson N (2010) Recombination and the maintenance of plant organelle genome stability. New Phytol 186: 299–317.

41. Xu Y-Z, Arrieta-Montiel MP, Virdi KS, de Paula WBM, Widhalm JR, et al. (2011) *MutS* HOMOLOG1 is a Nucleoid protein that alters mitochondrial and plastid properties and plant response to high light. Plant Cell.

42. Hofmann NR (2011) *MutS* HOMOLOG1 stabilizes plastid and mitochondrial genomes. Plant Cell.

CURRICULUM VITAE

Personal Data

Name:	Susann Wicke
Date of Birth	24. 07.1982 in Zwickau, Germany
Nationality	German
Address	Heisstr. 18, D-48145 Muenster, Germany
E-mail:	susann.wicke@uni-muenster.de

University Education

Since 5/2007	continued PhD position at the Department of Biogeography at the University of Vienna, Austria: - Dissertation research project: "*Plastid Genome Evolution in a group of non-photosynthetic parasitic angiosperms (Orobanchaceae)*" funded by the Austrian Science Fund (FWF), PhD-advisor: Prof. Dr. Gerald Schneeweiss
10/2002 – 5/2007	Studies of biology (Diplom-Biologie) at the Technische Universität Dresden, Germany with major focus on botany, genetics and microbiology - Diploma thesis: "*Genomic reorganization of the nuclear ribosomal DNA – Structural analyses of the nrDNA´s intergenic spacer region among major land plant lineages*", Advisor: Prof. Dr. Christoph Neinhuis
10/2001 – 09/2002	Undergraduate studies on geo-ecology at the Technische Universität Bergakademie Freiberg/Germany

School Education

1993 – 2001	Grammar school "Clara-Wieck-Gymnasium" in Zwickau, Germany, Abitur
1989 – 1993	Primary school „Friedrich-Engels-Grundschule in Zwickau, Germany

Studies & Research Abroad

Since 06/2010	Research associate in the Plant Evolution and Biodiversity group headed by Prof. Dr. Kai Müller at the Institute for Evolution and Biodiversity, University of Münster, Germany.
07/2009 – 10/2009	Research stay at the Pennsylvania State University (PSU), State College USA: *"Plastid genome evolution in a group of non-parasitic angiosperms (Orobanchaceae)"* under the supervision of Prof. Dr. Claude W. dePamphilis (PSU).
07/09 2008	Guest researcher at the Royal Botanical Garden (RJB) Madrid (CSIC), Madrid, Spain: - Project: *"Crossing the Atlantic by means of wind: a case study on Anacolia laevisphaera"* in collaboration with Dr. Jesus Munoz (RJB Madrid) and Prof. Dr. Dietmar Quandt (University of Bonn, Germany).
09/11 2007	Guest researcher at the Natural History Museum (NHM), London, Great Britain: - Project: *"Reorganization of the chloroplast genome in leptosporangiate ferns"* in collaboration with Dr. Harald Schneider (NHM London) and Prof. Dr. Dietmar Quandt (University of Bonn, Germany).
07/09 2005	Undergraduate research stay at the Royal Botanical Garden (RJB) Madrid (CSIC), Madrid, Spain: - Project: *"Molecular evolution and phylogenetic/-genomic utility of the nuclear ribosomal DNA in bryophytes"* , supervised by Dr. Jesùs Muñoz (RJB Madrid), and Prof. Dr. Dietmar Quandt (University of Bonn and TU Dresden, Germany).

Research Grants & Awards

2011	Travel grant from the German Academic Exchange Service (DAAD) for participation and oral presentation at the XVIII. International Botanical Congress in Melbourne, Australia, covering travel and accommodation expenses (EUR 2,000).
2010	SYNTHESYS (EU) grant (NL-TAF-1550) for travel support to the University of Leiden, The Netherlands covering travel and living expenses (EUR 1,000).
2009	Short-term research grant (KWA-grant) by the University of Vienna covering living-expenses at the Pennsylvania State University, USA (EUR 1.300).

continued 2009	Genetics Section Student Travel Award ($ 500) by the Botanical Society of America for the oral presentation: "*From the exception to the rule: the re-arrangement(s) of the nuclear ribosomal DNA in land plants*".
2008	SYNTHESYS (EU) grant (ES-TAF-5266) for short-term research at the Royal Botanical Garden Madrid (CSIC), Madrid, Spain. The grant covered travel and research costs plus living expenses (EUR ≈10.000).
2007	SYNTHESYS (EU) grant (GB-TAF-3745) for short-term research at the Natural History Museum London, Great Britain. The grant covered travel and research costs plus living expenses (EUR ≈12.000).

