Summer 7-26-2016

Comparative Evolutionary Analysis of Organellar Genomic Diversity in Green Plants

Weishu Fan
University of Nebraska - Lincoln

COMPARATIVE EVOLUTIONARY ANALYSIS OF ORGANELLAR GENOMIC DIVERSITY IN GREEN PLANTS

by

Weishu Fan

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Agronomy and Horticulture

Under the Supervision of Professor Jeffrey P. Mower

Lincoln, Nebraska

July, 2016

COMPARATIVE EVOLUTIONARY ANALYSIS OF ORGANELLAR GENOMIC DIVERSITY IN GREEN PLANTS

Weishu Fan, Ph.D.

University of Nebraska, 2016

Advisor: Jeffrey P. Mower

The mitochondrial genome (mitogenome) and plastid genome (plastome) of plants vary immensely in genome size and gene content. They have also developed several eccentric features, such as the preference for horizontal gene transfer of mitochondrial genes, the reduction of the plastome in non-photosynthetic plants, and variable amounts of RNA editing affecting both genomes. Different organismal lifestyles can partially account for the highly diverse organellar genomes across the tree of green plants. For example, endosymbiotic and parasitic lifestyles can dramatically affect the genomic architectures of plant mitochondria and plastids. In this study, the organellar genomes of several green plants with atypical lifestyles were investigated and compared with the breadth of organelle genomic diversity within green plants. Next-generation sequencing and comparative evolutionary analyses were performed on organellar genomes of parasitic plants in Orobanchaceae and endosymbiotic algae in Chlorellaceae. Comparative organellar genomic analysis from endosymbiotic green algae provided no evidence for genome reduction; instead the endosymbiont genomes are generally larger in genome size and richer in intron content. Similarly, facultative hemiparasitic species in Orobanchaceae revealed minimal organellar genome degradation, but some evidence for several horizontal transferred genes. In both groups, the lack of genomic reduction may be attributed to the retention of photosynthetic ability. In addition, the extent of RNA

editing was examined in the mitogenome of *Welwitschia*, a xerophytic plant. RNA editing sites in *Welwitschia* are extremely reduced compared with other gymnosperms, and may be caused by retroprocessing. Taken together, these results demonstrated that atypical lifestyle does not necessarily lead to the production of unusual genomic features and exhibited the convergence and divergence in green plants organelle genomes.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my special appreciation and thanks to my advisor Dr. Jeffrey P. Mower. He has been a tremendous mentor for me. I appreciate all his contributions of ideas and time to make my Ph.D. program stimulating and effective. His enthusiasm and continuous innovation on the research was contagious and motivational for me, especially during tough times in the journey. He has provided an excellent example for being a scientist and professor. My gratitude to him is magnified by his warm personality and constant encouragement and support.

Secondly, I would like to recognize my committee members: Dr. Heriberto Cerutti, Dr. Thomas E. Elthon and Dr. Wayne Riekhof for serving as my committee members, and for their time, interest, brilliant comments and insightful suggestions. I am thankful to Dr. Sally Mackenzie's lab, Dr. Kenneth Nickerson's lab, Dr. Wayne Riekhof's lab and Dr. Steve Harris' lab for their generous sharing of facilities and chemicals with us. I also sincerely thank my previous advisor Dr. Gaihe Yang for his support and encouragement.

My colleagues in the Mower Lab have contributed immensely to my professional and personal time at Lincoln. I am especially grateful to Dr. Wenhu Guo for his constant help, collaborations and constructive advice in my research and career. I also appreciate the comments and suggestions from Dr. Andan Zhu. I would like to acknowledge former and present lab members: Nancy Hepburn, Dr. Kanika Jain, Dr. Brandi Sigmon, Dr. Felix Grewe, Lexis Funk and Yan Li for their invaluable input to my dissertation and friendship.

I also acknowledge the funding source that made my Ph.D. work in U.S. possible. I was funded by a fellowship from the China Scholarship Council for four successive years. My work was also supported by the National Science Foundation and Center for Plant Science Innovation.

My time at University of Nebraska-Lincoln was made enjoyable in large part due to the many friends that became a part of my life. I cannot thank enough Jyothi Kumar, Sunil Kumar, Ruvini Pathirana, Jithesh, Amit, Yuan, Wenjin, Ting, Xin, Wenjia, Fei, Guangchao, Qian and many other friends for being so nice and bringing memorable times.

Lastly, I would like to thank my families, especially my mom Xiuzhi Li, my sister Chao Fan and my best friend Peng Wang, for all their unconditional love and encouragement, for their faithful support and patience. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Literature Review

1 Introduction

The endosymbiotic theory states that the mitochondrion, a key factor in energy generation, originated ~1.5 billion years ago from the α-proteobacteria endosymbiont (Gray et al. 1999), while the plastid, the major organelle associating with photosynthesis and carbon fixation, originated from cyanobacteria (Douglas and Turner 1991, Wolf 2012). The mitochondrial genome (mitogenome) and the plastid genome (plastome) are the DNA genomes are separate from the nuclear genome, and their main function is to generate some of the proteins required for mitochondrial and plastid activity, respectively. Genome sequencing technologies have revolutionized genomics and ushered in comparative analysis for functional and evolutionary purpose.

The mitogenome and plastome display broad variation in their structure, size, gene content and posttranscriptional modification. This is true in many eukaryotic lineages, and particularly in green plants (Viridiplantae), for which many organellar genomes are now available. The green plants include what have traditionally been called “green algae” and the land plants (Embryophytes), forming a monophyletic group of eukaryotic organisms with distinctive chloroplasts (Lang and Nedelcu 2012, Simpson 2010). Although most plants have retained the plastid genome, recent studies have suggested that the plastome has become lost in the holoparasite *Rafflesia* (Molina et al. 2014) and also in the non-photosynthetic unicellular alga *Polytomella* (Smith and Lee 2014). Similarly, the plant mitogenome exhibits an extreme range of sizes, from 13 kb to >11 Mb, along with several unique features, such as the insertion of foreign DNA and widespread RNA editing.

One of the reasons that contribute to the highly diverse organellar genomes is the different living styles. While most green plants are autotrophic and free-living, some plants have to live in symbiotic or parasitic relationships with the other organisms, obtaining partial or complete nutrients from their hosts. Another small portion of plants called xerophytes have adapted to survive in arid environments. All of these atypical plant lifestyles have the potential to contribute to the extensive differences in the organellar genomes.

2 Green Plants Organelle Genome Sequencing Progress

The first mitogenome fragments were sequenced from *Chlamydomonas reinhardtii* in 1985 (Boer et al. 1985), a pioneer point of genomic studies in green plants. The first complete genomes to be sequenced were the plastid genomes of *Marchantia polymorpha* and *Nicotiana tabacum* 30 years ago (in 1986) (Ohyama et al. 1986, Shinozaki et al. 1986). Six years later, the first complete land plant mitochondrial genome was reported from *M. polymorpha* (Oda et al. 1992). After another six years, the first complete mitogenome from green algae was reported from *C. reinhardtii* (Denovan-Wright et al. 1998). In 2000, the first plant nuclear genome was published from *Arabidopsis thaliana*, opening the floodgates for the plant evolutionary and biological research (Mayer et al. 1999, Salanoubat et al. 2000, Tabata et al. 2000, Theologis et al. 2000).

Over the next decade, aided by the fast development of next-generation sequencing technologies, complete genome sequences from plant organelles have grown enormously, driving the comparative genomic studies.

Currently, more than 1100 complete plastomes from the green plants are available (https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?opt=plastid&taxid=33090, last accessed on June 2016). Within the green algae, there are 18 complete plastomes from charophytes (the green algae most closely related to land plants) and more than 75 from chlorophytes (the group containing the majority of green algae). In land plants, the total released complete plastomes exceed 1000. In contrast, the total number of sequenced mitogenome is only approximately 10% as that of plastome (https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?opt=organelle&taxid=33090, last accessed June 2016), which could be attributed to the higher genomic complexity of many land plant mitochondrial genomes that increases the difficulty in assembly. The mitogenome dataset from green algae includes nine charophytes and 44 chlorophytes. Among land plants, representatives from all major groups have been sequenced or reported (>100 mitogenomes), demonstrating dramatic diversity in many aspects (details below). Within vascular plants, there are more than 70 mitogenomes from angiosperms, while many fewer are from the other groups, including only three from gymnosperms [*Cycas taitungensis* (Chaw et al. 2008), *Ginkgo biloba* and *Welwitschia mirabilis* (Guo et al. 2016)], three from lycophytes [*Isoetes engelmannii* (Grewe et al. 2009), *Selaginella moellendorffii* (Hecht et al. 2011) and *Huperzia squarrosa* (Liu et al. 2012)], and two from ferns [*Ophioglossum californicum* and *Psilotum nudum* (Guo, in press)].

Among all of the sequenced organellar genomes, it is worthwhile to mention that there have been efforts to sequence plants with a broad spectrum of lifestyles: free-living and endosymbiotic species, parasitic and non-parasitic plants, photosynthetic and non-photosynthetic taxa, xerophyte and mesophyte species. With these genomes elaborated, it

has become possible to compare evolutionary diversity among species to assess the genomic effects of these different lifestyles on characteristics such as the conserved and diverse genome size and gene content, shared or distinguished features.

3 Plant Organellar Genome Size and Genetic Content Diversity

3.1 Plant Mitochondrial Genome

The plant mitochondrion plays a major role in translocating protons and generating ATP (Rose and Sheahan 2007, Tovar et al. 1999). These functions are required by all plants, and therefore there are no examples of mitogenome loss from any species. However, there dramatic variation in genome size, gene arrangement, the amount of foreign DNA and the degree of posttranscriptional modification.

3.1.1 Genome Size

Genome size varies dramatically among major plant lineages. In green algae, *Polytomella capuana* was reported to have the smallest mitogenome at only 13 kb in size (Smith and Lee 2008). At the opposite extreme, due to the lower coding density and heavier repeat elements, the mitogenome of *Chlorokybus atmophyticus* has expanded to 201.8 kb (Turmel et al. 2007). In land plants, mitogenome size diversity is even greater, ranging from 58 kb in *I. engelmannii* to over 11 Mb in *Silene conica* (Grewe et al. 2009, Sloan et al. 2012). Strikingly, the massive mitogenome size in *S. conica* is even larger than some cyanobacterium genomes, which comprise thousands of genes (Welsh et al. 2008). In the gymnosperm subclade, although only three complete mitogenomes have been sequenced so far, the genome size varies from 346 to 979 kb (Chaw et al. 2008, Guo et al. 2016). In

angiosperms, in contrast to the extremely large mitogenomes in *S. conica* (>11 Mb) and *Cucurbita pepo* (983 kb), the hemiparasitic plant *Viscum scurruloideum* mitogenome is only 66 kb (Alverson et al. 2010, Skippington et al. 2015, Sloan et al. 2012). All of these previous studies have demonstrated the wide range of mitogenome size, but given the few genomes that have been sequenced so far, it is very possible that even more diversity remains to be discovered.

3.1.2 Gene Content

The mitochondrial genome encodes genes associated with electron transport (*nad*, *sdh*, *cob*, *cox*), ATP synthesis (*atp*), translation (ribosomal RNAs, transfer RNAs, *rpl* and *rps*), protein import and maturation (*ccm*, *mttB*/*tatC*) (Table 1-1). Most of the green algae retain most of these genes which are essential for respiration, although some species exhibit minor to substantial gene loss. For example, selective species in prasinophytes and chlorophyceae experienced massive gene loss, particularly the *atp*, *rps*, and *rpl* genes (Kroymann and Zetsche 1998, Smith et al. 2010, Turmel et al. 1999). The colorless green alga *Polytomella parva* encodes only seven genes (Fan and Lee 2002).

Land plants typically contain 20-41 protein-coding genes (Table 1-1). The angiosperm mitogenomes generally share the same set of 24 core protein genes but differ in the other 17 protein genes, most of which are ribosomal proteins (Adams and Palmer 2003, Adams et al. 2002). However, the recently reported mitogenome from the parasitic plant *Viscum* has lost 11 of the 24 core protein genes, and 11 of the 17 variably present genes, resulting in the lowest mitochondrial gene content in angiosperms (Petersen et al. 2015, Skippington et al. 2015). In hornworts and lycophytes, the absence of *ccm* genes and

ribosomal protein genes have been demonstrated many times (Grewe et al. 2009, Hecht et al. 2011, Xue et al. 2010), contributing to even more gene content diversity across land plants.

In terms of ribosomal RNAs (rRNA), all green plants possess large and small ribosomal RNA (*rnl* and *rns*), and almost all land plants have 5S rRNA (*rrn5*) except for *S. moellendorffii* (Hecht et al. 2011). The *rrn5* is also present in some of the green algae, although it was not found in Chlorophyceae and several lineages of prasinophytes (Burger and Nedelcu 2012, Mower et al. 2012).

3.1.3 Intron Content

In mitochondrial genes of green plants, introns are prevalent but their position and frequency are highly variable, from a single intron in some green algae (*Chlorella* sp. ArM0029B, *Pedinomonas minor*, etc) to 37 introns in the lycophyte *S. moellendorffii* (Jeong 2014, Turmel 1999, Hecht 2011). Although intron content is relatively conserved within the major land plant lineages, it can sometimes still show obvious variation. For example, the gymnosperms *Cycas* and *Ginkgo* contain 26 and 25 mitochondrial introns, respectively, while the *Welwitschia* mitogenome retains only 10 introns, less than half of the other two gymnosperms (Chaw et al. 2008, Guo et al. 2016). In vascular plants, the variation in intron content is due primarily to the loss of introns in particular lineages. Few examples of obvious intron gain exist. However, a well-studied example of intron gain is the mobile *cox1* intron in angiosperm, which is speculated to be horizontally transferred among lineages due to its mobile nature (Cho et al. 1998, Sanchez-Puerta et al. 2008).

3.2 Plant Plastid Genome

The plant plastid performs an essential role in carrying out photosynthesis and carbon fixation in plant metabolism. Many plastid-encoded genes, therefore, are involved in the photometabolic pathways (Bock Ralph 2007, Palmer 1997) while others generate components of the genetic apparatus, such as structural and transfer RNAs (tRNA), in order to translate the plastid genes.

3.2.1 Genome Structure and Genome Size Diversity

Generally, the plastid chromosomes are arranged in head-to-tail concatemers of multiple molecules in circularized or linear form (Scharff and Koop 2006, Wicke et al. 2011). In land plants, the plastid genome normally has a highly conserved structure with two inverted repeats (IR) that separate the genome into a large and small single-copy region (LSC and SSC region, respectively). In green algal lineages, most plastid DNAs are circular mapping with the occurrence of some fragments carrying repeat regions (Reyes-Prieto et al. 2007). Some of the green algae have the IR structure (e.g. *P. minor*) but the others like *Chlorella vulgaris* do not (Wakasugi et al. 1997). The mechanisms of generating or eliminating the repeat regions remain unclear.

The plastomes in green plants also have diversity in their genome size, which is closely related to the retention or loss of photosynthetic ability. The photosynthetic green algae exhibit the total length of 118-521 kb in *Mesostigma viride* and *Floydiella terrestris*, respectively (Brouard et al. 2010, Lemieux et al. 2000). For those nonphotosynthetic or endosymbiotic green algae, the genome size may be much smaller, such as the parasitic, nonphotosynthetic green alga *Helicosporidium* (de Koning and Keeling 2006). In

Polytomella spp., a free-living, nonphotosynthetic green alga related to *Chlamydomonas* (Smith and Lee 2014), the entire plastome was inferred to be lost. The plastid genome size of most land plants is between 120-170 kb and the genome size is strongly affected by the length of the IR regions (Chumley et al. 2006, Naumann et al. 2016, Ruhlman and Jansen 2014, Wakasugi et al. 1994, Wu et al. 2007). Again, the genome reduction or complete loss could occur in parasitic lineages, which will be discussed in more detail later.

3.2.2 Gene Content

The plastid genes in green plants have been classified in three categories, encoding components of the genetic system, components related to photosynthesis, and components of other pathways. A detailed catalogue of the protein-coding genes is displayed in Table 1-2. In total, the number of protein-coding genes in photosynthetic species is approximately 66-88 (Wicke et al. 2011). In contrast, the species that have lost photosynthetic capacity generally eliminate the genes related to photosynthesis (see below), such as the holoparasites (Funk et al. 2007, Molina et al. 2014). The ribosomal RNA genes (*rrn23*, *rrn16*, *rrn5* and *rrn4.5*) are present in most land plant plastomes studied so far, and often these genes are duplicated in each copy of the IR region. In green algae, generally there is only one set of ribosomal RNA with no *rrn4.5*. Similar to the protein-coding genes, nonphotosynthetic or minimal photosynthetic angiosperms will also lose some tRNAs (Funk et al. 2007, McNeal et al. 2007, McNeal et al. 2009, Morden et al. 1991). For transfer RNA genes, the standard number is 27-32, but unique cases could also occur in some photosynthetic specus, such as *Selaginella uncinata* which has only 12 tRNA genes (Tsuji et al. 2007).

3.2.3 Intron Content

The intron content in the plastome is extremely variable among all the different clades, and even within the same group it can vary from lineage to lineage. Land plants typically contain 15-20 group II introns and one group I intron in *trnL*. The plastid *matK* gene is usually located within the group II intron of *trnK*. In green algae, intron content can range from no introns in some species to 26 introns in *F. terrestris*. Although several highly sophisticated search algorithms have been developed and applied, it is still difficult to identify and classify all the introns, especially in tRNAs (Beck and Lang 2009, 2010, Wyman et al. 2004).

4 Peculiar Features of Plant Organellar Genomes

The plant mitochondrial genome exhibits some unusual features, such as frequent genomic recombination (Lonsdale et al. 1984, Marechal and Brisson 2010), incorporation of foreign DNA (Adams et al. 2002, Cho et al. 1998), widespread RNA editing (Giegé and Brennicke 1999, Hiesel et al. 1989) and *trans*-splicing of transcripts to remove introns (Chapdelaine and Bonen 1991, Wissinger et al. 1991). However, many of these peculiar features are absent from green algae and early branches of land plants. The plant plastid genome is generally very highly conserved in land plants, with some more diverse structures in green algae. However, some abnormal changes can occur to the plastome when parasitism evolves.

4.1 Horizontal Gene Transfer in Plant Mitochondrial Genome

Horizontal gene transfer (HGT) refers to the transfer of the genetic material between non-mating species (Bock 2010). Plant mitogenomes experience frequent horizontal transfer of genes acquired from evolutionarily diverse plants (Davis et al. 2005, Richardson and Palmer 2007). The exchange of genetic material is more often to occur by the transfer of DNA rather than RNA (Mower et al. 2010, Xi et al. 2013, 2012, Zhang et al. 2014). To date, most of the transferred genes thus far have resulted in the presence of both horizontally acquired and preexisting gene copies within mitogenome, such as observed for *Rafflesia* and *Amborella* (Rice et al. 2013, Xi et al. 2013). In a few cases, some of the transferred genes have functionally replaced the native copies (Sanchez-Puerta 2014).

The abundance of HGT in plant mitochondria has been suggested to be caused by several factors, including the ability of the mitochondrion to uptake DNA (Koulintchenko et al. 2003), frequent fusion and fission of mitochondrial during the plant cell cycle (Arimura et al. 2004) and the presence of massive intergenic regions in the mitogenome (Kitazaki and Kubo 2010). Furthermore, HGT occurs frequently in parasitic-host plant systems (Bergthorsson et al. 2004, Zhang et al. 2014). Evidence of HGT has been reported in 10 of the 11 parasitic lineages to date (Davis and Xi 2015). This preference is hypothesized mainly because of the haustorium, a physical connection between hosts' roots or shoots and the parasites. Studies have uncovered the fact of horizontal movement of genetic material from hosts to parasites and in the opposite direction (Davis and Wurdack 2004, Mower et al. 2004). In some cases, the frequency of HGT can be on a large scale. For example, Rafflesiaceae harbored 41% of its mitochondrial genes via HGT (Xi et al. 2013). Horizontal transfer can also involve intron transfer, such as the relatively well-

studied case of HGT comes from the *cox1* intron, being acquired in many angiosperm lineages by mitochondria-to-mitochondria movement. This also illustrated that the horizontal transfer in mitochondria is not limited to protein-coding genes (Alverson et al. 2010, Barkman et al. 2007, Cho et al. 1998, Sanchez-Puerta et al. 2008). Briefly, studies have implied that plant mitogenomes experience rampant HGT, and parasitic plants are particularly active in HGT.

4.2 RNA Editing in Plant Mitochondrial Genome

RNA editing is a widespread post-transcriptional process that changes the coding information of mRNAs. In angiosperms, RNA editing was first recognized in the mitogenome by comparing DNA and RNA sequences in 1989 (Hiesel et al. 1989, Takenaka et al. 2008). The most common RNA editing effect is the conversion from cytidines to uridines (C-to-U), which has been reported in all major land plant lineages but is absent from green algae (Chaw et al. 2008, Guo et al. 2016, Oda et al. 1992, Unseld et al. 1997). Another type of editing involves the U-to-C reverse RNA editing, which is distributed in some specific lineages in ferns, mosses, and lycophytes (Grewe et al. 2009, Kugita et al. 2003). RNA editing can create initiation and termination codons or remove premature stop codons, but most often it restores internal codons with strong functional relevance in protein-coding genes. RNA editing can also occur in non-coding regions of the mitogenome, which in some cases improve intron splicing efficiency and tRNA processing (Castandet et al. 2010, Grewe et al. 2011).

4.2.1 Variation in the Frequency of RNA Editing

The frequency of mitogenome RNA editing is significantly different among all the major lineages within land plants. All examined land plants examined to date perform C-to-U RNA editing except the complex thalloid liverworts (Marchantiidae) (Steinhauser 1999), whereas no editing has been reported yet in any green algae. Thus, we can speculate that C-to-U RNA editing originated from the common ancestor of land plants and was subsequently lost from the complex thalloid liverworts.

Angiosperm mitogenomes are diverse in RNA editing based on extensive studies across the clade. To date, 12 mitogenomes have been shown to have various editing numbers: *Arabidopsis* (441 sites) (Giege and Brennicke 1999), *Oryza* (491 sites) (Notsu et al. 2002), *Brassica* (427 sites) (Handa 2003), *Beta* (357 sites) (Mower and Palmer 2006), *Vitis* (401 sites) (Picardi et al. 2010), *Silene* (287 sites in *Silene latifolia* and 189 sites in *Silene noctiflora*) (Sloan et al. 2010), *Citrullus* (463 sites) (Alverson et al. 2010), *Cucurbita* (444 sites) (Alverson et al. 2010), *Amborella* (779 sites) (Rice et al. 2013), *Liriodendron* (781 sites) (Richardson et al. 2013), *Oenothera* (362 sites) (Richardson et al. 2013) and *Nicotiana* (463 sites) (Richardson et al. 2013). These studies show that the C-to-U RNA editing occurs ~200-800 sites per species in angiosperms.

Unlike the relatively rich sequencing data from angiosperms, RNA editing analyses from other vascular plants are very limited. In gymnosperms, only three species have been described the RNA editing frequency: in *Cycas*, it was estimated to retain approximately 1200 editing sites (Chaw et al. 2008); in *Ginkgo*, there was 1306 predicted editing sites; whereas in *Welwitschia*, only 226 sites were estimated in its coding genes (Guo et al.

2016). In addition to the gymnosperms, there are three reported lycophytes, *I. engelmannii*, *S. moellendorffii*, and *H. squarrosa* with editing counts of 1782, 2152 and ~300, respectively (Grewe et al. 2009, Hecht et al. 2011, Liu et al. 2012). In bryophytes, *Physcomitrella patens* has the minimal identified editing sites of only 11 (Rüdinger et al. 2009), while the liverwort lineage Marchantiidae have no edit sites (Groth-Malonek et al. 2007, Oda et al. 1992, Rudinger et al. 2008). Thus, although sequencing results from species outside of the angiosperms is limited, it can be deduced that the frequency of RNA editing is substantially different in land plants.

4.2.2 RNA Editing Loss with Retroprocessing

Retroprocessing usually refers to the re-integration of a reverse transcribed transcript into the genome. A gene that has been retroprocessed could result in the loss of introns, along with the elimination of RNA editing sites (Bowe and dePamphilis 1996, Ran et al. 2010). However, the extent of RNA editing and intron loss can occur to different levels. First, it is possible that retroprocessing could affect the entire mature transcript, which means all of the introns would be lost and all of the U's produced by editing would be seen as T's in the retroprocessed genes (Hepburn et al. 2012). For instance, in the *rps3* gene of many conifers, two introns shared by many gymnosperm species were missing, and no editing sites were detected from this gene (Ran et al. 2010). A second possible outcome of retroprocessing is that the process may affect only part of the gene, resulting in the loss of some introns and the nearby edit sites, but distant introns and edit sites would remain. This model was supported from several previous studies, such as the *cox2* and *nad7* genes in *Silene* and the *cox2*, *nad1*, and *nad2* genes in *Isoetes*, in which an intron and surrounding edit sites were lost, while other introns and edit sites remain (Grewe et al.

2011, Sloan et al. 2010). In the third outcome, reverse transcription of a spliced but mostly unedited transcript would result in a sporadic loss of edit sites (Hepburn et al. 2012). Evidence for this process is difficult to obtain, because it would not remove many edit sites.

4.3 Plastid Genome Reduction in Parasitic Plants

Starting with the investigation of the plastid genome from holoparasite *Epifagus virginiana* over 20 years ago (Wolfe et al. 1992), the pseudogenization and loss of the plastid genes from parasitic plants has been described many times (Krause 2012, Krause and Scharff 2014). The heterotrophic lifestyle has a dramatic impact on the plastome, particularly in holoparasites in which the photosynthetic ability has been lost. It also has been speculated that the early stage of plastome reduction was the loss of many noncoding and possibly unimportant parts of the plastome sequences (Funk et al. 2007). So far, more than 20 plastomes from parasitic plants have been published including 12 complete plastomes from Orobanchaceae (Li et al. 2013, Uribe-Convers et al. 2014, Wicke et al. 2013), which displays a varying degree of reduction. Previous studies have documented gene loss affecting *ndh* genes, photosystem genes, the *rbcL* gene, ribosomal protein genes, *rpo* genes, hypothetical conserved reading frames (*ycfs*) and tRNAs (Delannoy et al. 2011, McNeal et al. 2007, Morden et al. 1991, van der Kooij et al. 2000, Wolfe et al. 1992).

Overall, as the adaptations to parasitism becomes more pronounced, resulting in more pseudogenization and functional loss of the plastid genes (Krause 2012). For those parasitic plants which still retain the ability of photosynthesis, their plastomes generally

retain the genes coding for photosynthetic components and the transcriptional and translational apparatus. For instance, the *Schwalbea americana* (Orobanchaceae) plastome has experienced no gene loss with only several genes pseudogenized. For those non-photosynthetic plants, massive gene loss has been observed frequently. Based on the current sequencing results, typically 27-35 genes are retained in the plastome (Barrett et al. 2014, Wicke et al. 2011, 2013). For example, the holoparasite *Cuscuta* lacks the subunits of the NADH dehydrogenase complex (*ndh* genes), some ribosomal protein genes and a few more genes, such as *psaI*, *matK*, *rpo* genes, etc (Funk et al. 2007). A very recent study on the holoparastie *Hydnora visseri* (Hydnoraceae) also revealed the elimination of all photosynthesis-related genes (Naumann et al. 2016). One of the remarkable examples of plastome reduction is in the holoparasite *Rafflesia lagascae* (Rafflesiaceae). Whole genomic DNA studies from this particular organism tentatively suggested that the entire plastid genome has been lost (Molina et al. 2014).

5 Research Goals

Green plant organellar genomes exhibit widespread diversity across lineages over the millions of years of evolution. The mitogenomes have conservative features to some extent (e.g. gene content), some very divergent features (e.g. genome size), along with some unique features (e.g. frequent HGT and RNA editing). In some cases, this diversity may be attributable to lifestyle. For instance, the mitogenomes of parasitic mistletoes in *Viscum* and heterotrophic algae in *Polytomella* are extremely reduced in size and content. Plastome features are highly conserved in most green plants, such as quadripartite structure and the gene content related to photosynthesis, but there is some extreme

diversity based on plant lifestyles and living environments, most often associated with the loss of photosynthetic genes due to a non-photosynthetic lifestyle. In my research, I performed a comparative analysis of organellar genomes from plants with atypical lifestyles to better understand the effects of their lifestyles on organelle genome size, structure, and genetic content.

In Chapter 2, I discuss the organelle genomic diversity in free-living and endosymbiotic green algae using newly sequenced and assembled genomes from three green algae species: *Chlorella heliozoae*, *Chlorella* sp. ATCC30562 and *Micractinium conductrix*. These species are endosymbionts living in different hosts, although all of them can also survive after isolation from their hosts. By comparing their organelle genomes with those of selected free-living algae, I examined whether and how their endosymbiotic lifestyle affects the evolution of organellar genomes.

Chapter 3 examines the organelle genomic diversity in parasitic plants. The effect of a parasitic plant lifestyle has been relatively well carried out for the chloroplast genome, but there is little information for the mitogenome counterparts. The Orobanchaceae family contains non-parasitic, hemiparasitic and holoparasitic lineages, so I analyzed genomic data from representative species within this group in order to assess the organellar genomic diversity in parasitic plants.

The goal of Chapter 4 is to analyze one of the peculiar features of plant mitogenomes: RNA editing. In an earlier publication, I helped to show that there is a massive loss of RNA editing sites from the xerophyte *Welwitschia mirabilis*, which is a compelling

contrast to its closely related species. In this thesis chapter, I used this example to explore the various mechanisms to possibly explain this massive amount of RNA editing loss.

REFERENCES

Adams, Palmer. 2003. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Mol Phylogenet Evol 29: 380-395.

Adams, Qiu YL, Stoutemyer M, Palmer JD. 2002. Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. Proc Natl Acad Sci USA 99: 9905-9912.

Alverson AJ, Wei X, Rice DW, Stern DB, Barry K, Palmer JD. 2010. Insights into the evolution of mitochondrial genome size from complete sequences of Citrullus lanatus and Cucurbita pepo (Cucurbitaceae). Mol Biol Evol 27: 1436-1448.

Arimura S, Yamamoto J, Aida GP, Nakazono M, Tsutsumi N. 2004. Frequent fusion and fission of plant mitochondria with unequal nucleoid distribution. Proc Natl Acad Sci USA 101: 7805-7808.

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

Barrett CF, Freudenstein JV, Li J, Mayfield-Jones DR, Perez L, Pires JC, Santos C. 2014. Investigating the path of plastid genome degradation in an early-transitional clade of heterotrophic orchids, and implications for heterotrophic angiosperms. Mol Biol Evol 31: 3095-3112.

Beck N, Lang B. 2009. RNAweasel, a webserver for identification of mitochondrial, structured RNAs. Montreal (Quebec): University of Montreal.

Beck N, Lang B. 2010. MFannot, organelle genome annotation websever.

Bergthorsson U, Richardson AO, Young GJ, Goertzen LR, Palmer JD. 2004. Massive horizontal transfer of mitochondrial genes from diverse land plant donors to the basal angiosperm Amborella. Proc Natl Acad Sci USA 101: 17747-17752.

Bock. 2010. The give-and-take of DNA: horizontal gene transfer in plants. Trends Plant Sci 15: 11-22.

Bock R. 2007. Structure, function, and inheritance of plastid genomes. Pages 29-63. Cell and molecular biology of plastids, Springer.

Boer PH, Bonen L, Lee RW, Gray MW. 1985. Genes for respiratory chain proteins and ribosomal RNAs are present on a 16-kilobase-pair DNA species from Chlamydomonas reinhardtii mitochondria. Proc Natl Acad Sci USA 82: 3340-3344.

Bowe LM, dePamphilis CW. 1996. Effects of RNA editing and gene processing on phylogenetic reconstruction. Mol Biol Evol 13: 1159-1166.

Brouard JS, Otis C, Lemieux C, Turmel M. 2010. The exceptionally large chloroplast genome of the green alga Floydiella terrestris illuminates the evolutionary history of the Chlorophyceae. Genome Biol Evol 2: 240-256.

Burger G, Nedelcu AM. 2012. Mitochondrial genomes of algae. Pages 127-157. Genomics of Chloroplasts and Mitochondria, Springer.

Castandet B, Choury D, Bégu D, Jordana X, Araya A. 2010. Intron RNA editing is essential for splicing in plant mitochondria. Nucleic Acids Res: gkq591.

Chapdelaine Y, Bonen L. 1991. The wheat mitochondrial gene for subunit I of the NADH dehydrogenase complex: a trans-splicing model for this gene-in-pieces. Cell 65: 465-472.

Chaw S-M, Shih AC-C, Wang D, Wu Y-W, Liu S-M. 2008. The mitochondrial genome of the gymnosperm Cycas taitungensis contains a novel family of short interspersed elements, Bpu sequences, and abundant RNA editing sites. Mol Biol Evol 25: 603-615.

Cho Y, Qiu YL, Kuhlman P, Palmer JD. 1998. Explosive invasion of plant mitochondria by a group I intron. Proc Natl Acad Sci USA 95: 14244-14249.

Chumley TW, Palmer 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.

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

Davis CC, Xi Z. 2015. Horizontal gene transfer in parasitic plants. Curr Opin Plant Biol 26: 14-19.

Davis CC, Anderson WR, Wurdack KJ. 2005. Gene transfer from a parasitic flowering plant to a fern. Proc Biol Sci 272: 2237-2242.

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

Delannoy E, Fujii S, Colas des Francs-Small C, 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.

Denovan-Wright EM, Nedelcu AM, Lee RW. 1998. Complete sequence of the mitochondrial DNA of Chlamydomonas eugametos. Plant Mol Biol 36: 285-295.

Douglas SE, Turner S. 1991. Molecular evidence for the origin of plastids from a cyanobacterium-like ancestor. J Mol Evol 33: 267-273.

Fan J, Lee RW. 2002. Mitochondrial genome of the colorless green alga Polytomella parva: two linear DNA molecules with homologous inverted repeat Termini. Mol Biol Evol 19: 999-1007.

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

Giege P, Brennicke A. 1999. RNA editing in Arabidopsis mitochondria effects 441 C to U changes in ORFs. Proc Natl Acad Sci USA 96: 15324-15329.

Giegé P, Brennicke A. 1999. RNA editing in Arabidopsis mitochondria effects 441 C to U changes in ORFs. Proc Natl Acad Sci USA 96: 15324-15329.

Gray MW, Burger G, Lang BF. 1999. Mitochondrial evolution. Science 283: 1476-1481.

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 Isoetes engelmannii. Nucleic Acids Res 37: 5093-5104.

Grewe F, Herres S, Viehover P, Polsakiewicz M, Weisshaar B, Knoop V. 2011. A unique transcriptome: 1782 positions of RNA editing alter 1406 codon identities in mitochondrial mRNAs of the lycophyte Isoetes engelmannii. Nucleic Acids Res 39: 2890-2902.

Groth-Malonek M, Wahrmund U, Polsakiewicz M, Knoop V. 2007. Evolution of a pseudogene: exclusive survival of a functional mitochondrial nad7 gene supports Haplomitrium as the earliest liverwort lineage and proposes a secondary loss of RNA editing in Marchantiidae. Mol Biol Evol 24: 1068-1074.

Guo W, Grewe F, Fan W, Young GJ, Knoop V, Palmer JD, Mower JP. 2016. *Ginkgo* and *Welwitschia* Mitogenomes Reveal Extreme Contrasts in Gymnosperm Mitochondrial Evolution. Mol Biol Evol 33: 1448-1460.

Handa H. 2003. The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (Brassica napus L.): comparative analysis of the mitochondrial genomes of rapeseed and Arabidopsis thaliana. Nucleic Acids Res 31: 5907-5916.

Hecht J, Grewe F, Knoop V. 2011. Extreme RNA editing in coding islands and abundant microsatellites in repeat sequences of Selaginella moellendorffii mitochondria: the root of frequent plant mtDNA recombination in early tracheophytes. Genome Biol Evol 3: 344-358.

Hepburn NJ, Schmidt DW, Mower JP. 2012. Loss of two introns from the Magnolia tripetala mitochondrial cox2 gene implicates horizontal gene transfer and gene conversion as a novel mechanism of intron loss. Mol Biol Evol 29: 3111-3120.

Hiesel R, Wissinger B, Schuster W, Brennicke A. 1989. RNA editing in plant mitochondria. Science 246: 1632-1634.

Kitazaki K, Kubo T. 2010. Cost of having the largest mitochondrial genome: evolutionary mechanism of plant mitochondrial genome. Journal of Botany 2010.

Koulintchenko M, Konstantinov Y, Dietrich A. 2003. Plant mitochondria actively import DNA via the permeability transition pore complex. EMBO J 22: 1245-1254.

Krause K. 2012. Plastid genomes of parasitic plants: a trail of reductions and losses. Pages 79-103. Organelle genetics, Springer.

Krause K, Scharff LB. 2014. Reduced Genomes from Parasitic Plant Plastids: Templates for Minimal Plastomes? Pages 97-115. Progress in Botany, Springer.

Kroymann J, Zetsche K. 1998. The mitochondrial genome of Chlorogonium elongatum inferred from the complete sequence. J Mol Evol 47: 431-440.

Kugita M, Yamamoto Y, Fujikawa T, Matsumoto T, Yoshinaga K. 2003. RNA editing in hornwort chloroplasts makes more than half the genes functional. Nucleic Acids Res 31: 2417-2423.

Lang BF, Nedelcu AM. 2012. Plastid Genomes of Algae. 35: 59-87.

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.

Li X, Zhang TC, Qiao Q, Ren Z, Zhao J, Yonezawa T, Hasegawa M, Crabbe MJ, Li J, Zhong Y. 2013. Complete chloroplast genome sequence of holoparasite Cistanche deserticola (Orobanchaceae) reveals gene loss and horizontal gene transfer from its host Haloxylon ammodendron (Chenopodiaceae). PLoS One 8: e58747.

Liu Y, Wang B, Cui P, Li L, Xue J-Y, Yu J, Qiu Y-L. 2012. The mitochondrial genome of the lycophyte Huperzia squarrosa: the most archaic form in vascular plants. PLoS One 7: e35168.

Lonsdale DM, Hodge TP, Fauron CM-R. 1984. The physical map and organisatlon of the mitochondrial genome from the fertile cytoplasm of maize. Nucleic acids Res 12: 9249-9261.

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

Mayer K, Schüller C, Wambutt R, Murphy G, Volckaert G, Pohl T, Düsterhöft A, Stiekema W, Entian K-D, Terryn N. 1999. Sequence and analysis of chromosome 4 of the plant Arabidopsis thaliana. Nature 402: 769-777.

McNeal JR, Kuehl JV, Boore JL, de Pamphilis CW. 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, Kuehl JV, Boore JL, Leebens-Mack J, dePamphilis CW. 2009. Parallel loss of plastid introns and their maturase in the genus Cuscuta. PLoS One 4: e5982.

Molina J, et al. 2014. Possible loss of the chloroplast genome in the parasitic flowering plant Rafflesia lagascae (Rafflesiaceae). Mol Biol Evol 31: 793-803.

Morden CW, Wolfe KH, dePamphilis CW, Palmer JD. 1991. Plastid translation and transcription genes in a non-photosynthetic plant: intact, missing and pseudo genes. EMBO J 10: 3281-3288.

Mower, Palmer JD. 2006. Patterns of partial RNA editing in mitochondrial genes of Beta vulgaris. Mol Genet Genomics 276: 285-293.

Mower, Sloan DB, Alverson AJ. 2012. Plant mitochondrial genome diversity: the genomics revolution. Pages 123-144. Plant Genome Diversity Volume 1, Springer.

Mower, Stefanović S, Young GJ, Palmer JD. 2004. Plant genetics: gene transfer from parasitic to host plants. Nature 432: 165-166.

Mower, Stefanovic S, Hao W, Gummow JS, Jain K, Ahmed D, Palmer JD. 2010. Horizontal acquisition of multiple mitochondrial genes from a parasitic plant followed by gene conversion with host mitochondrial genes. BMC Biol 8: 150.

Naumann J, et al. 2016. Detecting and Characterizing the Highly Divergent Plastid Genome of the Nonphotosynthetic Parasitic Plant Hydnora visseri (Hydnoraceae). Genome Biol Evol 8: 345-363.

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.

Oda K, et al. 1992. Gene organization deduced from the complete sequence of liverwort Marchantia polymorpha mitochondrial DNA. A primitive form of plant mitochondrial genome. J Mol Biol 223: 1-7.

Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M, Chang Z. 1986. Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA. Nature 322: 572-574.

Palmer JD. 1997. Organelle genomes: going, going, gone! Science 275: 790-791.

Petersen G, Cuenca A, Moller IM, Seberg O. 2015. Massive gene loss in mistletoe (Viscum, Viscaceae) mitochondria. Sci Rep 5: 17588.

Picardi E, Horner DS, Chiara M, Schiavon R, Valle G, Pesole G. 2010. Large-scale detection and analysis of RNA editing in grape mtDNA by RNA deep-sequencing. Nucleic Acids Res 38: 4755-4767.

Ran J-H, Gao H, Wang X-Q. 2010. Fast evolution of the retroprocessed mitochondrial rps3 gene in Conifer II and further evidence for the phylogeny of gymnosperms. Mol. Phylogenet. Evol 54: 136-149.

Reyes-Prieto A, Weber AP, Bhattacharya D. 2007. The origin and establishment of the plastid in algae and plants. Annu. Rev. Genet. 41: 147-168.

Rice DW, et al. 2013. Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm Amborella. Science 342: 1468-1473.

Richardson, Palmer JD. 2007. Horizontal gene transfer in plants. Journal of experimental botany 58: 1-9.

Richardson, Rice DW, Young GJ, Alverson AJ, Palmer JD. 2013. The "fossilized" mitochondrial genome of Liriodendron tulipifera: ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. BMC Biol 11: 29.

Rose RJ, Sheahan MB. 2007. Plant Mitochondria. eLS.

Rudinger M, Polsakiewicz M, Knoop V. 2008. Organellar RNA editing and plant-specific extensions of pentatricopeptide repeat proteins in jungermanniid but not in marchantiid liverworts. Mol Biol Evol 25: 1405-1414.

Rüdinger M, Funk HT, Rensing SA, Maier UG, Knoop V. 2009. RNA editing: only eleven sites are present in the Physcomitrella patens mitochondrial transcriptome and a universal nomenclature proposal. Mol Genet Genomics 281: 473-481.

Ruhlman TA, Jansen RK. 2014. The plastid genomes of flowering plants. Methods Mol Biol 1132: 3-38.

Salanoubat M, et al. 2000. Sequence and analysis of chromosome 3 of the plant Arabidopsis thaliana. Nature 408: 820-822.

Sanchez-Puerta. 2014. Involvement of plastid, mitochondrial and nuclear genomes in plant-to-plant horizontal gene transfer. Acta Societatis Botanicorum Poloniae 83: 317-323.

Sanchez-Puerta, Cho Y, Mower JP, Alverson AJ, Palmer JD. 2008. Frequent, phylogenetically local horizontal transfer of the cox1 group I intron in flowering plant mitochondria. Mol Biol Evol 25: 1762-1777.

Scharff LB, Koop HU. 2006. Linear molecules of tobacco ptDNA end at known replication origins and additional loci. Plant Mol Biol 62: 611-621.

Shinozaki K, et al. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J 5: 2043-2049.

Simpson MG. 2010. Plant systematics: Academic press.

Skippington E, Barkman TJ, Rice DW, Palmer JD. 2015. Miniaturized mitogenome of the parasitic plant Viscum scurruloideum is extremely divergent and dynamic and has lost all nad genes. Proc Natl Acad Sci USA 112: E3515-E3524.

Sloan DB, MacQueen AH, Alverson AJ, Palmer JD, Taylor DR. 2010. Extensive loss of RNA editing sites in rapidly evolving Silene mitochondrial genomes: selection vs. retroprocessing as the driving force. Genetics 185: 1369-1380.

Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR. 2012. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol 10: e1001241.

Smith DR, Lee RW. 2008. Mitochondrial genome of the colorless green alga Polytomella capuana: a linear molecule with an unprecedented GC content. Mol Biol Evol 25: 487-496.

Smith DR, Lee RW. 2014. A plastid without a genome: evidence from the nonphotosynthetic green algal genus Polytomella. Plant Physiol 164: 1812-1819.