Research Interests & International Collaborations

Evolution of Parasitic Plants and Organellar Genome Evolution	- Plastid Genome Evolution in Parasitic Plants - Evolution and Phylogeny of Orobanchaceae - Evolution of Photosynthesis in Obligate Parasites of Orobanchaceae, Loranthaceae, and Viscaceae) - Evolution of Plastid DNA Repair Proteins - Plastid-to-Nucleus Gene Transfer in Parasitic Plants - Evolution of Carnivorous Plants (focus in Lentibulariaceae) - Optimization of Locus-Specific Assembly from High-Throughput Sequencing Data
Nuclear Ribosomal DNA and Land Plant Evolution	- Molecular Evolution of 5S ribosomal DNA/RNA in land plants - Mutational Dynamics and Concerted Evolution of Nuclear Ribosomal (nrDNA) - Mutational Dynamics and Molecular Evolution of plastid DNA in Early Vascular

Review Activities

Research Funding Agency	Czech Science Foundation
Peer reviewed journals (selected)	Plant Molecular Biology BMC Plant Biology Plant Ecology and Evolution

Publications in peer-reviewed journals

2011

Wicke S, Costa A, Muñoz J, and Quandt D. Restless 5S: The re-arrangement(s) and evolution of the nuclear ribosomal DNA in land plants. *Molecular Phylogeny and Evolution* 61(2): 321-332.

Wicke S, Schneeweiss GM, Müller KM, dePamphilis CW, and Quandt D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Molecular Biology* 76(3-5): 273-97.

2010

Preußing M, Olsson S, Schäfer-Verwimp A, Wickett NJ, **Wicke S**, Quandt D, and Nebel M. New insights in the evolution of the liverwort family Aneuraceae (Metzgeriales, Marchantiophyta) with special regards to the genus *Lobatiriccardia*. *Taxon* 59(5): 1424-1440.

2009

Wicke S and Quandt D. Universal primers for the amplification of the plastid *trnK/matK* region in land plants. *Anales del Jardín Botánico de Madrid* 66(2): 285-288.

Oral contributions to international congresses (* presenter)

2011

Wicke S*, Quandt D, Müller KF, dePamphilis CW and Schneeweiss GM. Organellar genome evolution in parasitic plants - Assessing patterns of genome reduction and rate acceleration under relaxed selective pressure. *IBC2011 – XVIII. International Botanical Congress*, Melbourne, Australia.

Wicke S*, Latvis M, Quandt D, Schneeweiss GM, Soltis PS, dePamphilis CW, Soltis DE, and Müller KF. Minimum requirements for *de novo* organelle genome reconstruction using whole genome shotgun sequencing. *BioSystematics* 2011, Berlin, Germany.

2010

Wicke S*, Quandt D, Müller KF, dePamphilis CW and Schneeweiss GM. Assessing the patterns of plastid genome reduction, pseudogenization and gene loss in (non-) photosynthetic parasitic flowering plants. *19th International Symposium "Biodiversity and Evolutionary Biology"*, Vienna, Austria. Invited presentation.

Damayanti L*, **Wicke S**, Symmank L, Muñoz J, Frahm JP, Shaw B and Quandt D. Gone with the wind: The phylo-geography of *Strictidium*. *19th International Symposium "Biodiversity and Evolutionary Biology"*, Vienna, Austria. Invited presentation.