Smith DR, Lee RW, Cushman JC, Magnuson JK, Tran D, Polle JE. 2010. The Dunaliella salina organelle genomes: large sequences, inflated with intronic and intergenic DNA. BMC Plant Biol 10: 83.

Tabata S, et al. 2000. Sequence and analysis of chromosome 5 of the plant Arabidopsis thaliana. Nature 408: 823-826.

Takenaka M, Verbitskiy D, van der Merwe JA, Zehrmann A, Brennicke A. 2008. The process of RNA editing in plant mitochondria. Mitochondrion 8: 35-46.

Theologis A, et al. 2000. Sequence and analysis of chromosome 1 of the plant Arabidopsis thaliana. Nature 408: 816-820.

Tovar J, Fischer A, Clark CG. 1999. The mitosome, a novel organelle related to mitochondria in the amitochondrial parasite Entamoeba histolytica. Mol Microbiol 32: 1013-1021.

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. 2007. An unexpectedly large and loosely packed mitochondrial genome in the charophycean green alga Chlorokybus atmophyticus. BMC Genomics 8: 137.

Turmel M, Lemieux C, Burger G, Lang BF, Otis C, Plante I, Gray MW. 1999. The complete mitochondrial DNA sequences of Nephroselmis olivacea and Pedinomonas minor. Two radically different evolutionary patterns within green algae. Plant Cell 11: 1717-1730.

Unseld M, Marienfeld JR, Brandt P, Brennicke A. 1997. The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924. Nat genet 15: 57-61.

Uribe-Convers S, Duke JR, Moore MJ, Tank DC. 2014. A long PCR-based approach for DNA enrichment prior to next-generation sequencing for systematic studies. Applications in Plant Science 2.

van der Kooij TA, Krause K, Dorr I, Krupinska K. 2000. Molecular, functional and ultrastructural characterisation of plastids from six species of the parasitic flowering plant genus Cuscuta. Planta 210: 701-707.

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.

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.

Welsh EA, et al. 2008. The genome of Cyanothece 51142, a unicellular diazotrophic cyanobacterium important in the marine nitrogen cycle. Proc Natl Acad Sci USA 105: 15094-15099.

Wicke, Schneeweiss GM, 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.

Wicke, Muller KF, de Pamphilis CW, Quandt D, Wickett NJ, Zhang Y, Renner SS, Schneeweiss GM. 2013. Mechanisms of functional and physical genome reduction in photosynthetic and nonphotosynthetic parasitic plants of the broomrape family. The Plant Cell 25: 3711-3725.

Wissinger B, Schuster W, Brennicke A. 1991. Trans splicing in Oenothera mitochondria: nad1 mRNAs are edited in exon and trans-splicing group II intron sequences. Cell 65: 473-482.

Wolf PG. 2012. Plastid genome diversity. Pages 145-154. Plant Genome Diversity Volume 1, Springer.

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.

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.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20: 3252-3255.

Xi Z, Wang Y, Bradley RK, Sugumaran M, Marx CJ, Rest JS, Davis CC. 2013. Massive mitochondrial gene transfer in a parasitic flowering plant clade. PLoS Genet 9: e1003265.

Xi Z, Bradley RK, Wurdack KJ, Wong K, Sugumaran M, Bomblies K, Rest JS, Davis CC. 2012. Horizontal transfer of expressed genes in a parasitic flowering plant. BMC Genomics 13: 227.

Xue JY, Liu Y, Li L, Wang B, Qiu YL. 2010. The complete mitochondrial genome sequence of the hornwort Phaeoceros laevis: retention of many ancient pseudogenes and conservative evolution of mitochondrial genomes in hornworts. Curr Genet 56: 53-61.

Zhang M, Pereira e Silva Mde C, Chaib De Mares M, van Elsas JD. 2014. The mycosphere constitutes an arena for horizontal gene transfer with strong evolutionary implications for bacterial-fungal interactions. FEMS Microbiol Ecol 89: 516-526.

TABLES

Table 1-1. Overview of gene content in green plants mitogenome

Function	Genes
NADH dehydrogenase subunits (Complex I)	*nad1, nad2, nad3, nad4, nad4L, nad5, nad6, nad7, nad9, nad10**
Succinate dehydrogenase (Complex II)	*sdh3, sdh4*
Cytochrome bc_1 complex subunits (Complex III)	*cob*
Cytochrome *c* oxidase subunits (Complex IV)	*cox1, cox2, cox3*
ATP synthase subunits (Complex V)	*atp1, atp4, atp6, atp8, atp9*
Cytochrome *c* maturation proteins	*ccmB, ccmC, ccmFn, ccmFc*
Ribosomal proteins	*rpl2, rpl5, rpl6, rpl10, rpl14*, rpl16, rpl31*, rps1, rps2, rps3, rps4, rps7, rps8, rps10, rps11, rps12, rps13, rps14, rps19*
A putative protein transporter	*mttB/tatC*
Maturase-related protein	*matR*

*green algae only

Table 1-2. Plastid protein-encoding genes in green plants

Function		Genes
Photosynthetic dark reaction related proteins		
	inner membrane protein	*cemA*
	Protochloro-phylide reductase	*chlB, chlI*, chlL, chlN*
	Cytochrome c biogenesis protein	*ccsA*
	large subunit of RuBisCO	*rbcL*
Photosynthetic light reactions related proteins		
	ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	Photosystem I	*psaA, psaB, psaC, psaI, psaJ, psaM**
	Photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ*
	NADH dehydrogenase+	*ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
	Cytochrome complex	*petA, petB, petD, petG, petL, petN*+
	Photosystem I assembly factor	*ycf3, ycf4*
Proteins not related to photosynthesis		
	involved in lipid acid synthesis	*accD*
	cell division	*ftsh**
	putative ABC-containing sulfate transporter genes	*cysA, cyst*
	protein quality control	*clpP*
	homing Endonuclease	*I-CvuI**
	rRNA Ile-lysidine synthetase	*tilS**
Genetic apparatus		*matK*+
Translation and post-transcriptional modification		
	DNA-dependent RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
Translation and protein-modifying enzymes		
	translation factor	*infA, tufA**
	division factor	*minD*, minE**
	large ribosomal proteins	*rpl2, rpl5*, rpl12*, rpl14, rpl16, rpl19*, rpl20, rpl21, rpl22, rpl23, rpl32, rpl33, rpl36*
	small ribosomal proteins	*rps2, rps3, rps4, rps7, rps8, rps9*, rps11, rps12, rps14, rps15, rps16, rps18, rps19*
Proteins of unknown function		*ycf1, ycf2, ycf12*, ycf20*, ycf47**

* algae only

\+ land plants only

CHAPTER 2

Comparative Organellar Genomics of Chlorellales: Genomic Effects of an Endosymbiont Lifestyle?

Weishu Fan, Wenhu Guo, James L. Van Etten and Jeffrey P. Mower

ABSTRACT

Endosymbiotic bacteria have been reported with extraordinary reduced genome in numerous cases. Many endosymbiotic green algae also show extreme genomic reduction of their nuclear genomes, but they may retain a fully functional plastid genome if they maintain photosynthetic ability or if they can survive outside of their host. In order to better understand how the endosymbiotic lifestyle has affected the organellar genomes of photosynthetic green algae, we generated the complete organellar genome sequences from three green algal endosymbionts (*Chlorella heliozoae*, *Chlorella* sp. ATCC30562 and *Micractinium conductrix*) isolated from a ciliate and heliozoan. Compared with other trebouxiophycean green algae, the three newly sequenced species have regular genome size and gene content in both the mitochondria and the plastid genome, providing no evidence for organellar genomic reduction in these endosymbionts. Instead, the organellar genomes of the three endosymbionts are generally larger and more intron rich than other species of *Chlorella*. Phylogenetic analysis of plastid and mitochondrial genes demonstrated that *M. conductrix* clusters together with *Chlorella* strains, suggesting that it should be considered a species of *Chlorella*. In addition, the three endosymbionts do not form a monophyletic group, indicating that the endosymbiotic lifestyle has evolved multiple times within *Chlorella*.

INTRODUCTION

The mitochondrion originated ~1.5 billion years ago from an α-Proteobacterial endosymbiont. The ancestral genome of the endosymbiotic organism contained thousands of genes whereas the mitogenome of eukaryotes today typically retain <100 genes, indicating massive reduction of gene content over time. The origin of plastids formed somewhat later (~1 billion years ago) through the primary endosymbiosis of a cyanobacterium, which is also widely recognized as the origin of the photosynthetic organelles. Today, only 5-10% of the genes have been retained in the plastome compared with their cyanobacteria ancestor (Martin et al. 2002, Wolf 2012). Another cyanobacterial endosymbiont, which was established as a novel photosynthetic organelle (the "chromatophore") of the amoeba *Paulinella chromatophora*, has a genome that was extensively reduced. The *P. chromatophora* chromatophore has retained only 26% of the genescompared with its free-living relative. Genome-wide reduction was also reported from other bacterial endosymbionts, such as *Candidatus* Carsonella ruddii and*Candidatus* Tremblaya princeps, which possess only ~2-30% of the regular genome size (Husnik et al. 2013, McCutcheon et al. 2009, Nakabachi et al. 2006, Sloan et al. 2014).

Some green and red algal lineages also live as endosymbionts of other eukaryotes. In many cases, the algal live as endosymbionts due to their photosynthetic abilities. This relationship has led to the retention of the algal plastid genome, while the remnant nucleus genome (nucleomorph) has lost most of its genes except for the chloroplast-located proteins, and the mitogenome of endosymbionts get lost. For example, the non-photosynthetic host in the cryptomonads captured a photosynthetic red alga, which serves

as an endosymbiont with a functional chloroplast and highly reduced nucleomorph genome (Douglas et al. 2001). Similarly, a green alga was captured by the chlorarachniophyte *Bigelowiellanatans*, and the endosymbiont alga has retained essential plastid protein genes (Gilson et al. 2006, Lane et al. 2007). Some green algal species form lichen together with fungi, and their organelle genomes have regular features similar to free-living algae (Del Campo et al. 2010). Conversely, reduction or even loss of the entire plastid genome tends to occur if the algae are no longer required to be photosynthetic. For example, the non-photosynthetic trebouxiophyte *Helicosporidium* sp.contains the smallest known plastid genome (37.5 kb), resulting from the lack of all photosynthesis-related genes and tiny intergenic spaces (de Koning and Keeling 2006). The malaria parasite *Plasmodium falciparum* also surprisingly has 35 kb circular apicoplast genome (Arisue et al. 2012). One notable example of plastid reduction in non-photosynthetic green algae is the free-living, freshwater unicellular green algal genus, *Polytomella*. Although it is closely related to the model algae species *Chlamydomonas reinhardtii*, no plastid genome was detectable through next generation sequencing (Smith et al. 2013).

Many green algae form symbiotic relationships with ciliates (e.g. *Paramecium*), heliozoans (e.g. *Acanthocystis*), and invertebrates (e.g. *Hydra*) (Reisser 1992, 1994). Irrespective of the mutual relationship between the endosymbionts and their host, the endosymbiotic algae retain the ability to live without the host, and are thus facultative endosymbionts. One well-studied endosymbiotic association involves the ciliate *Paramecium bursaria* and several species of green algae in *Chlorella*. Previous studies have shown that *P. bursaria* harbors many endosymbiont species, and these

endosymbiont infections can affect the gene expression of the host (Kodama et al. 2014). However, there are no studies that have examined the effect of this relationship on the organellar genomes of the endosymbionts and their ciliate hosts, which raises several outstanding questions: will the organellar genomes of the endosymbiotic green algae be reduced, similar to the cases in bacterial genomes and algal nucleomorph genomes? Alternatively, will they remain intact due to the requirement of full mitochondria and plastid function?

Trebouxiophyceae is a class in the Chlorophyta clade of green algae, comprising species forming symbioses with fungi to form lichens (e.g. *Trebouxia* and *Myrmecia*), photosynthetic symbionts in ciliates or plants (e.g. *Chorella* and *Elliptochloris*), non-photosynthetic species (e.g. *Prototheca* and *Helicosporodium*) and also free-living representatives (e.g. *Chlorella* sp. ArM0029B) (de Koning and Keeling 2006, Friedl and Rokitta 1997, Jeong et al. 2014, Perez-Ortega et al. 2010, Pombert et al. 2014, Proschold et al. 2011, Ueno et al. 2003). To date, a total of 35 plastomes and nine mitogenomes from Trebouxiophyceae have been fully sequenced. Plastid genome (plastome) sizes range from 37.5 kb to ~300 kb containing 54-114 genes, while mitochondrial genome (mitogenome) sizes vary from ~49 kb to ~85 kb with some ~65 genes. To explore organelle genome interactions in the endosymbiotic lifestyles, the complete plastid and mitochondrial genomes were sequenced from three unicellular, endosymbiotic green algae: *Chlorella* sp. ATCC30562 (isolated from *P. bursaria*), *Chlorella heliozoae* (isolated from *Acanthocystis turfacea*) and *Micractinium conductrix* (isolated from *P. bursaria*). Their genomes sequences were compared to the organellar genomes of other

trebouxiophycean green algae in order to assess the effects of an endosymbiont lifestyle on the organellar genomes of green algae.

MATERIALS AND METHODS

Source of species

Pelleted cell cultures of *Chlorella heliozoae, Chlorella* sp. ATCC30562 and *Micractinium conductrix* were obtained from Dr. Van Etten's lab (Morrison Center, University of Nebraska-Lincoln). All cells were stored at -80˚C prior to DNA extraction.

DNA extraction and genome sequencing

DNA extraction was performed using a modified preparation protocol (Anwaruzzaman et al. 2004, Lin et al. 2010). About 0.2g of the harvested cells were resuspended in 375 µL SDS-EB buffer (2% SDS, 100 mM Tris-HCl pH 8.0, 400 mM NaCl, 40 mM EDTA pH8.0), and then an equal volume of water was added, followed by 750 µL phenol:chloroform:isoamyl alcohol (25 : 24 : 1). To break the cell wall, 25 mg glass beads were added in the same tube, and the sample was vortexed for 5 min. Cellular debris was pelleted by centrifuging for 5 min at 12,000 ×g. The aqueous solution was transferred to a new tube, and 2 µL RNase (10 mM) was added, followed by 30 min incubation at 37 ˚C. The phenol:chloroform:isoamyl alcohol and centrifugation steps were repeated without using glass beads, and the aqueous solution was then treated with 750 µL chloroform. The supernatant was transferred to a new tube, and then twice the volume of 100% ethanol was added and incubated 1 hour at -20 ˚C. The DNA was

collected by centrifugation for 20 min at 12,000 followed by washing with 70% ethanol. The DNA was air dried and then resuspended in 50 µL water.

The DNA samples were sent to the High-Throughput DNA Sequencing and Genotyping Core Facility at UNMC (Omaha, NE). For each sample, 20–30 M reads of 100 bp were sequenced from a 500 bp paired-end library on an Illumina HiSeq 2500.

Genome assembly and annotation

The organellar genomes of *C.* sp. ATCC30562, *C. heliozoae* and *M. conductrix* were assembled from the Illumina sequence reads by running Velvet version 1.2.03 (Zerbino and Birney 2008) using different pairwise combinations of Kmer (61, 71, 81, 91) and expected coverage (50, 100, 200, 500, 1000) values, as described previously (Grewe et al. 2014, Zhu et al. 2014). Scaffolding was turned off with paired-end containing data set and 10% of expected coverage was set as the minimum coverage parameter. For each assembly, plastid and mitochondrial contigs were detected by blastn searches with known organellar gene sequences from related Chlorellales species used as queries. The final consensus sequence for each species was constructed by aligning the mitochondrial and plastid contigs from the best draft assemblies (that maximized the average length of plastid or mitochondrial contigs). Circular genomes were confirmed by aligning the overlapping terminal regions of the contigs, which was further supported by read pairs that spanned both ends of the assembly. Using this strategy, a single completed circular chromosome was assembled for the plastome and mitogenome of each species. To evaluate the depth of coverage of the genome assemblies, read pairs were mapped onto respective consensus sequences with Bowtie 2.0 (Langmead and Salzberg 2012).

Protein-coding genes from *C.* sp. ATCC30562, *C. heliozoae* and *M. conductrix* complete mitochondrial genomes were annotated by blast against the database from National Center for Biotechnology Information. The protein genes from plastome of the three organisms were initially annotated by using the Web-based annotation package Dual Organellar Genome Annotator (DOGMA) (Wyman et al. 2004) with a 60% cutoff and a Blast E-value of 1e-5, followed by manual adjustment as necessary. Genes coding for ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) were identified by blastn searches (de Koning and Keeling 2006) and tRNAscan-SE (Lowe and Eddy 1997), respectively. To identify potentially novel genes, blastn and blastx searches were also applied to all noncoding regions but no additional genes were identified. The annotated mitogenomes and plastomes will deposit in GenBank.

RNA extraction and cDNA sequencing

To verify the intron content in the mitochondrial large ribosomal RNA (*rrnL*), experimental approaches were applied to the three organisms. Total RNA isolation from *C. heliozoae*, *C.* sp. ATCC30562 and *M. conductrix* cells (~1×10^9 cells) was performed with the following modified Trizol protocol. Harvested cells were spread on the wall of the Eppendorf tube before freezing in liquid nitrogen for 1 min, then 3 ml of Trizol was immediately added to the frozen pellet and the tube was vortexed for 10 min. The homogenized sample was incubated for another 5 min at room temperature followed by centrifuge at 12,000 ×g for 10 min at 4˚C to remove insoluble material and polysaccharides. The supernatant was transferred to a new tube, 0.75 ml of chloroform was added with vigorously shaking for 30 seconds, and then the tube was vortexed for 2 min. An additional 5 min incubation was undertaken at room temperature before

centrifugation at 12,000 ×g for 15 min at 4˚C. The aqueous phase was transferred to a clean tube, an equal volume of phenol:chloroform:isoalcohol (25:24:1) was added, and then vortexed for 2 min. The centrifugation process was repeated, and then the aqueous phase of the sample was transferred to a new tube and an equal volume of isopropanol was added to the solution, followed by incubation at -20˚C for 30 min. The RNA was collected by centrifugation, washed, air dried and redissolved following the manufacturer's protocol.

With the isolated RNA as template, RT-PCR and cDNA sequencing were carried out by the approaches described previously (Hepburn et al. 2012). Species-specific primers were designed for various regions of *rrnL* to amplify the total length of this cDNA. The PCR-amplified *rrnL* cDNAs were Sanger sequenced on both strands at Genscript (NJ, GenScript USA Inc.).

Genome structural analyses

To compare the plastome and mitogenome structural organization of the three species sequenced in this study, alignments of whole genomes from these three species and three additional Chlorellales species were carried out using the ProgressiveMauve algorithm of Mauve 2.3.1 (Darling et al. 2010).

Phylogenetic analysis

Both plastid and mitochondrial phylogenies were generated in this study. In addition to the three newly sequenced species, organellar genomes from 22 representative chlorophytes and six streptophytes (Table S1) were collected from GenBank as ingroup and outgroup, respectively. Individual protein-coding genes were extracted with a

customized perl script and then manually checked to avoid misannotation. Those genes (74 plastid genes and 32 mitochondrial genes) that that were present in more than half of the taxa were aligned by codons using MUSCLE (version 3.8.31) (Edgar 2004), and manually adjusted in BioEdit (version 7.2.0) if necessary. Plastid and mitochondrial protein gene data sets were concatenated seperately by FASconCAT (version 1.0) (Kuck and Meusemann 2010). The ambiguously aligned regions in the concatenated alignments were excluded using Gblocks (version 0.91b) (Castresana 2000) with relaxed parameters (t=c, b2=15, b4=5, b5=half).

Phylogenetic analyses were inferred from plastid and mitochondrial data sets using the Maximum Likelihood (ML) approach in PhyML version 3.0 (Guindon et al. 2010) and Bayesian inference (BI) in MrBayes version 3.2 (Ronquist et al. 2012). ML trees were estimated with the GTR+G+I model and confidence of branching was estimated by bootstrap (BS) analyses with 1000 replicates. For BI analyses, the GTR+G model was used and other default parameters were applied to the runs. To ensure convergence during the BI runs, 100,000 and 200,000 generations were set for plastid and mitochondrial data sets, respectively, in order to make the standard deviation of split frequencies below 0.01.

RESULTS

Comparative analysis of Chlorellales mitochondrial genomes

The three newly sequenced endosymbiotic green algae *C. heliozoae*, *C.* sp. ATCC30562 and *M. conductrix* have circular mitochondrial genomes of 62477 bp, 79601 bp and 74708 bp in length, respectively (Figure S1). A broader comparison of mitogenomes

from species in Trebouxiophyceae shows that genome size varies from 49 kb to 79 kb (Table 2-1). Notably, the four endosymbionts (*C. heliozoae*, *C.* sp. ATCC30562, *C. variabilis* and *M. conductrix*) tend to have relatively larger mitogenome sizes compared with free-living individuals (*C.* sp.ArM0029B, *C. sorokiniana*, *Coccomyxa* sp. C-169 and *Trebouxiophyceae* sp.) and substantially larger mitogenome sizes compared with parasitic, non-photosynthetic green algae (*Helicosporidium* sp. and *Prototheca wickerhamii*). In terms of adenine and thymine (AT) content, all three newly sequenced algae are ~70% similar to that of *C.* sp. ArM0029B (71.5%), *C. sorokiniana* (70.9%) and *C. variabilis* (71.8%), but lower than the two AT-rich species *Helicosporidium* sp. (74.4%) and *P. wickerhamii* (74.2%), and in contrast to that of *Coccomyxa* sp. (46.9%) and *Trebouxiophyceae* sp. (46.6%) (Table 2-1).

Despite the wide range of sizes, it is noteworthy that the Trebouxiophyceae mitogenomes carry very similar gene content (Table 2-1). In fact, the mitogenomes from all five *Chlorella* species and *M. conductrix* share the same set of 32 protein-coding genes, three rRNAs, and 27 tRNAs, indicating that mitochondrial gene content is not drastically affected by the different lifestyles. In contrast to the stable gene content, introns are highly variable among Trebouxiophyceae mitogenomes (Table 2-2), ranging from a minimum of one intron (*C.* sp. ArM0029B and *C. sorokiniana*) to a maximum of 11 introns (*Trebouxiophyceae* sp.). The large ribosomal RNA (*rrnL*) contains the most introns, although there is substantial variation in content among the *Chlorella* and *M. conductrix* species (Table 2-2 and Figure 2-1). The free-living *C.* sp. ArM0029B does not have any introns, and the free-living *C. sorokiniana* has only one, while the four endosymbiotic Chlorellales species have 3-7 introns (Figure 2-1). Homology searches

against all Trebouxiophyceae and Pedinophyceae mitochondrial and plastid genomes and against all nuclear rRNA introns shows many matches between mitochondrial and plastid rRNA introns, suggesting the transfer of some rRNA introns between organelles.

To compare the structural diversification of the study group, we examined the syntenic segments of the genomes (Figure 2-2A), which showed there were a large number of inversion and/or translocation events occurred during Chlorellales evolution. However, the *C.* sp. ATCC30562 and *C. variabilis* displayed a highly conserved genome order in addition to the similarities in terms of genome size, gene content, AT content and intron content. This implies that they share the most recent common ancestor and are very closely related species.

Comparative analysis of Chlorellales plastid genomes

The plastid DNA sequences from *C. heliozoae*, *C.* sp.ATCC30562 and *M. conductrix* also have circularly mapping structures, with lengths of 124,353 bp, 124,881 bp, and 129,436 bp, respectively (Figure S2). They harbored 79 (*C. heliozoae* and *C.* sp. ATCC30562) or 78 (*M. conductrix*) protein-coding genes (Table 2-1). The missing gene in *M. conductrix* is tRNA(Ile)-lysidine synthetase (*tilS*), which is responsible for modifying the CAU anticodon of a unique tRNA that allows the amino acid change to isoleucine (Fabret et al. 2011). Comparing the general plastome structure features of these endosymbionts with other selected Trebouxiophyceae with various lifestyles, there is no indications of reduction in genome size or gene content. These endosymbiont plastid genomes show more similarities with free-living collections instead of parasitic counterparts according to plastome size and gene content. The significant gene loss

occurred in *Helicosporidium* sp.and *P.wickerhamii* due to loss of photosynthesis ability (de Koning and Keeling 2006, Pombert and Keeling 2010).

The AT content has the irregular distribution in this group: two of the four free-living algae (*C.* sp. C-169 and *Trebouxiophyceae* sp.) have less than 50% AT content; whereas the free-living *C.* sp. ArM0029Band *C. sorokiniana*, (66.1% and 65.9% in AT content, respectively) have AT content that is very close to the endosymbiotic Chlorellales lineages (64.7%-66.1%) lineages. The parasitic, non-photosynthetic algae (*Heliosporidium* sp. and *P. wickerhamii*) have the highes AT together with a drastic decrease in genome size (Table 2-1). Altogether, this suggests that the AT content does not significantly relate to their lifestyles.

A comparison of intron content among species also indicates substantial variation (Table 2-2). *M. conductrix* was determined to be extremely intron rich with 10 introns and four of them (*petB, psaC*, *psbD* and *rps12*) were not found in other comparable trebouxiophytes. The other taxa in Chlorellales contain 1-3 introns and the unclassified *Trebouxiophyceae* sp. retains five introns (Table 2-2).

Plastome structural and organization of the recently sequenced endosymbiotic chlorophytes and *C. variabilis* were compared with two free-living green algae. All of the six comparable species do not carry the inverted repeat (IR) region (Figure 2-2B) (Jeong et al. 2014). Overall, the plastomes are highly conserved with large blocks having complete synteny and a few inversion events (Figure 2-2B). It is also difficult to tell the discrepancies correlated to their living styles. Notably, *C.* sp. ATCC30562 and *C. variabilis* display exactly the same gene order which is consistent with the highly

similarity in their mitogenome. Moreover, most of the gene clusters conserved in green algae (Turmel et al. 2009) are also conserved in these three endosymbionts. Our result also shows that the gene order of "*trnC-rpoB-rpoC1-rpoC2-rbcL-rps14*" matches *M. conductrix* and *C.* sp. ATCC30562, but not in *C. heliozoae*, which contradicts a previous study suggesting that this cluster is well conserved and may be specific to *Chlorella* species (Jeong et al. 2014). The overlap of the 5' coding region of the *psbC* with the 3' coding region of the *psbD* gene, which occurs in most of the Trebouxiophyceae sequences (Jeong et al. 2014), also exists in all three sequenced endosymbiotic algae.

Phylogenetic analysis

In order to evaluate and understand evolution within Chlorophyta, we performed ML (GTR+G+I model) and BI (GTR+G model) phylogenetic analyses on data sets containing 32 mitochondrial genes or 74 plastid genes from the three new Chlorellaceae genomes plus selected green algae of which complete organelle genome are available (Table S1). The phylogenetic relationships of the three endosymbionts within the Chlorophyta clade was investigated using the phylogenetic trees inferred from 74 plastid genes (Figure 2-3A) and 32 mitochondrial genes (Figure 2-3B) of 22 chlorophytes and six streptophytes.

Consistent phylogenies constructed by either the plastid or mitochondrial genes indicate that all five *Chlorella* species and *M. conductrix* group together with maximal support (100%). In other words, both trees, using ML and BI method, show strong support that *M. conductrix* is nested in the clade of *Chlorella*. This indicates that *Micractinium* should not be considered as a separate genus, but should be considered a species of *Chlorella* (Hoshina Ryo et al. 2010). The grouping of *C. variabilis* and *C.* sp. ATCC30562 also

received maximal statistical support in both trees (Figure2-3), which is accordant with the strong similarities in their plastomes and mitogenomes.

The phylogenies also suggest that the endosymbiont lifestyle evolved multiple times in *Chlorella.* The common ancestor of four endosymbionts and the two free-living *Chlorella* species received maximal support in both analyses, as reflected by the BS value and posterior probability (PP) of 100% (Figure 2-3). In both trees, *C.* sp. ArM0029B, a free-living green alga, was positioned as sister taxa with an endosymbiotic one, *M. conductrix* with strong support (BS>85%, PP=100%); while the free-living *C. sorokiniana*, was sister to the endosymbiotic *C. heliozoae* (BS=73%, PP=100% inferred from plastid genes). The closer genus to *Chlorella* spp. was the two parasitic non-photosynthetic *P. wickerhamii* and *Helicosporidium* sp. with strong support from the mitochondrial dataset (100% BS/PP support) (Figure 2-3B). Given the phylogenetic affiliation, the four endosymbiont species do not group together in a single clade. Therefore, the non-monophyletic endosymbionts assemblage implies that the endosymbiont lifestyle evolved more than once in *Chlorella*. Alternative topologies constrain the endosymbiotic lineages to a single clade forming a monophyletic group were rejected by the Shimodaira-Hasegawa (SH) Test (Shimodaira 2002) with $P<0.05$ ($P=0.000^*$). This result further supports the multiple evolution events, consistent with the previous study of symbioses in *P. bursaria* (Hoshina and Imamura 2008).

DISCUSSION

A lack of organellar genome reduction in the green algal endosymbionts

In this study, we determined the complete mitochondrial and plastid genomes from *C.* sp. ATCC30562, *C. heliozoae* and *M. conductrix* to elucidate the genomic effects of an endosymbiotic lifestyle and their phylogenetic positions within Trebouxiophyceae. These three newly sequenced species are all endosymbionts, isolated from *P. bursaria* (*C.* sp. ATCC30562 and *M. conductrix*) or *A. turfacea* (*C. heliozoae*). However, they retained a regular genome size with an intact organellar gene content (Table 2-1), providing no evidence of functional degeneration as seen in the cases of algal nucleomorph genomes or endosymbiotic bacteria. In the plastid genome, the newly sequenced green algal species retained 78 or 79 genes, which is the same number compared with free-living trebouxiophyte species (*C.* sp. ArM0029B, *C. sorokiniana, C.* sp. C-169 and *Trebouxiophyceae* sp.) and another endosymbiotic *Chlorella* species (*C. variabilis*), but contrasts with the two heterotrophic algae (*Helicosporidium* sp. and *P. wickerhamii*) that have been sequenced from trebouxiophytes. *P. wickerhamii* is a nonphotosynthetic, predominantly free-living alga that is also an opportunistic vertebrate parasite, and it has lost all photosynthesis-related genes. *Helicosporidium* sp. is an obligate parasite of invertebrates and it lacks all genes for function in photosynthesis. Other than the selected species in our study, the loss of plastid-encoded photosynthesis-related genes or even the complete plastid genome has also been documented many times, such as *Cryptomonas paramecium* and *Polytomella* (Donaher et al. 2009, Smith et al. 2013). These examples suggest that the loss of photosynthesis causes the loss of most or all of the related genes, leading to dramatic changes in plastomes. Conversely, our three endosymbiont species

did not show a reduction in plastome gene content. Thus, these endosymbiotic algae appear to need the ability to create energy for themselves or for their hosts by photosynthesis. This ability also allows them to grow outside of their host cells, although they may need the additional resources, such as nitrogen and vitamins, in order to survive.

As to the mitogenome, based on the comparison among species in this study, there is a very similar set of protein-coding genes within the various trebouxiophytes with different lifestyles. This is to be expected since the mitochondrial genes are essential for respiration, and this metabolism process is required for algae in all living styles. In fact, there are limited examples of massive reduction or loss of the mitochondrial genome in eukaryotes. The mitogenome of the colorless green alga *Polytomella parva* encodes only seven protein genes in two linear mitochondrial DNA components (Fan and Lee 2002). Several other green algae, like *C. reinhardtii* and other members of the Chlamydomonadales, also have a reduced mitogenome relative to other green algae (Gray and Boer 1988). In addition, the recently described mitogenomes from hemiparasitic mistletoes (*Viscum* spp.) have lost all genes for complex I of the mitochondrial electron transport chain (Petersen et al. 2015, Skippington et al. 2015). Moreover, mitogenome reduction has also occurred in parasitic protists, such as the human malarial parasite *Plasmodium falciparum,* whose mitogenome harbors only three protein genes with only ~6 kb in genome size (Feagin 1992). Loss of respiration-related genes also occurred in those lineages in which the mitochondrion was converted to a mitochondrion-related organelle (e.g. hydrogenosome, mitosome), which retains very limited metabolic capacity. *Trichomonas vaginalis* has a relic mitochondrion called a

"hydrogenosome", which retains the incomplete TCA cycle and electron transport chain (Bui et al. 1996, Lindmark and Müller 1973). Another mitochondrion-derived organelle called a "mitosome" has very limited metabolic capacity, and a genome that is even more reduced than hydrogenomes. These mitosomes have been found in species such as *Entamoeba histolytica* (Clark and Roger 1995), *Mastigamoeba balamuthi* (Gill et al. 2007), *Encephalitozoon cuniculi* (Goldberg et al. 2008).

Overall, our results did not show any evidence of genome reduction for the endosymbiotic green algae, in contrast to the reduction in genome content observed for endosymbiotic bacteria and other endosymbiotic algae. It is possible that the degenerated organellar genomes will only occur once the endosymbiotic relationship becomes fully obligatory for both endosymbiont and host, or if the relationship becomes parasitic rather than symbiotic, as observed in parasitic land plants and parasitic algae. Thus, further investigation with broader and deeper sampling of organisms with various lifestyles will be necessary to better address this question.

Intron variation in organelle genome

In the mitochondrial genome of Trebouxiophyceae species, two protein-coding genes (*cob* and *cox1*), the *rrnL* ribosomal RNA gene and three tRNAs were split by intron(s). The intron located in *cob* gene is unique to *C. heliozoae* and was not detected in all the other trebouxiophytes, assuming that this intron propagated only in the specific lineage. Interestingly, the *cox1* gene was split into two exons by the group I intron in *C. heliozoae* and *M. conductrix*, whereas no intron was detected in *C.* sp. ATCC30562 and *C. sorokiniana*. Previous studies have demonstrated that the *cox1* intron was found in *C.* sp.

ArM0029B, *Chlorella vulgaris*, *Helicosporidium* sp. and *P. wickerhamii* but not in *C.* sp. C-169 and *Trebouxiophyceae* sp. (Jeong et al. 2014, Pombert and Keeling 2010, Servin-Garciduenas and Martinez-Romero 2012, Smith et al. 2011, Wolff et al. 1994). The *rrnL* introns are also sporadically distributed among algae. All the selected trebouxiophyceans except *C.* sp. ArM0029B have intron(s) in *rrnL*, with their intron numbers varying among species (1-7 introns) (Table 2-2). Notably, the endosymbiotic lineages generally retain more introns than the free-living taxa in *rrnL*. On the contrary, the introns in tRNAs are distributed only in the two free-living species *C.* sp. C-169 and *Trebouxiophyceae* sp. (Table 2-2). The sporadic distribution of these introns suggest that some green algal species have acquired these introns horizontally, which has been observed for several introns in other plant lineages (Sanchez-Puerta et al. 2008). If horizontal transfer is the explanation for the distribution of introns within Chlorellales, it raises the question of the intron donors: are they other algal species, other genomes within the algae, the endosymbiotic hosts or some other evolutionarily distant species? On the other hand, we cannot completely eliminate the possibility that these introns were gained at the common ancestor of the Chlorellales and then lost multiple times in the specific lineages. Thus, it is still unclear how introns spread in green algae and comprehensive analysis of intron distribution among green algae may shed light on this mystery.

In the plastid genome, the three endosymbiotic species tend to be intron rich, possessing 10 introns in *M. conductrix*, three in *C.* sp. ATCC30562 and two in *C. heliozoae*. The presence or absence of introns in Trebouxiophyceae plastomes appears to be sporadic and diverse. Group I introns were supposed to be found in diverse lineages and at extremely unbalanced frequencies (Haugen et al. 2005), so our results consistent with this

statement. Moreover, *C. heliozoae* and *M. conductrix* both have *rrnL* introns in the mitogenome and plastome, but the introns do not show any similarity between the two organelles. This suggests that there was no horizontal transfer of the *rrnL* introns between organelles. Further phylogenetic and structural analyses with increased sampling of new algal mitogenomes are needed to assess whether these introns were acquired horizontally.

Endosymbionts evolve from different origins

The endosymbiotic lifestyle has evolved multiple times in green algae, as evidenced by the multiple independent clades of endosymbiont algae in a phylogenetic tree based on the ITS2 or 18S rRNA sequences (Hoshina et al. 2006). Even within the *Chlorella* clade, endosymbionts did not cluster together in molecular phylogenies, suggesting that this lifestyle evolved multiple times (Hoshina and Imamura 2008, Krienitz et al. 2004, Summerer et al. 2008). However, the main limitation in these previous studies was the very small datasets used to construct the tree, such as the dataset containing only 1687 bp of 18S rRNA sequences. In our study, phylogenies inferred from the plastid (74 genes) and mitochondrial (32 genes) dataset revealed that the recent sequence-generated endosymbionts among *Chlorella* species are polyphyletic. Two paramecian endosymbionts *C.* sp. ATCC30562 and *M. conductrix* did not cluster together, but instead each of them were nested with another free-living organism, providing evidence for multiple origins of the endosymbiotic lifestyle. Overall, the present results combined with previous studies (Hoshina and Imamura 2008, Summerer et al. 2008) strongly suggest that the endosymbiotic lifestyle arose multiple times in Chlorellales. Similarly, some algal symbionts in lichens were found to be closely related, yet these lichen

endosymbioses still appear to have independent origins (Piercey-Normore and Depriest 2001, Zoller and Lutzoni 2003).

Taxonomic treatment of Micractinium

The *M. conductrix* species used in our study is an algal endosymbiont of the ciliate *P. bursaria*, but the classification and identification of the *P. bursaria* endosymbionts are still unclear or controversial (Hoshina and Imamura 2008). Based on the phylogenetic analyses inferred from almost all mitochondrial and plastid genes, our study has indicated that *M. conductrix* clusters within the group of Chlorellales, grouping specifically as the sister taxon to *C.* sp. ArM0029B. If this is the true classification, why was *M. conductrix* given a separate genus name and not named a species of *Chlorella*? Initially, the genus name "*Micractinium*" was given to those hydra symbionts which show the genetic distinct features from the known algae (Hoshina 2011). Then, there was an argument between "*Micractinium resseri*" and "*Micractinium conductrix*" because the *Micractinium reisseri* is the *P. bursaria* endosymbiotic alga while the name "*conductrix*" originally referred to the endosymbionts of *Hydra*. Interestingly, the species used in our study was isolated from *P. bursaria* as well.

To classify the species into genus *Chlorella*, it should meet the requirements in three criteria: morphology, cytology and molecular phylogeny. Hoshina *et al.*(2010) has demonstrated that the *M. reisseri* fulfill all of these requirements. Therefore, it is possible that the so called "*M. conductrix*" is the "*M. reisseri*" which may nest together with true *Chlorella* species. However, previous study has differentiated *Micractinium* and *Chlorella* by their ITS sequences (Luo et al. 2010).

We speculated that the *M. conductrix* species used in our study should be a *Chlorella* species mainly due to the robust organellar genomic phylogenies. Because of the difficulties and limitations of algal identification based on the morphological and ultrastructural characters, our results suggest that organellar genome phylogenies may be the preferred method for classification (Hoshina et al. 2004, Mattox 1984, Turmel et al. 2009). Similarly, the newly described chloroplast genomes in Trebouxiophyceae and Pedinophyceae have provided valuable insights into the relationships within the Trebouxiophyceae (Lemieux et al. 2014). However, our speculation was limited by the single sample from the genus *Micractinium*. To more fully address the issues discussed above, broad sequencing of organellar genomes from multiple *Micractinium* taxa may help to identify their taxonomic position.

ACKNOWLEDGMENTS

We are grateful to Dr. Joseph Msanne for providing *Chlorella* RNA extraction protocol, and to James Gurnon and Eric Noel for growing and preparing algae culture.

REFERENCES

Anwaruzzaman M, Chin BL, Li XP, Lohr M, Martinez DA, Niyogi KK. 2004. Genomic analysis of mutants affecting xanthophyll biosynthesis and regulation of photosynthetic light harvesting in Chlamydomonas reinhardtii. Photosynth Res 82: 265-276.

Arisue N, Hashimoto T, Mitsui H, Palacpac NM, Kaneko A, Kawai S, Hasegawa M, Tanabe K, Horii T. 2012. The Plasmodium apicoplast genome: conserved structure and close relationship of P. ovale to rodent malaria parasites. Mol Biol Evol 29: 2095-2099.

Bui E, Bradley PJ, Johnson PJ. 1996. A common evolutionary origin for mitochondria and hydrogenosomes. Proc. Natl. Acad. Sci. USA 93: 9651-9656.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17: 540-552.

Clark CG, Roger AJ. 1995. Direct evidence for secondary loss of mitochondria in Entamoeba histolytica. Proc. Natl. Acad. Sci. USA 92: 6518-6521.

Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PloS one 5: e11147.

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

Del Campo EM, Casano LM, Gasulla F, Barreno E. 2010. Suitability of chloroplast LSU rDNA and its diverse group I introns for species recognition and phylogenetic analyses of lichen-forming Trebouxia algae. Mol Phylogenet Evol 54: 437-444.

Donaher N, Tanifuji G, Onodera NT, Malfatti SA, Chain PS, Hara Y, Archibald JM. 2009. The complete plastid genome sequence of the secondarily nonphotosynthetic alga Cryptomonas paramecium: reduction, compaction, and accelerated evolutionary rate. Genome Biol Evol 1: 439-448.

Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu X, Reith M, Cavalier-Smith T, Maier UG. 2001. The highly reduced genome of an enslaved algal nucleus. Nature 410: 1091-1096.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792-1797.

Fabret C, Dervyn E, Dalmais B, Guillot A, Marck C, Grosjean H, Noirot P. 2011. Life without the essential bacterial tRNAIle2–lysidine synthetase TilS: a case of tRNA gene recruitment in Bacillus subtilis. Mol Microbiol 80: 1062-1074.

Fan J, Lee RW. 2002. Mitochondrial genome of the colorless green alga Polytomella parva: two linear DNA molecules with homologous inverted repeat Termini. Mol Biol Evol 19: 999-1007.

Feagin JE. 1992. The 6-kb element of Plasmodium falciparum encodes mitochondrial cytochrome genes. Mol Biochem Parasitol 52: 145-148.

Friedl T, Rokitta C. 1997. Species relationships in the lichen alga Trebouxia (Chlorophyta, Trebouxiophyceae): molecular phylogenetic analyses of nuclear-encoded large subunit rRNA gene sequences. Symbiosis, Philadelphia, Pa.(USA).

Gill EE, Diaz-Triviño S, Barberà MJ, Silberman JD, Stechmann A, Gaston D, Tamas I, Roger AJ. 2007. Novel mitochondrion-related organelles in the anaerobic amoeba Mastigamoeba balamuthi. Mol Microbiol 66: 1306-1320.

Gilson PR, Su V, Slamovits CH, Reith ME, Keeling PJ, McFadden GI. 2006. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. Proc. Natl. Acad. Sci. USA 103: 9566-9571.

Goldberg AV, Molik S, Tsaousis AD, Neumann K, Kuhnke G, Delbac F, Vivares CP, Hirt RP, Lill R, Embley TM. 2008. Localization and functionality of microsporidian iron-sulphur cluster assembly proteins. Nature 452: 624-628.

Gray MW, Boer PH. 1988. Organization and expression of algal (Chlamydomonas reinhardtii) mitochondrial DNA. Philos Trans R Soc Lond B Biol Sci 319: 135-147.

Grewe F, Edger PP, Keren I, Sultan L, Pires JC, Ostersetzer-Biran O, Mower JP. 2014. Comparative analysis of 11 Brassicales mitochondrial genomes and the mitochondrial transcriptome of Brassica oleracea. Mitochondrion 19 Pt B: 135-143.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59: 307-321.

Haugen P, Simon DM, Bhattacharya D. 2005. The natural history of group I introns. Trends Genet. 21: 111-119.

Hepburn NJ, Schmidt DW, Mower JP. 2012. Loss of two introns from the Magnolia tripetala mitochondrial cox2 gene implicates horizontal gene transfer and gene conversion as a novel mechanism of intron loss. Mol Biol Evol 29: 3111-3120.

Hoshina. 2011. Comments on the taxonomic treatment of Micractinium reisseri (Chlorellaceae, Trebouxiophyceae), a common endosymbiont in Paramecium. Phycological Res. 59: 269-272.