2010 - continued	**Wicke S***, Quandt D, Müller KF, Wickett NJ, dePamphilis CW, and Schneeweiss GM. Plastid Genome Evolution - What's so different between autotrophs, semi- and non-autotrophic flowering plants. *Botany 2010*, Providence (Rhode Island), USA.
	Wicke S*, Quandt D, Müller KF, dePamphilis CW, and Schneeweiss GM. Plastid Genome Evolution in (non-) photosynthetic flowering plants. *2nd International Symposium on Chloroplast Genomics and Genetic Engineering (ISCGGE)*. Maynooth, Co. Kildare, Ireland. Invited presentation.
2009	**Wicke S*****,** Quandt D and Schneeweiss GM. Plastid genome evolution in a group of non-photosynthetic angiosperms (Orobanchaceae). *Botany & Mycology 2009*, Snowbird (Utah), USA. Invited symposium presentation.
	Wicke S*, Costa A, Muñoz J, Neinhuis C, and Quandt D. From the exception to the rule: The re-arrangement(s) of the nuclear ribosomal DNA in land plants. *Botany & Mycology 2009*, Snowbird (Utah), USA. Awarded presentation.
	Worberg A, Quandt D, Korotkova N, Müller K, **Wicke S*****,** and Borsch T. `Very large´, `poorly understood´ but `well supported´ – The Phylogeny of Rosids based on fast evolving and non-coding chloroplast markers. *Botany & Mycology 2009*, Snowbird (Utah), USA.
2008	**Wicke S***, Costa A, Bauer F, Muñoz J, Neinhuis C, and Quandt D. Organization(s) of the nuclear ribosomal DNA in land plants. *Systematics 2008* - first joint meeting of GfBS and DBG, Goettingen, Germany. Invited presentation.
2007	**Wicke S***, Costa A, Bauer F, Muñoz J, Neinhuis C and Quandt D. Organization(s) of the nuclear ribosomal DNA in land plants. *International Botanical Congress* (DBG), Hamburg, Germany.
2006	Quandt D*, Borsch T, **Wicke S**, Renner SS, Hilu KW, and Neinhuis C. Seed plant evolution: sequence based cladistics vs. micro-structural changes. *17th International Symposium on "Biodiversity and Evolutionary Biology"*, Bonn, Germany.
	Quandt D*, Borsch T, **Wicke S**, Won H, Renner SS, and Hilu KW. Land plant evolution: a perspective from fast evolving chloroplast regions. *Botany 2006*, BSA international conference Chico, USA. Invited presentation.
	Quandt D*, Prieskorn S, Petzold K, **Wicke S,** and Neinhuis C. Do the old European camellias have a common origin? A molecular approach. *17th Exhibition of Ancient Camellias*, Lucca, Italy.

Poster contributions to international congresses (* presenter)

2012	**Wicke S**, Quandt D, Müller KF, dePamphilis CW, and Schneeweiss GM*. Assessing patterns of plastid genome reduction under relaxed selective pressure in a group of non-photosynthetic angiosperms. Plant and Animal Genome XX Conference, San Diego, CA, USA.
2010	Ataei N*, **Wicke S**, Schneeweiss H, Schneeweiss GM, and Quandt D. Phylogeography and genome evolution of the non-photosynthetic parasitic plant *Cistanche* (Orobanchaceae). *19th International Symposium "Biodiversity and Evolutionary Biology"*, Vienna, Austria.
2008	**Wicke S***, Costa A, Bauer F, Muñoz J, Neinhuis C, and Quandt D. From the exception to the rule: the co-localization of the nuclear ribosomal DNA (nrDNA) in land plants represents the ancestral state. *SMBE 2008 – Annual Meeting of the Society for Molecular Biology and Evolution*, Barcelona, Spain.
	Wicke S*, Quandt D, and Schneeweiss GM. Plastid genome evolution in a group of non-photosynthetic angiosperms. *SMBE 2008 – Annual Meeting of the Society for Molecular Biology and Evolution*, Barcelona, Spain.
	Wicke S*, Schneider H, and Quandt D. Structure and evolution of the *trnL*UAA-intron in Monilophytes. *SMBE 2008 – Annual Meeting of the Society for Molecular Biology and Evolution*, Barcelona, Spain.
	Schneeweiss GM*, Schiller D, and **Wicke S**. Heteroplasmy and horizontal gene transfer in a non-photosynthetic flowering plant: Where do the two distinct plastid large single copy regions come from? *SMBE 2008 – Annual Meeting of the Society for Molecular Biology and Evolution*, Barcelona, Spain.

Vienna, January 5th 2012

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe.

Wien, Januar 2012 x_____