Hoshina, Imamura N. 2008. Multiple Origins of the Symbioses in Paramecium bursaria. Protist 159: 53-63.

Hoshina, Kamako S-I, Imamura N. 2004. Phylogenetic position of endosymbiotic green algae in Paramecium bursaria Ehrenberg from Japan. Plant Biol. 6: 447-453.

Hoshina, Hayashi S, Imamura N. 2006. Intraspecific genetic divergence of Paramecium bursaria and re-construction of the paramecian phylogenetic tree. Acta Protozool 45: 377-386.

Hoshina R, Iwataki M, Imamura N. 2010. Chlorella variabilis and Micractinium reisseri sp. nov.(Chlorellaceae, Trebouxiophyceae): Redescription of the endosymbiotic green algae of Paramecium bursaria (Peniculia, Oligohymenophorea) in the 120th year. Phycological Res. 58: 188-201.

Husnik F, et al. 2013. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. Cell 153: 1567-1578.

Jeong H, Lim JM, Park J, Sim YM, Choi HG, Lee J, Jeong WJ. 2014. Plastid and mitochondrion genomic sequences from Arctic Chlorella sp. ArM0029B. BMC genomics 15: 286.

Kodama Y, Suzuki H, Dohra H, Sugii M, Kitazume T, Yamaguchi K, Shigenobu S, Fujishima M. 2014. Comparison of gene expression of Paramecium bursaria with and without Chlorella variabilis symbionts. BMC Genomics 15: 183.

Krienitz L, Hegewald EH, Hepperle D, Huss VA, Rohr T, Wolf M. 2004. Phylogenetic relationship of Chlorella and Parachlorella gen. nov.(Chlorophyta, Trebouxiophyceae). Phycologia 43: 529-542.

Kuck P, Meusemann K. 2010. FASconCAT: Convenient handling of data matrices. Mol Phylogenet Evol 56: 1115-1118.

Lane CE, van den Heuvel K, Kozera C, Curtis BA, Parsons BJ, Bowman S, Archibald JM. 2007. Nucleomorph genome of Hemiselmis andersenii reveals complete intron loss and compaction as a driver of protein structure and function. Proc. Natl. Acad. Sci. USA 104: 19908-19913.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359.

Lemieux C, Otis C, Turmel M. 2014. Chloroplast phylogenomic analysis resolves deep-level relationships within the green algal class Trebouxiophyceae. BMC Evol Biol 14: 211.

Lin H, Kwan AL, Dutcher SK. 2010. Synthesizing and salvaging NAD: lessons learned from Chlamydomonas reinhardtii. PLoS Genet 6: e1001105.

Lindmark DG, Müller M. 1973. Hydrogenosome, a cytoplasmic organelle of the anaerobic flagellate Tritrichomonas foetus, and its role in pyruvate metabolism. J. Biol. Chem 248: 7724-7728.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25: 955-964.

Luo W, Proschold T, Bock C, Krienitz L. 2010. Generic concept in Chlorella-related coccoid green algae (Chlorophyta, Trebouxiophyceae). Plant Biol (Stuttg) 12: 545-553.

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.

Mattox K. 1984. Classification of the green algae: a concept based on comparative cytology. Systematics of the green algae: 29-72.

McCutcheon JP, McDonald BR, Moran NA. 2009. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. Proc. Natl. Acad. Sci. USA 106: 15394-15399.

Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, Hattori M. 2006. The 160-kilobase genome of the bacterial endosymbiont Carsonella. Science 314: 267.

Perez-Ortega S, Rios Ade L, Crespo A, Sancho LG. 2010. Symbiotic lifestyle and phylogenetic relationships of the bionts of Mastodia tessellata (Ascomycota, incertae sedis). Am J Bot 97: 738-752.

Petersen G, Cuenca A, Moller IM, Seberg O. 2015. Massive gene loss in mistletoe (Viscum, Viscaceae) mitochondria. Scientific Reports 5: 17588.

Piercey-Normore MD, Depriest PT. 2001. Algal switching among lichen symbioses. Am J Bot 88: 1490-1498.

Pombert, Keeling PJ. 2010. The mitochondrial genome of the entomoparasitic green alga helicosporidium. PloS one 5: e8954.

Pombert, Blouin NA, Lane C, Boucias D, Keeling PJ. 2014. A Lack of Parasitic Reduction in the Obligate Parasitic Green Alga Helicosporidium. PLoS genetics 10: e1004355.

Proschold T, Darienko T, Silva PC, Reisser W, Krienitz L. 2011. The systematics of Zoochlorella revisited employing an integrative approach. Environ Microbiol 13: 350-364.

Reisser W. 1992. Algae and symbioses: Biopress Limited.

Reisser W. 1994. Enigmatic chlorophycean algae forming symbiotic associations with ciliates. Pages 87-95. Evolutionary Pathways and Enigmatic Algae: Cyanidium caldarium (Rhodophyta) and Related Cells, Springer.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61: 539-542.

Sanchez-Puerta MV, Cho Y, Mower JP, Alverson AJ, Palmer JD. 2008. Frequent, phylogenetically local horizontal transfer of the cox1 group I intron in flowering plant mitochondria. Mol Biol Evol 25: 1762-1777.

Servin-Garciduenas LE, Martinez-Romero E. 2012. Complete mitochondrial and plastid genomes of the green microalga Trebouxiophyceae sp. strain MX-AZ01 isolated from a highly acidic geothermal lake. Eukaryot Cell 11: 1417-1418.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol 51: 492-508.

Skippington E, Barkman TJ, Rice DW, Palmer JD. 2015. Miniaturized mitogenome of the parasitic plant Viscum scurruloideum is extremely divergent and dynamic and has lost all nad genes. Proc. Natl. Acad. Sci. USA 112: E3515-E3524.

Sloan DB, Nakabachi A, Richards S, Qu J, Murali SC, Gibbs RA, Moran NA. 2014. Parallel histories of horizontal gene transfer facilitated extreme reduction of endosymbiont genomes in sap-feeding insects. Mol Biol Evol 31: 857-871.

Smith DR, Hua J, Archibald JM, Lee RW. 2013. Palindromic genes in the linear mitochondrial genome of the nonphotosynthetic green alga Polytomella magna. Genome Biol Evol 5: 1661-1667.

Smith DR, Burki F, Yamada T, Grimwood J, Grigoriev IV, Van Etten JL, Keeling PJ. 2011. The GC-rich mitochondrial and plastid genomes of the green alga Coccomyxa give insight into the evolution of organelle DNA nucleotide landscape. PLoS One 6: e23624.

Summerer M, Sonntag B, Sommaruga R. 2008. Ciliate-symbiont specificity of freshwater endosymbiotic Chlorella (Trebouxiophyceae, Chlorophyta) 1. J. Phycol 44: 77-84.

Turmel M, Otis C, Lemieux C. 2009. The chloroplast genomes of the green algae Pedinomonas minor, Parachlorella kessleri, and Oocystis solitaria reveal a shared ancestry between the Pedinomonadales and Chlorellales. Mol Biol Evol 26: 2317-2331.

Ueno R, Urano N, Suzuki M. 2003. Phylogeny of the non-photosynthetic green micro-algal genus Prototheca (Trebouxiophyceae, Chlorophyta) and related taxa inferred from SSU and LSU ribosomal DNA partial sequence data. FEMS microbiology letters 223: 275-280.

Wolf PG. 2012. Plastid genome diversity. Pages 145-154. Plant Genome Diversity Volume 1, Springer.

Wolff G, Plante I, Lang BF, Kuck U, Burger G. 1994. Complete sequence of the mitochondrial DNA of the chlorophyte alga Prototheca wickerhamii. Gene content and genome organization. J Mol Biol 237: 75-86.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20: 3252-3255.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821-829.

Zhu A, Guo W, Jain K, Mower JP. 2014. Unprecedented heterogeneity in the synonymous substitution rate within a plant genome. Mol Biol Evol 31: 1228-1236.

Zoller S, Lutzoni F. 2003. Slow algae, fast fungi: exceptionally high nucleotide substitution rate differences between lichenized fungi Omphalina and their symbiotic green algae Coccomyxa. Mol Phylogenet Evol 29: 629-640.

TABLES

Table 2-1. General features comparison among selected Trebouxiophyceae

	Mitochondrial Genome						Plastid Genome					
	Size (bp)	A+T (%)	Protein-coding	tRNAs	rRNAs	Introns	Size (bp)	A+T (%)	Protein-coding	tRNAs	rRNAs	Introns
Chlorella ArM0029B	65049	71.5	32	27	3	1	119989	66.1	79	32	3	1
Chlorella heliozoae	62477	68.3	32	27	3	7	124353	64.7	79	31	3	2
Chlorella sp. ATCC30562	79601	71.9	32	27	3	7	124881	66.1	79	32	3	3
Chlorella sorokiniana	52528	70.9	32	27	3	1	109811	65.9	78	31	3	2
Chlorella variabilis	78500	71.8	32	27	3	6	124579	66.1	79	32	3	3
Coccomyxa sp. C-169	65497	46.9	30	26	3	5	175731	49.3	79	33	3	1
Helicosporidium sp.	49343	74.4	32	25	3	4	37454	73.1	26	25	3	1
Micractinium conductrix	74708	70.6	32	27	3	5	129436	65.2	78	32	3	10
Prototheca wickerhamii	55328	74.2	30	26	3	5	55636	68.8	40	28	3	1
Trebouxiophyceae sp.	74423	46.6	30	23	3	11	149707	43.8	79	33	3	5

Table 2-2. Comparison of mitochondrial and plastid genome intron content

Table 2. Comparison of mitochondrial and plastid genome intron content

	Mitogenome							Plastome										
	cob	*cox1*	*rrnL*	*trnH*	*trnS**	*trnW*	Total	*chlL*	*ftsh*	*petB*	*psaC*	*psbA*	*psbB*	*psbC*	*psbD*	*rps12*	*rrnL*	*trnL*	Total
Chlorella ArM0029B	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1
Chlorella heliozoae	1	2	4	0	0	0	7	0	0	0	0	0	0	1	0	1	0	2
Chlorella sp. ATCC30562	0	0	7	0	0	0	7	0	0	0	0	1	0	1	0	0	1	3
Chlorella sorokiniana	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	1	2
Chlorella variabilis	0	0	6	0	0	0	6	0	0	0	0	1	0	1	0	0	1	3
Coccomyxa sp. C-169	0	0	1	1	2	1	5	0	0	0	0	0	1	0	0	0	0	1
Helicosporidium sp.	0	2	2	0	0	0	4	×	0	0	0	0	0	0	0	0	1	1
Micractinium conductrix	0	2	3	0	0	0	5	0	0	1	1	0	1	0	1	1	4	1	10
Prototheca wickerhamii	0	3	2	0	0	0	5	×	0	0	0	0	0	0	0	0	0	1	1
Trebouxiophyceae sp.	0	1	6	1	2	1	11	0	1	0	0	1	0	0	0	0	3	0	5

**two copies of this gene each has one intron*

× Gene loss

FIGURES

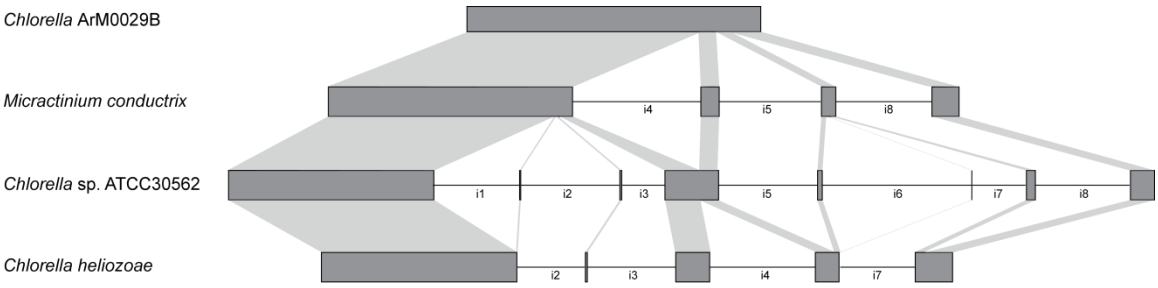

Figure 2-1. Comparison of mitogenome *rrnL* intron content. The exons are marked by gray rectangular boxes, connecting by the intron regions (black lines). Introns are ordered by name labelling under lines. Homologous exon regions are highlighted by gray shadow. Species name are shown on the left and the maps are drawn approximately to scale.

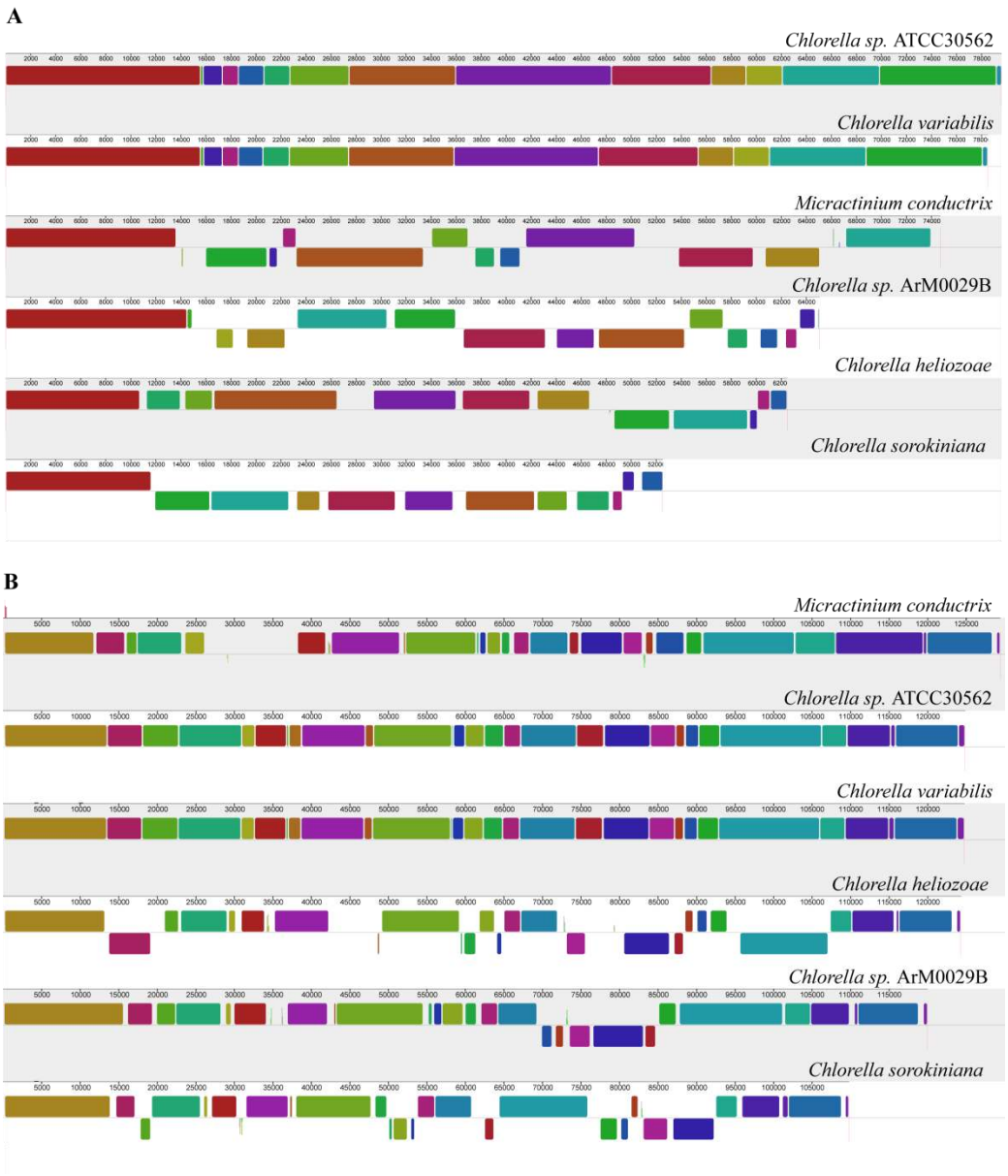

Figure 2-2. Selected green algae organellar genome synteny. (A) Mitochondrial genome synteny. (B) Plastid genome synteny. Images were generated using Mauve with default settings (Darling 2004). Color-coded syntenic blocks indicate conserved segments identified by Mauve. Plots of sequences similarity are shown within each syntenic block. Regions with no color indicate no detectable homology between the two genomes with the settings used in Mauve. The species names were labeled on the right.

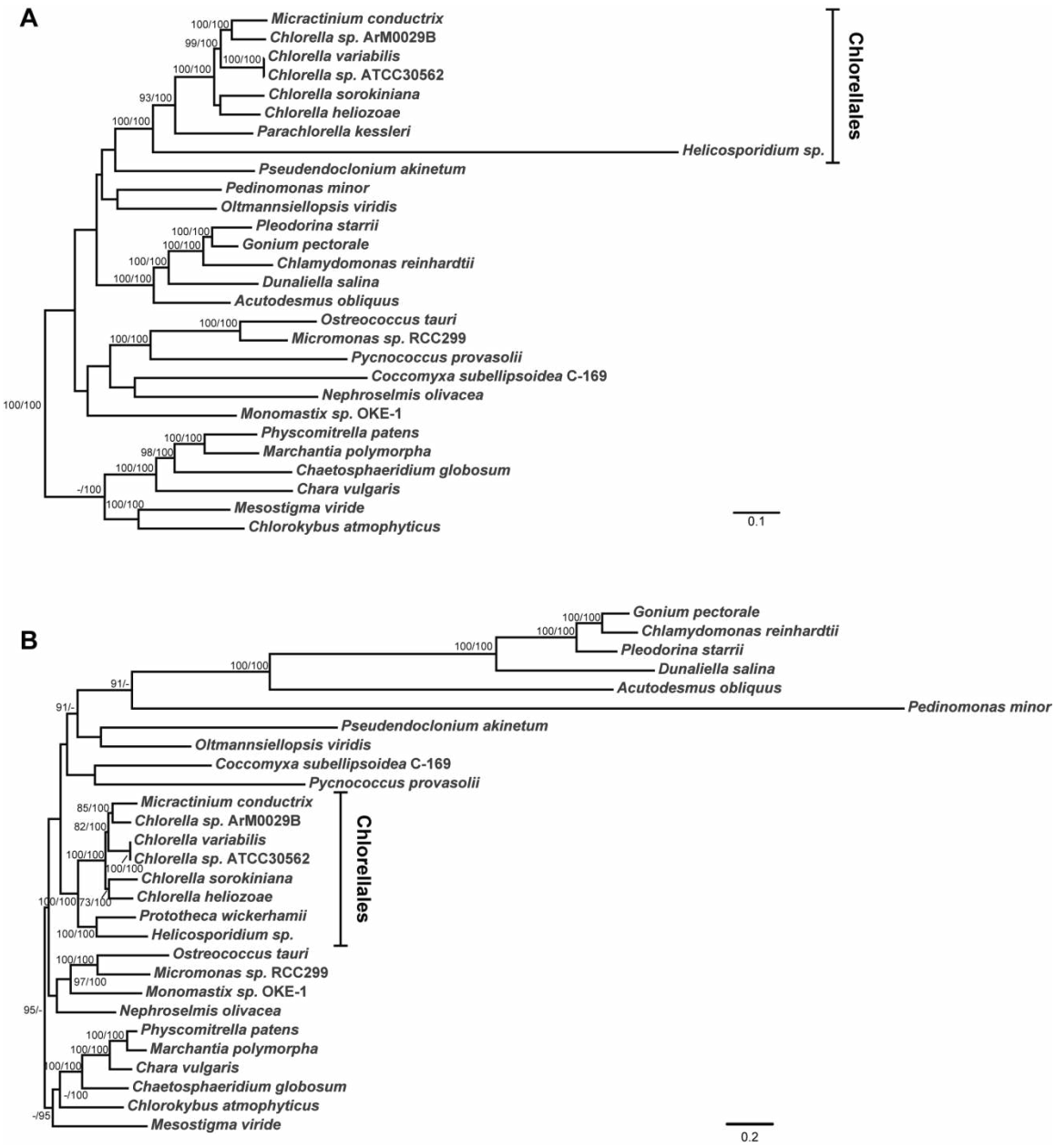

Figure 2-3. Phylogenetic analysis of selected trebouxiophytes by (A) 74 plastid genes and (B) 32 mitochondrial genes. The trees shown were generated by maximum likelihood (left) and Bayesian (right) inference indicated at each node. Weak support values (<50%) were eliminated from the figure. Trees were rooted on streptophytes. The scale bars were shown at the bottom right for each tree.

SUPPORTING INFORMATION

Table S1. GenBank accession numbers for taxa used in analysis

Class	Family	Species	Plastome Acc. No.	Mitogenome Acc. No.
Bryopsida	Funariaceae	*Physcomitrella patens*	AP005672	NC_007945
Charophyceae	Characeae	*Chara vulgaris*	NC_008097	NC_005255
Chlorokybophyceae	Chlorokybaceae	*Chlorokybus atmophyticus*	NC_008822	NC_009630
Chlorophyceae	Chlamydomonadaceae	*Chlamydomonas reinhardtii*	NC_005353	NC_001638
Chlorophyceae	Dunaliellaceae	*Dunaliella salina*	NC_016732	NC_012930
Chlorophyceae	Volvocaceae	*Gonium pectorale*	NC_020438	NC_020437
Chlorophyceae	Volvocaceae	*Pleodorina starrii*	NC_021109	NC_021108
Chlorophyceae	Scenedesmaceae	*Scenedesmus obliquus*	NC_008101	NC_002254
Coleochaetophyceae	Chaetosphaeridiaceae	*Chaetosphaeridium globosum*	NC_004115	NC_004118
Mamiellophyceae	Mamiellaceae	*Micromonas sp.* RCC299	NC_012575	NC_012643
Mamiellophyceae	Bathycoccaceae	*Ostreococcus tauri*	NC_008289	NC_008290
Mamiellophyceae	Monomastigaceae	*Monomastix sp.* OKE-1	NC_012101	NC_022797
Marchantiopsida	Marchantiaceae	*Marchantia polymorpha*	NC_001319	NC_001660
Mesostigmatophyceae	Mesostigmataceae	*Mesostigma viride*	NC_002186	NC_008240
Nephroselmidophyceae	Pycnococcaceae	*Nephroselmis olivacea*	NC_000927	NC_008239
Pedinophyceae	Pedinomonadaceae	*Pedinomonas minor*	NC_016733	NC_000892
Prasinophyceae	Pycnococcaceae	*Pycnococcus provasolii*	NC_012097	NC_013935
Trebouxiophyceae	Chlorellaceae	*Chlorella sp.* ArM0029B	KF554427	KF554428
Trebouxiophyceae	Chlorellaceae	*Chlorella sorokiniana*	NC_023835	KM241869
Trebouxiophyceae	Chlorellaceae	*Chlorella variabilis*	NC_015359	KP271968
Trebouxiophyceae	Coccomyxaceae	*Coccomyxa sp* . C-169	NC_015084	NC_015316
Trebouxiophyceae	Chlorellales incertae sedis	*Helicosporidium sp.*	NC_008100	NC_017841
Trebouxiophyceae	Chlorellaceae	*Parachlorella kessleri*	NC_012978	
Trebouxiophyceae	Chlorellaceae	*Prototheca wickerhamii*	KJ001761	NC_001613
Trebouxiophyceae	unclassified Trebouxiophyceae	*Trebouxiophyceae sp.*	NC_018569	NC_018568
Ulvophyceae	Kornmanniaceae	*Pseudendoclonium akinetum*	NC_008114	NC_005926
Ulvophyceae	Oltmannsiellopsidaceae	*Oltmannsiellopsis viridis*	NC_008099	NC_008256

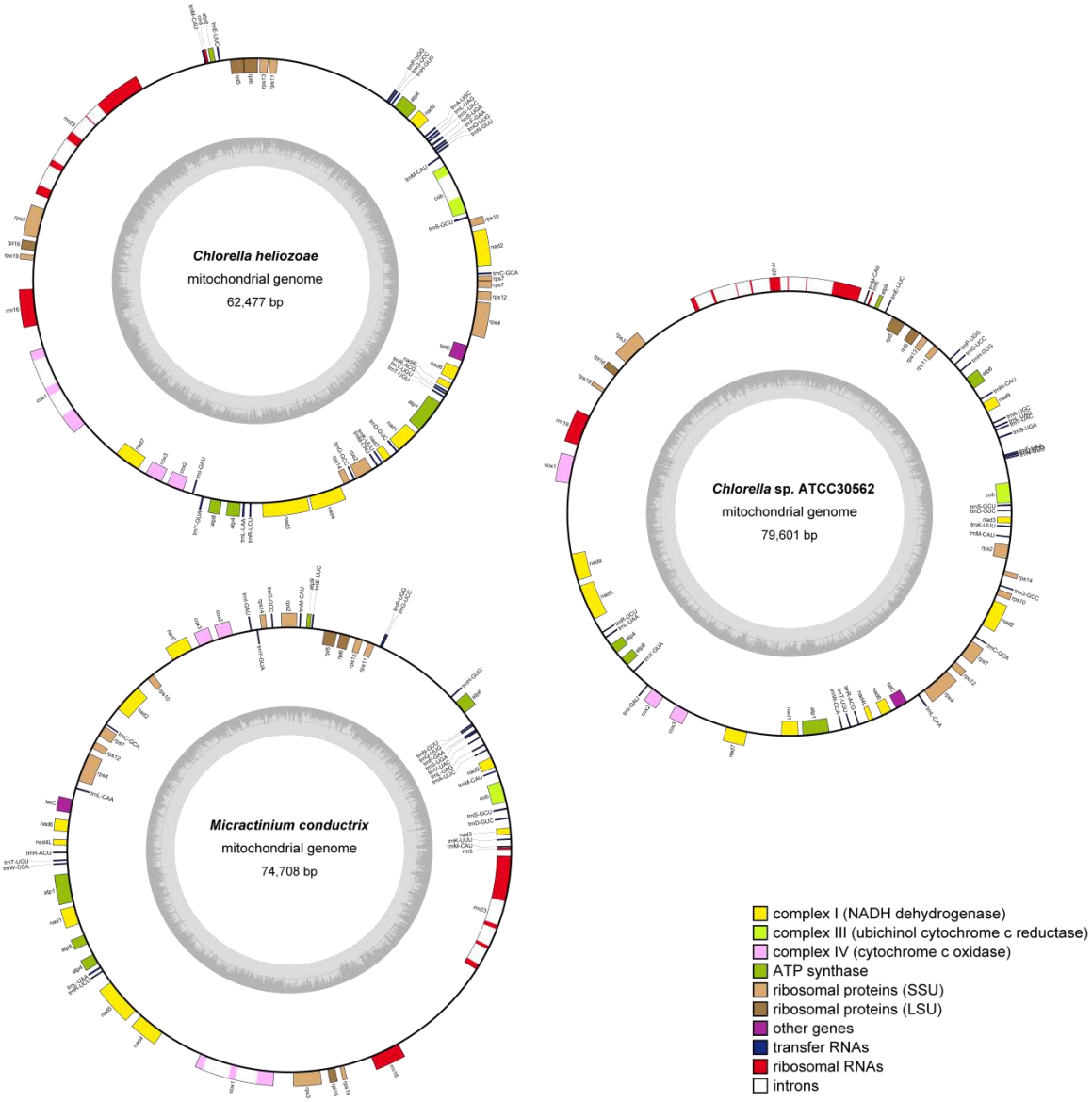

Figure S1. Mitogenome maps of the *C. heliozoae*, *C.* sp. ATCC30562 and *M. conductrix*. Outer genes are transcribed counter-clockwise; inner genes are transcribed clockwise. Gene and intron colors correspond to the functional categories listed in the key at the bottom right. GC content is shown on the inner circle by dark grey bars. The map was drawn with OgDraw (http://ogdraw.mpimp-golm.mpg.de/).

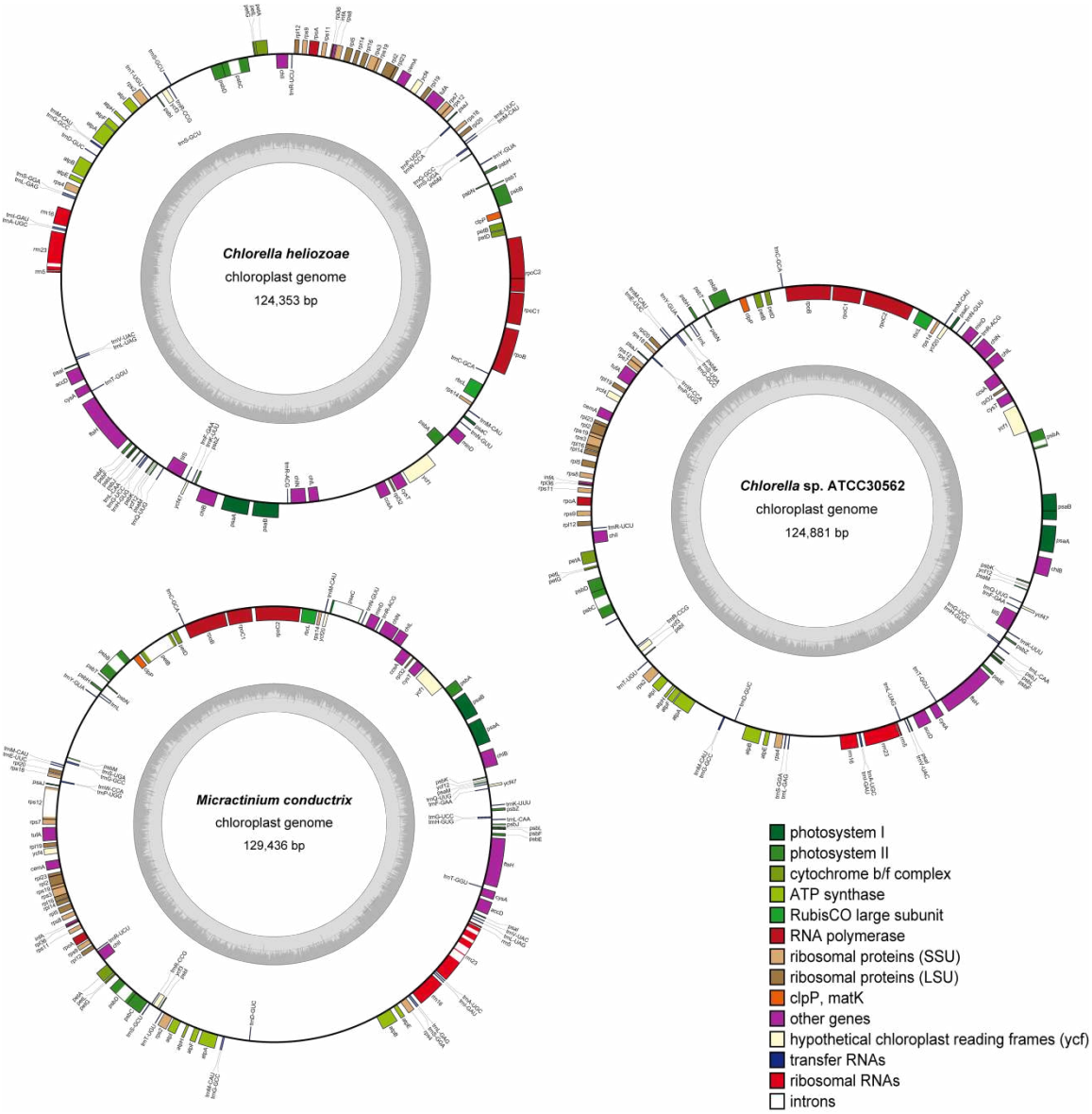

Figure S2. Plastid genome maps of the *C. heliozoae*, *C.* sp. ATCC30562 and *M. conductrix*. Outer genes are transcribed counter-clockwise; inner genes are transcribed clockwise. Gene and intron colors correspond to the functional categories listed in the key at the bottom right. GC content is shown on the inner circle by dark grey bars. The map was drawn with OgDraw (http://ogdraw.mpimp-golm.mpg.de/).

CHAPTER 3

Evaluation of Genetic Degradation and Horizontal Transfer in Mitochondrial Genomes from Hemiparasites and Holoparasites in Orobanchaceae

Weishu Fan, Andan Zhu, Melisa Kozaczek, Neethu Shah, Natalia Pabón-Mora, Favio González and Jeffrey P. Mower

ABSTRACT

In parasitic plants, the reduction in plastid genome (plastome) size and content is driven predominantly by the loss of photosynthetic genes. The first completed mitochondrial genomes (mitogenomes) from parasitic mistletoes also exhibit significant degradation, but the generality of this observation for other parasitic plants is unknown. We sequenced the complete mitogenome and plastome of the hemiparasite *Castilleja paramensis* (Orobanchaceae) and compared them with additional holoparasitic, hemiparasitic and nonparasitic species from Orobanchaceae. Comparative mitogenomic analysis revealed minimal gene loss among the seven Orobanchaceae species, indicating the retention of typical mitochondrial function among Orobanchaceae species. Phylogenetic analysis provided evidence for horizontal transfer of six genes, lending further support that the parasite lifestyle facilitates transfer among species. However, the mobile *cox1* intron was acquired vertically from a nonparasitic ancestor, arguing against a role for plant parasitism in the horizontal acquisition or distribution of this intron. The *C. paramensis* plastome has retained nearly all genes except for the recent pseudogenization of four subunits of the NAD(P)H dehydrogenase complex, indicating a very early stage of plastome degradation. These results lend support to the notion that loss of *ndh* gene function is the first step of plastome degradation in the transition to a parasitic lifestyle.

INTRODUCTION

One of the defining characteristics of plants is the presence of a plastid, which enables the fixation of carbon to produce organic molecules via photosynthesis. Parasitic plants represent a dramatic departure from the typical autotrophic lifestyle of plants because they obtain organic carbon sources heterotrophically, using specialized organs called haustoria to make direct connections with the vascular tissue in the roots or shoots of a host plant. Parasitic plants, which comprise approximately 1% of all angiosperms (Westwood et al. 2010), can be subdivided based on the extent of their reliance on heterotrophy: hemiparasites retain the ability to photosynthesize and obtain only some of their nutrients from their hosts, while holoparasites have lost photosynthetic ability and must obtain all of their nutrition from hosts. Mycoheterotrophic plants, which obtain nutrients from fungi associated with other plants (Merckx et al. 2009), do not utilize haustoria for obtaining nutrients are therefore distinct from the haustorial parasitic plants under consideration in this study.

The transition from an autotrophic to a heterotrophic lifestyle has had a dramatic impact on the plastid genomes (plastome) of parasitic plants. Studies of parasitic plant plastomes have established a wide range of genomic degradation, defined primarily by the presence or absence of photosynthetic activity. For example, the hemiparasite *Schwalbea americana* (Orobanchaceae) possesses a large 160 kb plastome with minimal pseudogenization/loss of only six *ndh* genes, which encode subunits of the plastid NAD(P)H dehydrogenase complex (Wicke et al. 2013). Plastomes of hemiparasitic mistletoes (Viscaceae) are slightly more degraded, exhibiting both a reduction in size

(down to 126–147 kb) and the loss of all 11 *ndh* genes plus a small number (1–6) of non-photosynthetic genes (Petersen et al. 2015a). Within *Cuscuta* (Convolvulaceae), the four sequenced plastomes range from 85 to 125 kb in size and have experienced more extensive gene loss, yet they still retain all (or all but one) photosynthetic genes (Funk et al. 2007, McNeal J. R. et al. 2007), which is consistent with at least low levels of photosynthetic activity (van der Kooij et al. 2000). Other *Cuscuta* species are clearly non-photosynthetic and their plastomes have lost numerous photosynthetic and non-photosynthetic genes (Braukmann et al. 2013, van der Kooij et al. 2000). Plastomes in the holoparasitic species of Orobanchaceae are also heavily degraded (Cusimano and Wicke 2016, Li et al. 2013, Wicke et al. 2013, Wolfe et al. 1992), most extensively in *Conopholis americana* whose plastome is only 46 kb in size with just 21 intact protein-coding genes. Even greater genomic reduction was reported in *Pilostyles* (Apodanthaceae), whose plastomes are reduced to just 11–15 kb and may contain only five or six functional genes(Bellot and Renner 2016). In some holoparasites, such as *Rafflesia lagascae* (Rafflesiaceae), the entire plastome may have been lost (Molina et al. 2014).

Much less is known about the effects of a parasitic lifestyle on the mitochondrial genomes (mitogenomes) of plants. In fact, only a single genus of parasitic plants has a completely sequenced mitogenome, from the hemiparasitic mistletoes *Viscum scurruloideum* and *Viscum album*, along with draft genomes from two additional *Viscum* species (Petersen et al. 2015b, Skippington et al. 2015). Compared with other land plants, *V. scurruloideum* has the smallest mitogenome (66kb) and all four *Viscum* sequences have lost functional copies of all nine *nad* genes encoding subunits of the mitochondrial

NADH dehydrogenase complex I, the first reported loss of this complex from any multicellular eukaryote (Skippington et al. 2015). In contrast, the draft mitogenome from the holoparasite *R. lagascae* has a typical size (estimated at >300 kb) for an angiosperm and contains a nearly complete set of protein-coding genes, including at least seven of nine *nad* genes (Molina et al. 2014).

Despite the limited mitogenomic information for parasitic plants, it is well established that their mitochondrial DNA undergoes frequent horizontal transfer, which is likely facilitated by the direct physical connection between parasitic and host plants (Davis and Xi 2015, Mower Jeffrey P et al. 2012a, Sanchez-Puerta Maria Virginia 2014). Perhaps the best studied example of plant horizontal transfer involves the mobile group I intron of the cytochrome oxidase subunit 1 (*cox1*) gene. This intron was originally acquired from fungi and has been subsequently transferred many times during angiosperm evolution (Cho et al. 1998, Sanchez-Puerta M. V. et al. 2008, Vaughn et al. 1995). Intriguingly, this *cox1* intron is highly overrepresented in the parasitic plants that have been examined to date, suggesting that parasitic plants may serve as mediators of horizontal intron transfer among angiosperms (Barkman et al. 2007). Although this hypothesis was not supported in an analysis with limited sampling of parasitic plants (Barkman et al. 2007), denser sampling of parasites and closely related nonparasitic taxa is needed before the hypothesis should be rejected.

The Orobanchaceae is an ideal family for studies on parasitic plant evolution because it contains the full range of trophic specialization, including a nonparasitic lineage (*Lindenbergia*), numerous hemiparasitic lineages with varying degrees of photosynthetic activity (e.g., *Bartsia*, *Castilleja*, *Schwalbea*, *Striga*), and at least three transitions to

holoparasitism (e.g., *Lathraea*, *Orobanche*, *Hyobanche*) resulting in a complete loss of photosynthesis (Bennett and Mathews 2006, McNeal Joel R et al. 2013, Young et al. 1999). Complete plastome sequences are available from 12 species in Orobanchaceae, but only from a single hemiparasite (*S. americana*), while data from the mitogenome in this family is lacking. To improve our understanding of organellar genomic evolution in hemiparasitic plants, we sequenced the complete mitochondrial and plastid genomes from the hemiparasite *Castilleja paramensis*. Furthermore, to assess mitogenomic diversity within the Orobanchaceae, we generated draft mitogenome sequences from six additional species representing the range of trophic diversity: the autotroph *Lindenbergia philippensis*, the hemiparisites *Bartsia pedicularioides* and *S. americana*, and the holoparasites *Orobanche crenata*, *Orobanche gracilis*, and *Phelipanche ramosa*. These sequences were compared to assess the degree of genomic degradation and the extent of horizontal gene transfer resulting from the parasitic lifestyle.

MATERIALS AND METHODS

Sample collection and organellar genome sequencing

A *C. paramensis* individual was collected from a páramo in the department of Boyacá, Colombia on March 21, 2014 (voucher *N. Pabón-Mora et al. 299*, HUA). A *B. pedicularioides* individual was collected from a páramo in Cajas National Park, Ecuador on December 17, 2010 (voucher *J. P. Mower et al. 2064*, QCA). Total genomic DNA was extracted from silica-dried leaves using the Plant DNeasy Kit (Qiagen). DNA samples were sequenced on the Illumina HiSeq2000 platform at BGI (Shenzhen, China),

which generated 6 Gb (for *B. pedicularioides*) or 8 Gb (for *C. paramensis*) of 100-bp paired-end reads from an 800-bp library.

Genome assembly and annotation

Draft organellar genomes of *C. paramensis* and *B. pedicularioides* were assembled from the Illumina sequence reads with Velvet version 1.2.03 (Zerbino and Birney 2008) using multiple combinations of kmer (61, 71, 81, 91) and expected coverage (50, 100, 200, 500, 1000) values, as described previously (Guo et al. 2014, Zhu et al. 2014). Organellar contigs were identified in each assembly by using default blastn searches with known organellar gene sequences from related Lamiales species as queries. For each targeted genome, the best assembly that maximized total mitochondrial or plastid length in the fewest number of contigs was used for further scaffolding. Scaffolding was performed by mapping read pairs onto the contig sequences using blastn (e-value $\leq 1x10^{-10}$, hit length ≥90 bp, sequence identity ≥90%), and read pairs spanning two different contigs were used to infer contig joins and repeat regions. Using this strategy, circular-mappingplastid and mitochondrial genomes were assembled for *C. paramensis*, and a draft mitogenome was assembled for *B. pedicularioides*. The *C. paramensis* and *B. pedicularioides* mitogenome assemblies were annotated as described previously (Guo et al. 2016, Mower et al. 2012b, Zhu et al. 2014). The *C. paramensis* plastid genome was annotated using DOGMA (Wyman et al. 2004) followed by manual adjustment as necessary.

To survey mitochondrial gene content in additional Orobanchaceae species, 454 pyrosequencing data from a previous study (Piednoel et al. 2012) were downloaded from the NCBI sequence read archive (accession SRA047928) for one hemiparasite (*S.*

americana), three holoparasites (*O. crenata*, *O. gracilis*, *P. ramosa*)and one nonparasite (*L. philippensis*). The downloaded 454 data were assembled with Velvet 1.2.03 as described above using various pairwise combinations of kmer (41, 51, 61, 71) and expected coverage (5, 10, 20, 50, 100) values. Lower kmer and expected coverage values were required for these data sets given the lower amount of data available (<1 Gb total genomic DNA for each species), resulting in assemblies with 5–10x depth of mitochondrial sequence coverage for each species. Scaffolding was not performed because the reads were unpaired. The presence of mitochondrial genes and introns was scored by using blastn searches with mitochondrial gene sequences from other Lamiales species as queries against the best 454 assemblies. Gene and intron sequences of interest were manually extracted from these 454 assemblies for further analysis.

Genes identified from each assembly were assessed for potential loss of function by searching for frameshifting indels and/or premature stop codons. Genes were scored as pseudogenes if the mutations disrupted >20% of their conserved domain structure, as defined by a search of the NCBI Conserved Domain Database (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi), or if >30% of the gene was disrupted overall.

Phylogenetic evaluation of horizontal transfer

Gene sequences were extracted from the Orobanchaceae mitochondrial assemblies in this study and from 42 additional seed plant mitogenomes available in GenBank (Table S1). These DNA sequences were translated using the standard genetic code and then aligned using MUSCLE version 3.8.31 (Edgar 2004). Individual protein alignments were reverse

translated into codon-based nucleotide alignments using PAL2NAL (Suyama et al. 2006), and each codon alignment was trimmed of poorly aligned regions with Gblocks 0.91b (Castresana 2000) using relaxed parameters (t=c, b2=half+1, b4=5, b5=half). Phylogenetic analyses of trimmed codon alignments were performed using maximum-likelihood in PhyML 3.0 (Guindon et al. 2010). A GTR+G+I model with four substitution rate categories was employed. Tree topologies, branch lengths, and rate parameters were optimized during the run. Branch support was calculated from 500 bootstrap replicates. The resulting phylogenetic trees were examined for strong phylogenetic incongruence (≥80% bootstrap support) involving one or more Orobanchaceae sequences.

Trees exhibiting strong incongruence of Orobanchaceae species in the initial survey were further evaluated to reduce the potentially artefactual effects of low taxon sampling and the presence of RNA editing in the gene sequences, both of which can negatively affect the accuracy of phylogenetic results (Bowe and dePamphilis 1996, Zwickl and Hillis 2002). To increase taxon sampling, additional gene sequences from phylogenetically diverse eudicot species were obtained from GenBank and added to the alignments (Table S1). To mitigate the effects of RNA editing, experimentally determined RNA sequences from *Arabidopsis thaliana* (Giegé and Brennicke 1999), *Oryza sativa* (Notsu et al. 2002), *Beta vulgaris* (Mower J. P. and Palmer 2006), *Citrullus lanatus*(Alverson et al. 2010), *Cycas taitungensis* (Salmans et al. 2010), *Liriodendron tulipifera* (Richardson et al. 2013), and *Amborella trichopoda*(Rice et al. 2013) were used to predict edit sites in the alignments using the PREP-Aln server (Mower J. P. 2009) with a cutoff value of 0.2. The predicted RNA sequences were aligned and trimmed and phylogenetic trees were

constructed with PhyML, as described above. Alternative topologies constraining the putative horizontally transferred sequences to their expected organismal position in the tree were compared to the ML tree using the SH Test as implemented in PAUP* (Swofford 2002).

An angiosperm *cox1* intron alignment containing sequences used in previous studies (Sanchez-Puerta M. V. et al. 2008, Sanchez-Puerta Maria V et al. 2011), including the intron from the Orobanchaceae parasite *E. virginiana*, was provided by Dr. Virginia Sanchez-Puerta. Additional Orobanchaceae *cox1* intron sequences were extracted from their best assemblies generated in this study and then manually aligned to the data set. Alignments were trimmed of poorly aligned regions with Gblocks 0.91b using relaxed parameters (b2=half+1, b4=5, b5=half). The final trimmed data set contained 958 aligned nucleotide positions and 194 intron sequences, representing 191 angiosperm species from 60 families (Table S2). The *cox1* intron alignment was then used to construct a phylogenetic tree with PhyML as described above.

RESULTS

The mitochondrial genome of the hemiparasite *Castilleja paramensis*

The complete mitogenome of *C. paramensis* maps as a single circular chromosome that is 495,499 bp in length (Figure 3-1A). The genome includes a total of 67 genes (34 protein-coding, 3 rRNA, and 30 tRNA) and 23 introns (17 *cis*-spliced and 6 *trans*-spliced). In addition to these functional elements, repeats and MIPTs (mitochondrial DNA of plastid origin) comprise a substantial component of this genome (Figure 3-1B). There is one

large repeat of 8.7 kb, 15 intermediate repeats from 100 to 447 bp, and 32 small repeats between 50 and 100 bp. Together, these repeats cover 2.7% (13,525 bp) of the genome. A total of 43 MIPTs were also identified. Ranging in size from 116 bp to 7.7 kb, these MIPTs cover 16.6% (82,133 bp) of the mitogenome, which is the highest MIPT percentage observed in any plant yet sequenced. The MIPTs contain 55 full-length or nearly full-length plastid genes, about half of which are pseudogenes based on the presence of frameshifting indels and/or premature stop codons. With the exception of the MIPTs and repeats, the depth of sequencing coverage is consistently at ~50x throughout most portions of the mitogenome. The greatly increased coverage depth at most MIPTs is likely due to mismapping of reads in the data set that were derived from the plastome. The two-fold increase in coverage depth of the 8.7 kb repeat relative to the rest of the mitogenome is an indication that this region is in fact present in two copies in the genome, consistent with its status as a repeat.

Limited gene and intron loss from the parasitic Orobanchaceae mitogenomes

In contrast to the extensive mitochondrial gene and intron loss observed in mistletoe, comparative mitogenomic analysis of seven Orobanchaceae species—including a nonparasite (*L. philippensis*), three hemiparasites (*B. pedicularioides*, *C. paramensis*, *S. americana*), and three holoparasites (*O. crenata, O. gracilis, P. ramosa*)—revealed only minor variation in gene and intron content (Figure 3-2). The mitogenomes of all seven Orobanchaceae species share 29 protein-coding genes. This conserved set encompasses 23 of the 24 core genes that are nearly universally present in angiosperm mitogenomes (Adams et al. 2002), including nine subunits for the NADH dehydrogenase complex (*nad1*, *2*, *3*, *4*, *4L*, *5*, *6*, *7*, *9*), the apocytochrome *b* gene for the cytochrome bc_1 complex

(*cob*), three subunits for the cytochrome *c* oxidase complex (*cox1*, *2*, *3*), four of the five subunits for the ATP synthase complex (*atp1*, *4*, *6*, *8*), the four cytochrome *c* maturation factors (*ccmB*, *C*, *Fc*, *Fn*), an intron maturase (*matR*), and a protein translocase (*mttB/tatC*).For the remaining ATP synthase subunit (*atp9*), the gene was detected in all species except *L. philippensis*. However, the lack of detection of this very short gene (only 225 bp) should be interpreted with caution because it could be an artefact of an incomplete draft assembly.

There is more variability in the presence of genes encoding subunits of the ribosomal protein and succinate dehydrogenase complexes among the Orobanchaceae species (Figure 3-2). Six protein members of the large (*rpl10*, *16*) and small (*rps3*, *4*, *12*, *14*) ribosomal subunits were conserved in all seven Orobanchaceae mitogenomes, whereas the remaining seven ribosomal proteins and both succinate dehydrogenase genes were lost or pseudogenized in at least one species. Also, several genes (*O. gracilis rps7*, *O. crenata* and *P. ramosa rps13*, *C. paramensis* and *L. philippensis sdh3*) were tentatively scored as present and putatively functional in this study, although they are truncated by 20–30% and may be pseudogenes. Further analysis is required to assess whether they retain functionality.

In terms of intron content, all examined Orobanchaceae species contain either 22 or 23 introns (Figure 3-2). In all seven Orobanchaceae species, there are 15 introns removed by *cis* splicing and 6 by *trans* splicing. All seven species lack *cox2*-i1, *nad7*-i3, and *rpl2*-i1, as do other sequenced Lamiales species (e.g., *Boea*, *Mimulus*), suggesting that the introns were lost early in Lamiales evolution prior to the radiation of Orobanchaceae. Within the Orobanchaceae, the *cox2*-i2 intron was uniquely lost from *B. pedicularioides*, while in

the *rps10* pseudogenes from *O. crenata* and *O. gracilis*, remnants of the *rps10* intron are still retained.

Evidence for horizontal transfer of Orobanchaceae mitochondrial genes

We used phylogenetic analysis to investigate horizontal transfer involving Orobanchaceae mitochondrial genes. In an initial genomic survey, we identified six genes (*nad4L*, *rpl5*, *rps1*, *rps7*, *rps10*, *sdh3*) exhibiting phylogenetic incongruence with strong bootstrap support (87–100%) for the position of one or more Orobanchaceae parasites (Figure S2). To evaluate the robustness of these initial results, we increased taxon sampling by adding gene sequences from GenBank to the data sets (Table S1) and used predicted RNA sequences instead of DNA sequences to eliminate the potentially confounding effects of RNA edit sites in a second round of phylogenetic analyses (Figure 3-3).

For most genes, the results between the DNA (Figure S2) and RNA (Figure 3-3) analyses consistently provided moderate to strong support for phylogenetic incongruence, implicating a horizontal origin for these Orobanchaceae mitochondrial genes. In the *rpl5* tree, *O. crenata* grouped unexpectedly with *Salvia* in both the DNA analysis (87%) and the RNA analysis(72%). For *rps1*, the *O. crenata* and *Phelipanche* pseudogenes grouped strongly with *Daucus* in Apiales (98% DNA; 96% RNA) rather than with Solanales, the closest relatives to Orobanchaceae in the data set. In the *rps7* tree, three Orobanchaceae species (*Bartsia*, *O. gracilis*, and *Schwalbea*)were strongly excluded (89% DNA, 81% RNA) from their expected position within the lamiids clade, represented by *Asclepias*, *Daucus*, *Panax*, and *Rhazya*. Instead, *O. gracilis* grouped unexpectedly with *Vitis* (85%

in both analyses), while *Bartsia* and *Schwalbea* did not associate strongly with any taxa. The *O. gracilis rps10* gene was moderately to strongly excluded from its expected position within the Lamiales clade (88% DNA; 76% RNA), but it did not form a strongly supported association with any other taxa. Similarly, the *Phelipahe sdh3* gene was strongly excluded from the Lamiales clade (96% in both analyses) but did not associate closely with any other taxa. Alternative topologies constraining the putative horizontally transferred genes to their expected phylogenetic position based on vertical transfer (shown as red dots in Figure 3-3) were rejected by an SH Test for *rps1* (P=0.005), *rps7* (P=0.008), *rps10* (P=0.002), and *sdh3* (P=0.004), but could not be rejected for the *rpl5* gene (P=0.118).

For the *nad4L* gene, the DNA analysis recovered, with maximal support (100%), an unusual long-branched clade that included the two *Orobanche* sequences plus *Malus* (Figure S2). However, examination of RNA editing distribution in this alignment revealed that this anomalous clade was caused by the convergent loss of nearly all RNA editing sites, from 15 sites in most species to only a single site in the *Orobanche* and *Malus* sequences, most likely as a result of retroprocessing (i.e., the genomic integration of a mature, edited transcript). In the RNA analysis, this anomalous clade was not recovered (Figure 3-3). Instead, the two *Orobanche* sequences clustered unexpectedly with *Symphoricarpos* with weak support (54%). Overall, the *nad4L* DNA and RNA trees are highly unresolved with generally poor bootstrap support, suggesting that there is little phylogenetic information provided by this small gene. Nevertheless, the phylogenetic position expected for a vertical transfer scenario (i.e., *Orobanche* with *Phelipanche*) was rejected by an SH Test (P=0.012), lending further support for a horizontal transfer event.

The Orobanchaceae *cox1* intron was acquired vertically from a non-parasitic ancestor

Previous studies have identified the mobile *cox1* intron in a small fraction of angiosperms, between 4% and 25% of the hundreds of examined species in the two most extensive analyses (Sanchez-Puerta M. V. et al. 2008, Sanchez-Puerta Maria V et al. 2011). In contrast to the general scarcity of this intron among angiosperms, it was previously observed that a large fraction ofparasitic plants (15 out of 17 examined species, representing 12 distinct parasitic lineages) possess the intron, including *E. virginiana*, the only Orobanchaceae parasite to be examined thus far (Barkman et al. 2007). In agreement with this observation, we confirmed the presence of this intron in all six parasitic Orobanchaceae species examined in the current study, and also in the nonparasitic *L. philippensis* (Figure 3-2).

The mobile nature of the *cox1* intron, coupled with the overrepresentation of this intron in parasitic plants and the fact that parasitic plants are known to facilitate the horizontal transfer of mitochondrial DNA among species (Davis and Xi 2015, Mower Jeffrey P et al. 2012a, Sanchez-Puerta Maria Virginia 2014), raises two possibilities: 1) parasitic plants may frequently transfer this intron to other angiosperms, explaining the abundant horizontal transmission of the intron among angiosperms, and 2) parasitic plants may frequently acquire this intron from other angiosperms, explaining the overrepresentation of the intron in parasitic plants. Both hypotheses can be tested phylogenetically. If parasitic plants are frequent donors of the intron to other angiosperms, then we would expect to find the introns of recipient angiosperms nested within the parasitic plant clade of introns. If parasitic plants are frequently receiving the intron from other angiosperms,

then we would expect to see the horizontally acquired introns of parasitic plants cluster with the donating angiosperm clades rather than in the expected organismal position for Orobanchaceae species within Lamiales.

Phylogenetic analysis of the *cox1* intron from Orobanchaceae sequences and a diverse collection of other angiosperms demonstrated that neither hypothesis is correct for the parasitic plants in this family (Figure 3-4; Figure S3). Within the tree, there is a clade that comprises all Orobanchaceae parasites, which is nested within a larger clade of Lamiales that includes the nonparasitic *Lindenbergia*,also from Orobanchaceae, plus other species (*Catalpa*, *Paulownia, Rehmannia*) from families that are closely related to the Orobanchaceae. Support for most relationships within this larger Lamiales clade is generally weak (<50% bootstrap support for most branches). Nevertheless, the monophyletic clustering of Orobanchaceae species in the more-or-less expected position within Lamiales indicates that this *cox1* intron was most likely acquired vertically in parasitic Orobanchaceae from a nonparasitic ancestor. Furthermore, the absence of any unexpected species nested within the parasitic Orobanchaceae clade indicates that the parasites did not donate the *cox1* intron to any of the other angiosperm species sampled in the analysis.

Minimal degeneration of the *Castilleja paramensis* plastid genome

The *C. paramensis* plastome (Figure S4) is 152,926 bp in length, with a typical quadripartite structure that includes the large and small single-copy regions separated by two copies of an inverted repeat. Relative to the gene and intron content present in a typical asterid, the *C. paramensis* plastome contains nearly a full set of protein-coding

genes, a full set of 4 rRNAs and 31 tRNAs, and a full set of 21 introns (Figure S4 and S5). The few exceptions involve the pseudogenization of *ndhD* and *ndhF* due to frameshifting indels and *ndhH* and *ndhJ* due to the presence of premature stop codons (Figure 3-5). The pseudogenization of *ndhF* does not apply to all *Castilleja* species, as an intact gene was sequenced from *Castilleja linariifolia* in a previous study (Refulio-Rodriguez and Olmstead 2014). Like *C. paramensis*, the obligate hemiparasite *S. americana* has also lost functionality of several *ndh* genes (pseudogenization of *ndhA*, *ndhD*, *ndhF*, *ndhG*, *ndhJ* and loss of *ndhI*). By contrast, Orobanchaceae holoparasites *in Cistanche*, *Conopholis*,and *Orobanche* have lost ~70% of all of their genes, including nearly all of the photosynthesis-related genes and numerous tRNAs (Figure S5; Cusimano and Wicke 2016, Li et al. 2013, Wicke et al. 2013).

DISCUSSION

Gene loss from Orobanchaceae mitogenomes is unrelated to parasitism

In this study, we generated one complete mitogenome from the hemiparasite *C. paramensis* and draft mitogenomes from six additional Orobanchaceae species, including two more hemiparasites (*B. pedicularioides* and *S. americana*), three holoparasites (*O. crenata, O. gracilis* and *P. ramosa*), and a nonparasite (*L. philippensis*). Despite the wide range of trophic strategies among the examined Orobanchaceae species, their mitogenomes display no evidence of functional degeneration that can be attributed to the adoption of a parasitic lifestyle. The relatively few mitochondrial genes that were lost or pseudogenized are limited to ribosomal proteins and succinate dehydrogenase subunits (Figure 3-2). The loss of these genes is unlikely attributable to the adoption of a parasitic

lifestyle because these same genes have also been lost from the mitogenomes of many non-parasitic land plants (Adams et al. 2002, Guo et al. 2016, Hecht et al. 2011, Skippington et al. 2015). Importantly, their loss is unlikely to have a detrimental effect on mitochondrial activity, as each loss event is usually preceded by the establishment of a homolog in the nucleus that maintains a functional product (e.g., Adams et al. 2001, Kobayashi et al. 1997, Mower J. P. and Bonen 2009, Nugent and Palmer 1991). Like in these other examples, we suggest that the functions of the missing Orobanchaceae mitochondrial genes have been supplanted by nuclear-encoded homologs, although sequencing of the nuclear genome will be required to test this prediction.

In addition to the Orobanchaceae data reported here, large-scale mitogenomic data from a parasitic plant is available from four hemiparasitic mistletoes (Petersen et al. 2015b, Skippington et al. 2015) and three holoparasitic members of Rafflesiaceae(Xi et al. 2013). As in the Orobanchaceae parasites, the three Rafflesiaceae holoparasites contain a nearly complete set of the expected mitochondrial genes, although a substantial fraction were reported to have been acquired horizontally (Xi et al. 2013). By contrast, in the hemiparasitic *V. scurruloideum*, the mitogenome has been greatly reduced in size, and in all four mistletoes the coding content has undergone extreme reduction, including the pseudogenization or loss of all nine *nad* genes encoding subunits of mitochondrial complex I, a NADH dehydrogenase (Petersen et al. 2015b, Skippington et al. 2015). The coordinated loss of functional copies of all nine *nad* genes, which has not been reported for any other multicellular eukaryote, argues against a nuclear transfer scenario and instead suggests that the entire complex I was lost (Skippington et al. 2015), with

nuclear-encoded alternative dehydrogenases(Rasmusson et al. 2004) compensating for the loss of complex I activity.

Thus, while the reduced mitochondrial sequences from mistletoe suggested the possibility of general mitochondrial upheaval in parasitic plants, this does not appear to be the case, at least in Orobanchaceae and Rafflesiaceae. Overall, based on the available data from parasitic plants, there does not appear to be any clear correlation between mitogenomic degradation and the degree of host dependence. This is perhaps not surprising as the mitochondrion is essential for respiration and the production of ATP, and these processes are still required by parasitic plants to generate amino acids and other essential organic molecules. The putative loss of complex I from *Viscum* may reflect the first step in mitogenomic degradation in a parasitic plant, which may be tolerated due to the partially overlapping abilities of the alternative dehydrogenases (Skippington et al. 2015). Regardless, unusual mitogenomic features observed for *Viscum* are clearly not representative of all parasitic plants. Whether this complex or any other seemingly essential mitochondrial genes have been lost in other parasitic lineages awaits further investigation.

More evidence for the role of parasitic plants in facilitating horizontal gene transfer

It is now well established that horizontal transfer occurs with high frequency among plants, and that parasitic plants are often involved as donors or recipients (Davis and Xi 2015, Mower Jeffrey P et al. 2012a, Sanchez-Puerta Maria Virginia 2014). With complete mitogenomes now available from a diverse collection of angiosperms (Figure

S1), it is possible to investigate the extent of horizontal transfer between angiosperms for any mitochondrial gene. Combining these available genome data with the Orobanchaceae genomes generated in this study, we identified six genes with phylogenetic incongruence for the placement of one or more Orobanchaceae parasites, suggestive of horizontal transfer (Figure S2). Expanded taxonomic sampling and alternative topology testing of these genes generally corroborated the suggestion of horizontal transfer (Figure 3-3), although bootstrap support for the incongruence fell below 80% for some of the results from the expanded analyses, weakening the evidence for horizontal transfer. Also, for the *rpl5* analysis, the SH test could not reject the alternative topology that forced the *O. crenata* sequence to cluster with *O. gracilis*, which is the expected position based on vertical transfer.

In addition to the phylogenetic support (bootstrap support and/or SH Tests) for horizontal transfer, there are some additional indications that horizontal transfer, rather than some phylogenetic artifact, is the more likely scenario. First, of the species found to be closely allied with the putative horizontally transferred genes in Orobanchaceae,most of them (*Daucus*, *Salvia*, *Cannabis*, and *Symphoricarpos*) represent larger lineages (Apiaceae, Lamiaceae, Cannabaceae, Dipsacales) that are commonly used as hosts by *Orobanche* and *Phelipanche* (Rubiales et al. 2009, Schneeweiss 2007). Second, no functional native copy of *rps1* and *rps7* has been sequenced from any Lamiales species, suggesting that these genes were lost in the common ancestor of the order. Thus, the detection of these genes in several Orobanchaceae species is consistent with a regain via horizontal transfer. Third, other than the putative cases of horizontal transfer involving Orobanchaceae sequences, there are very few examples of phylogenetic incongruence in these data sets.

The most notable is the weakly supported positioning of *Liriodendron* within monocots (68% bootstrap support) in the *nad4L* analysis and *Amborella* with *Batis* and *Mahonia* (59% bootstrap support) in the *rps1* analysis (Figure 3-3). However, *Liriodendron* and *Amborella* are sole representatives from their respective groups (magnoliids and Amborellales), suggesting that these results could be due to the low taxon sampling of these two groups. The Orobanchaceae results are less likely to be due to poor taxon sampling because at least two Orobanchaceae taxa, plus multiple representatives from other closely related lineages (Lamiales, Solanales, Gentianales), are present in each analysis.

If we accept that horizontal gene transfer has occurred, it remains to be determined whether any of these transfer events resulted in the establishment of a functional gene. Several of the transferred genes are clearly nonfunctional pseudogenes, while others are intact and potentially expressed. Because we did not detect any situations in which more than one intact gene was present in a genome, any foreign genes that are in fact functional would thus be examples of replacement transfer, in which the foreign copy replaced the function of the native copy. Further sequencing of Orobanchaceae genomes will be necessary to corroborate these horizontal transfer cases, particularly from *Orobanche* and *Phelipanche* which are involved in most of the cases identified here.

Finally, it should be noted that taxon sampling in these analyses is not deep enough to unambiguously identify the donor species, even in cases where an Orobanchaceae parasite groups strongly with an unexpected species (e.g., *O. crenata rpl5* with *Salvia*; *O. crenata* and *P. ramosa rps1* with *Daucus*; *O. gracilis rps7* with *Vitis*). *Vitis*, for example, was the only species sampled from all of Vitales, which includes >800 species. In the

rpl5 tree, *Salvia* was the lone representative of the >7000 species in Lamiaceae. And *Daucus*, as the lone campanulid representative in the *rps1* tree,is representing the >33000 species in this group. Denser sampling of these large clades, as well as clades of common hosts, will be important in defining the most likely sources of these transferred genes.

No evidence that the *cox1* intron was acquired or distributed horizontally by Orobanchaceae parasites

Studies have indicated that the angiosperm *cox1* intron was acquired from a fungal donor and then horizontally transferred numerous times among species, evidenced primarily by the sporadic distribution of the intron among species and extensive phylogenetic incongruence in the intron tree (Cho et al. 1998, Sanchez-Puerta M. V. et al. 2008, Vaughn et al. 1995). Barkman *et al*. (2007) made the intriguing observation that nearly all examined parasitic plants possess this intron, but they found no evidence that the intron was acquired from their putative hosts. Alternatively, parasitic plants, particularly those with nonspecific host preferences, may serve as key players in the horizontal spread of the intron.

Using the multiple Orobanchaceae *cox1* introns assembled in this study, we demonstrated that the Orobanchaceae introns were acquired vertically from a nonparasitic ancestor (Figure 3-4; Figure S3), consistent with the initial results of Barkman *et al.*(2007) using a single Orobanchaceae intron sequence. Furthermore, we found no evidence that the Orobanchaceae parasites facilitated the spread of the intron to any of the other intron-containing species included in the phylogeny. Overall, there are no indications that the parasitic lifestyle has had any influence on the presence of the *cox1* intron in

Orobanchaceae or its transfer to other species. Thus, it remains unclear why parasitic plants tend to have the *cox1* intron, or whether the proclivity of parasitic plants for horizontal transfer plays any role in the intron's spread. Broad sampling from additional parasitic plant lineages may help to shed light on any potential connections between the parasitic lifestyle and the distribution *cox1* intron.

Plastid genome degeneration in parasitic plants

Unlike the mitogenome of *C. paramensis*, which exhibits few signs of functional degradation, the *C. paramensis* plastome has frameshift mutations or premature stop codons in four subunits of the plastid NAD(P)H dehydrogenase complex (Figure 3-5). These mutations in the well-conserved *ndh* genes are likely to lead to a reduction or loss of gene function. This pattern of *ndh*-specific degradation in the *C. paramensis* plastome is similar to observations in other hemiparasites such as *S. americana* and some species of *Cuscuta* (Figure S5; Funk et al. 2007, McNeal J. R. et al. 2007, Wicke et al. 2013). The draft plastome from the hemiparasite *Bartsia inaequalis* also lacks intact, full-length copies of several *ndh* genes (*ndhD*, *ndhE*, *ndhG*, and *ndhI*), although it cannot be ruled out that these genes were missed due to the incomplete nature of the genome (Uribe-Convers et al. 2014). Compared with other sequenced hemiparasites, the *C. paramensis* plastome appears to be in the very earliest stages of degradation, as indicated by the small number of genes so-far affected, the limited number of deleterious mutations that have accumulated in each affected gene and the lack of any genes that were deleted completely. Furthermore, an intact *ndhF* gene was detected in another *Castilleja* species, indicating that the pseudogenization of the *C. paramensis ndhF* gene occurred recently within the genus, at some point after *C. paramensis* diverged from other members of the

genus. The *C. paramensis* plastome thus provides strong support for the idea that loss of the NAD(P)H dehydrogenase complex is the first step of plastome degradation in the evolution of heterotrophy in plants (Barrett and Davis 2012, Barrett et al. 2014, Wicke et al. 2013). By contrast, the plastomes from nonphotosynthetic holoparasites are generally much more degraded than those of hemiparasites, affecting not only the full spectrum of photosynthetic genes but also many genes not directly related to photosynthesis (Figure S5; Braukmann et al. 2013, Wicke et al. 2013, Wolfe et al. 1992). Taken together, the collective evidence from available parasitic plastomes provides a clear connection between the degree of plastomic degeneration and heterotrophic dependence.

Although it is possible that these genes have been functionally transferred to the nuclear genome in *C. paramensis*, there has been no demonstration of functional *ndh* gene transfer for any seed plants that have lost the plastid *ndh* genes. Fragments of some *ndh* genes were identified in the nucleus of several Orobanchaceae species (Cusimano and Wicke 2016), but there is no indication that these fragments produce functional proteins. Instead, mounting evidence in multiple lineages—including the pine family, gnetophytes, several orchids, and several species of *Erodium* (Geraniaceae)—has shown that these lost plastid genes were not relocated to the nucleus, and furthermore, that many of the nuclear-encoded subunits of this complex have also been lost (Ruhlman et al. 2015). These results strongly suggest that the entire NAD(P)H complex has been eliminated from these species.

ACKNOWLEDGMENTS

The authors thank Nancy Hepburn and Danilo Minga for assistance with plant collection in Ecuador, Katya Romoleroux and Hugo Navarrete for helping to obtain exportation permits from Ecuador, Kanika Jain and Yizhong Zhang for preparation of DNA samples, Susann Wicke and Susanne Renner for providing access to 454 pyrosequencing data from several Orobanchaceae species prior to publication, and Virginia Sanchez-Puerta for providing an alignment of *cox1* introns. This work was supported by the National Science Foundation (awards IOS 1027529 and MCB 1125386 to JPM) and by a Chinese Scholarship Council award (to WF).

AUTHOR CONTRIBUTIONS

J.P.M. planned and designed the research. W.F., A.Z., M.K., N.S., and J.P.M. performed experiments and analyzed results. N.P.M. and F.G. conducted field work and analyzed results. W.F. and J.P.M. wrote the paper, with contributions from all other authors. All authors read and approved the final version of the text.

ADDITIONAL INFORMATION

Accession codes: Assembled sequences were deposited in GenBank under accession numbers KP940485–KP940493 (draft *B. pedicularioides* mitogenome), KT959111 (*C. paramensis* plastome), KT959112 (*C. paramensis* mitogenome), and KT961690–KT961693 (Orobanchaceae *cox1* genes and introns).

Completing financial interests: The authors declare no competing financial interests.

REFERENCES

Adams KL, Ong HC, Palmer JD. 2001. Mitochondrial gene transfer in pieces: fission of the ribosomal protein gene *rpl2* and partial or complete gene transfer to the nucleus. Molecular Biology and Evolution 18: 2289-2297.

Adams KL, Qiu YL, Stoutemyer M, Palmer JD. 2002. Punctuated evolution of mitochondrial gene content: High and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. Proceedings of the National Academy of Sciences USA 99: 9905-9912.

Alverson AJ, Wei X, Rice DW, Stern DB, Barry K, Palmer JD. 2010. Insights into the evolution of mitochondrial genome size from complete sequences of Citrullus lanatus and Cucurbita pepo (Cucurbitaceae). Molecular Biology and Evolution 27: 1436-1448.

Barkman TJ, McNeal JR, Lim S-H, Coat G, Croom HB, Young ND, dePamphilis CW. 2007. Mitochondrial DNA suggests at least 11 origins of parasitism in angiosperms and reveals genomic chimerism in parasitic plants. BMC Evolutionary Biology 7: 248.

Barrett CF, Davis JI. 2012. The plastid genome of the mycoheterotrophic Corallorhiza striata (Orchidaceae) is in the relatively early stages of degradation. American Journal of Botany 99: 1513-1523.

Barrett CF, Freudenstein JV, Li J, Mayfield-Jones DR, Perez L, Pires JC, Santos C. 2014. Investigating the path of plastid genome degradation in an early-transitional clade of heterotrophic orchids, and implications for heterotrophic angiosperms. Molecular Biology and Evolution 31: 3095-3112.

Bellot S, Renner SS. 2016. The plastomes of two species in the endoparasite genus Pilostyles (Apodanthaceae) each retain just five or six possibly functional genes. Genome Biology and Evolution 8: 189-201.

Bennett JR, Mathews S. 2006. Phylogeny of the parasitic plant family Orobanchaceae inferred from phytochrome A. American Journal of Botany 93: 1039-1051.

Bowe LM, dePamphilis CW. 1996. Effects of RNA editing and gene processing on phylogenetic reconstruction. Molecular Biology and Evolution 13: 1159-1166.

Braukmann T, Kuzmina M, Stefanovic S. 2013. Plastid genome evolution across the genus Cuscuta (Convolvulaceae): two clades within subgenus Grammica exhibit extensive gene loss. Journal of Experimental Botany 64: 977-989.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution 17: 540-552.

Cho Y, Qiu Y-L, Kuhlman P, Palmer JD. 1998. Explosive invasion of plant mitochondria by a group I intron. Proceedings of the National Academy of Sciences USA 95: 14244-14249.

Cusimano N, Wicke S. 2016. Massive intracellular gene transfer during plastid genome reduction in nongreen Orobanchaceae. New Phytologist 210: 680-693.

Davis CC, Xi Z. 2015. Horizontal gene transfer in parasitic plants. Current Opinion in Plant Biology 26: 14-19.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792-1797.

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

Giegé P, Brennicke A. 1999. RNA editing in *Arabidopsis* mitochondria effects 441 C to U changes in ORFs. Proceedings of the National Academy of Sciences USA 96: 15324-15329.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology 59: 307-321.

Guo W, Grewe F, Fan W, Young GJ, Knoop V, Palmer JD, Mower JP. 2016. Ginkgo and Welwitschia mitogenomes reveal extreme contrasts in gymnosperm mitochondrial evolution. Molecular Biology and Evolution: in press.

Guo W, Grewe F, Cobo-Clark A, Fan W, Duan Z, Adams RP, Schwarzbach AE, Mower JP. 2014. Predominant and substoichiometric isomers of the plastid genome coexist

within Juniperus plants and have shifted multiple times during cupressophyte evolution. Genome Biology and Evolution 6: 580-590.

Hecht J, Grewe F, Knoop V. 2011. Extreme RNA editing in coding islands and abundant microsatellites in repeat sequences of Selaginella moellendorffii mitochondria: the root of frequent plant mtDNA recombination in early tracheophytes. Genome Biology and Evolution 3: 344-358.

Kobayashi Y, Knoop V, Fukuzawa H, Brennicke A, Ohyama K. 1997. Interorganellar gene transfer in bryophytes: the functional nad7 gene is nuclear encoded in Marchantia polymorpha. Molecular and General Genetics 256: 589-592.

Li X, Zhang TC, Qiao Q, Ren Z, Zhao J, Yonezawa T, Hasegawa M, Crabbe MJ, Li J, Zhong Y. 2013. Complete chloroplast genome sequence of holoparasite Cistanche deserticola (Orobanchaceae) reveals gene loss and horizontal gene transfer from its host Haloxylon ammodendron (Chenopodiaceae). PLoS One 8: e58747.

McNeal JR, Kuehl JV, Boore JL, de Pamphilis CW. 2007. Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus Cuscuta. BMC Plant Biology 7: 57.

McNeal JR, Bennett JR, Wolfe AD, Mathews S. 2013. Phylogeny and origins of holoparasitism in Orobanchaceae. American Journal of Botany 100: 971-983.

Merckx V, Bidartondo MI, Hynson NA. 2009. Myco-heterotrophy: when fungi host plants. Annals of Botany 104: 1255-1261.

Molina J, et al. 2014. Possible loss of the chloroplast genome in the parasitic flowering plant Rafflesia lagascae (Rafflesiaceae). Molecular Biology and Evolution 31: 793-803.

Mower JP. 2009. The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. Nucleic Acids Research 37: W253-W259.

Mower JP, Palmer JD. 2006. Patterns of partial RNA editing in mitochondrial genes of Beta vulgaris. Molecular Genetics and Genomics 276: 285-293.

Mower JP, Bonen L. 2009. Ribosomal protein L10 is encoded in the mitochondrial genome of many land plants and green algae. BMC Evolutionary Biology 9: 265.

Mower JP, Jain K, Hepburn NJ. 2012a. The role of horizontal transfer in shaping the plant mitochondrial genome. Advances in Botanical Research 63: 41-69.

Mower JP, Case AL, Floro ER, Willis JH. 2012b. Evidence against equimolarity of large repeat arrangements and a predominant master circle structure of the mitochondrial genome from a monkeyflower (Mimulus guttatus) lineage with cryptic CMS. Genome Biology and Evolution 4: 670-686.

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. Molecular Genetics and Genomics 268: 434-445.

Nugent JM, Palmer JD. 1991. RNA-mediated transfer of the gene coxII from the mitochondrion to the nucleus during flowering plant evolution. Cell 66: 473-481.

Petersen G, Cuenca A, Seberg O. 2015a. Plastome evolution in hemiparasitic mistletoes. Genome Biology and Evolution 7: 2520-2532.

Petersen G, Cuenca A, Moller IM, Seberg O. 2015b. Massive gene loss in mistletoe (Viscum, Viscaceae) mitochondria. Scientific Reports 5: 17588.

Piednoel M, Aberer AJ, Schneeweiss GM, Macas J, Novak P, Gundlach H, Temsch EM, Renner SS. 2012. Next-generation sequencing reveals the impact of repetitive DNA across phylogenetically closely related genomes of Orobanchaceae. Molecular Biology and Evolution 29: 3601-3611.

Rasmusson AG, Soole KL, Elthon TE. 2004. Alternative NAD(P)H dehydrogenases of plant mitochondria. Annual Review of Plant Biology 55: 23-39.

Refulio-Rodriguez NF, Olmstead RG. 2014. Phylogeny of Lamiidae. American Journal of Botany 101: 287-299.

Rice DW, et al. 2013. Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. Science 342: 1468-1473.

Richardson AO, Rice DW, Young GJ, Alverson AJ, Palmer JD. 2013. The "fossilized" mitochondrial genome of Liriodendron tulipifera: ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. BMC Biology 11: 29.

Rubiales D, FernÁNdez-Aparicio M, Wegmann K, Joel DM. 2009. Revisiting strategies for reducing the seedbank of Orobanche and Phelipanche spp. Weed Research 49: 23-33.

Ruhlman TA, et al. 2015. NDH expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. BMC Plant Biology 15: 100.

Salmans ML, Chaw SM, Lin CP, Shih AC, Wu YW, Mulligan RM. 2010. Editing site analysis in a gymnosperm mitochondrial genome reveals similarities with angiosperm mitochondrial genomes. Current Genetics 56: 439-446.

Sanchez-Puerta MV. 2014. Involvement of plastid, mitochondrial and nuclear genomes in plant-to-plant horizontal gene transfer. Acta Societatis Botanicorum Poloniae 83: 317-323.

Sanchez-Puerta MV, Cho Y, Mower JP, Alverson AJ, Palmer JD. 2008. Frequent, phylogenetically local horizontal transfer of the cox1 group I intron in flowering plant mitochondria. Molecular Biology and Evolution 25: 1762-1777.

Sanchez-Puerta MV, Abbona CC, Zhuo S, Tepe EJ, Bohs L, Olmstead RG, Palmer JD. 2011. Multiple recent horizontal transfers of the cox1 intron in Solanaceae and extended co-conversion of flanking exons. BMC Evolutionary Biology 11: 277.

Schneeweiss GM. 2007. Correlated evolution of life history and host range in the nonphotosynthetic parasitic flowering plants Orobanche and Phelipanche (Orobanchaceae). Journal of Evolutionary Biology 20: 471-478.

Skippington E, Barkman TJ, Rice DW, Palmer JD. 2015. Miniaturized mitogenome of the parasitic plant Viscum scurruloideum is extremely divergent and dynamic and has lost all nad genes. Proceedings of the National Academy of Sciences USA 112: E3515-E3524.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Research 34: W609-W612.

Swofford DL. 2002. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sunderland, MA: Sinauer Associates.

Uribe-Convers S, Duke JR, Moore MJ, Tank DC. 2014. A long PCR-based approach for DNA enrichment prior to next-generation sequencing for systematic studies. Applications in Plant Science 2.

van der Kooij TA, Krause K, Dorr I, Krupinska K. 2000. Molecular, functional and ultrastructural characterisation of plastids from six species of the parasitic flowering plant genus Cuscuta. Planta 210: 701-707.

Vaughn JC, Mason MT, Sper-Whitis GL, Kuhlman P, Palmer JD. 1995. Fungal origin by horizontal transfer of a plant mitochondrial group I intron in the chimeric coxI gene of Peperomia. Journal of Molecular Evolution 41: 563-572.

Westwood JH, Yoder JI, Timko MP, dePamphilis CW. 2010. The evolution of parasitism in plants. Trends in Plant Science 15: 227-235.

Wicke S, Muller KF, de Pamphilis CW, Quandt D, Wickett NJ, Zhang Y, Renner SS, Schneeweiss GM. 2013. Mechanisms of functional and physical genome reduction in photosynthetic and nonphotosynthetic parasitic plants of the broomrape family. The Plant Cell 25: 3711-3725.

Wolfe KH, Morden CW, Palmer JD. 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proceedings of the National Academy of Sciences USA 89: 10648-10652.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20: 3252-3255.

Xi Z, Wang Y, Bradley RK, Sugumaran M, Marx CJ, Rest JS, Davis CC. 2013. Massive mitochondrial gene transfer in a parasitic flowering plant clade. PLoS Genetics 9: e1003265.

Young ND, Steiner KE, Depamphilis CW. 1999. The evolution of parasitism in Scrophulariaceae/Orobanchaceae: plastid gene sequences refute an evolutionary transition series. Annals of the Missouri Botanical Garden 86: 876-893.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18: 821-829.

Zhu A, Guo W, Jain K, Mower JP. 2014. Unprecedented heterogeneity in the synonymous substitution rate within a plant genome. Molecular Biology and Evolution 31: 1228-1236.

Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. Systematic Biology 51: 588-598.

FIGURES

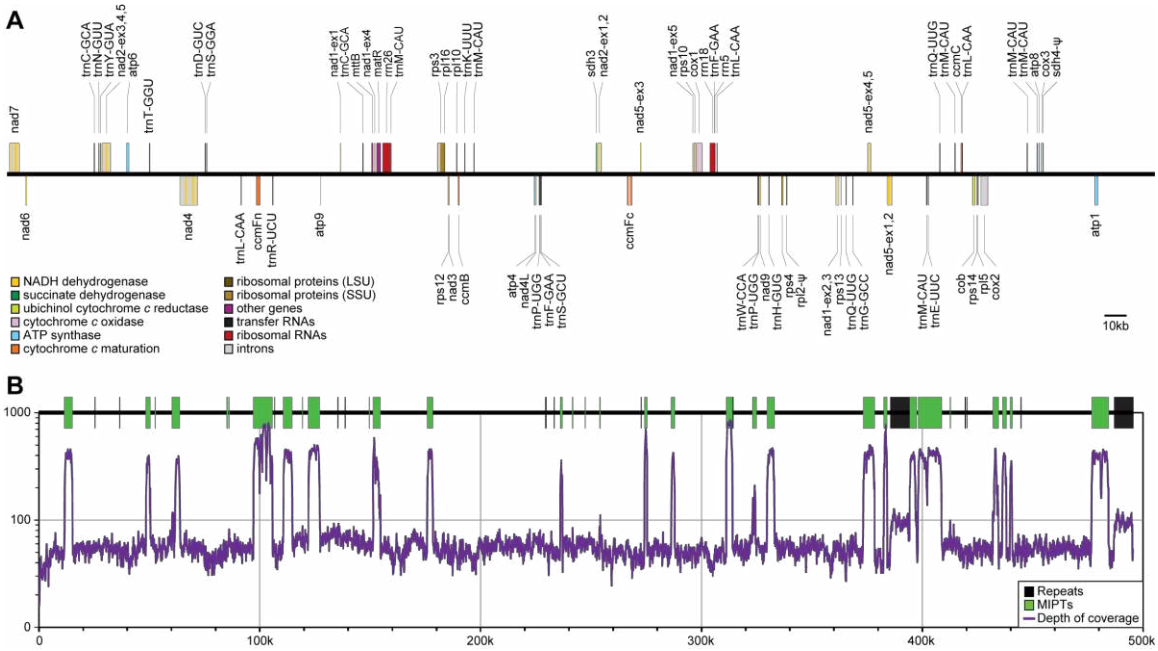

Figure 3-1. The *Castilleja paramensis* mitogenome. A) Gene and intron map. Top genes are transcribed in the forward direction; bottom genes are transcribed in the reverse direction. Colors correspond to the functional categories listed in the key. B) Correlation of repeats and MIPTs with depth of sequencing coverage. The location of all repeats (black) and MIPTs (green) >100 bp in length are shown. Genome maps were drawn with OgDraw (http://ogdraw.mpimp-golm.mpg.de/).

	Orobanchaceae							Other Asterids			
	Cpa	Bpe	Ocr	Ogr	Pra	Sam	Lph	Mgu	Bhy	Nta	Dca
Protein genes											
28 genes	+	+	+	+	+	+	+	+	+	+	+
atp9	+	+	+	+	+	+	–	+	+	+	+
rpl2	ψ	+	–	–	ψ	+	+	ψ	+	+	–
rpl5	+	+	+	+	+	–	+	+	+	+	+
rps1	–	–	ψ	–	ψ	–	–	–	–	+	+
rps7	–	ψ	–	+	–	ψ	–	–	–	–	+
rps10	+	+	ψ	ψ	+	+	+	+	+	+	–
rps13	+	+	+	ψ	+	+	+	+	+	+	+
rps14	+	+	+	+	+	+	+	+	+	ψ	–
rps19	–	–	–	–	–	–	–	–	–	+	–
sdh3	+	–	–	–	+	ψ	+	+	+	+	–
sdh4	ψ	ψ	ψ	ψ	–	+	+	+	+	+	ψ
Introns											
14 cis	+	+	+	+	+	+	+	+	+	+	+
6 trans	+	+	+	+	+	+	+	+	+	+	+
cox1-i1	+	+	+	+	+	+	+	–	+	–	–
cox2-i1	–	–	–	–	–	–	–	–	–	+	+
cox2-i2	+	–	+	+	+	+	+	+	+	–	+
nad7-i3	–	–	–	–	–	–	–	–	–	–	+
rpl2-i1	–	–	–	–	–	–	–	–	–	+	x
rps10-i1	+	+	ψ	ψ	+	+	+	+	+	+	x
total genes	34	34	32	32	34	34	35	35	36	37	33
total introns	23	22	22	22	23	23	23	22	23	23	23

Figure 3-2. Mitochondrial gene and intron content in Orobanchaceae and selected asterids. Genes and introns present in each genome are marked with a plus symbol ("+"). Lost genes and introns ("–"), pseudogenized genes and introns ("ψ"), and missing introns due to loss of the host gene ("x") are shaded gray. The 28 genes include *atp1*, *atp4*, *atp6*, *atp8*, *ccmB*, *ccmC*, *ccmFc*, *ccmFn*, *cob*, *cox1*, *cox2*, *cox3*, *matR*, *mttB*, *nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad7*, *nad9*, *rpl10*, *rpl16*, *rps3*, *rps4*, and *rps12*. The 14 *cis*-arrangedintrons include *ccmFc*-i1, *nad1*-i2, *nad2*-i1, *nad2*-i3, *nad2*-i4, *nad4*-i1, *nad4*-i2, *nad4*-i3, *nad5*-i1, *nad5*-i4, *nad7*-i1, *nad7*-i2, *nad7*-i4, and *rps3*-i1. The six *trans*-

arrangedintrons include *nad1*-i1*, nad1*-i3*,nad1*-i4, *nad2*-i2*, nad5*-i2, and *nad5*-i3. Cpa = *Castilleja paramensis*; Bpe = *Bartsia pedicularioides*; Ocr = *Orobanche crenata*; Ogr = *Orobanche gracilis*; Pra = *Phelipanche ramosa*; Sam = *Schwalbea americana*; Lph = *Lindenbergia philippensis*; Mgu = *Mimulus guttatus*; Bhy = *Boea hygrometrica*; Nta = *Nicotiana tabacum*; Dca = *Daucus carota*.

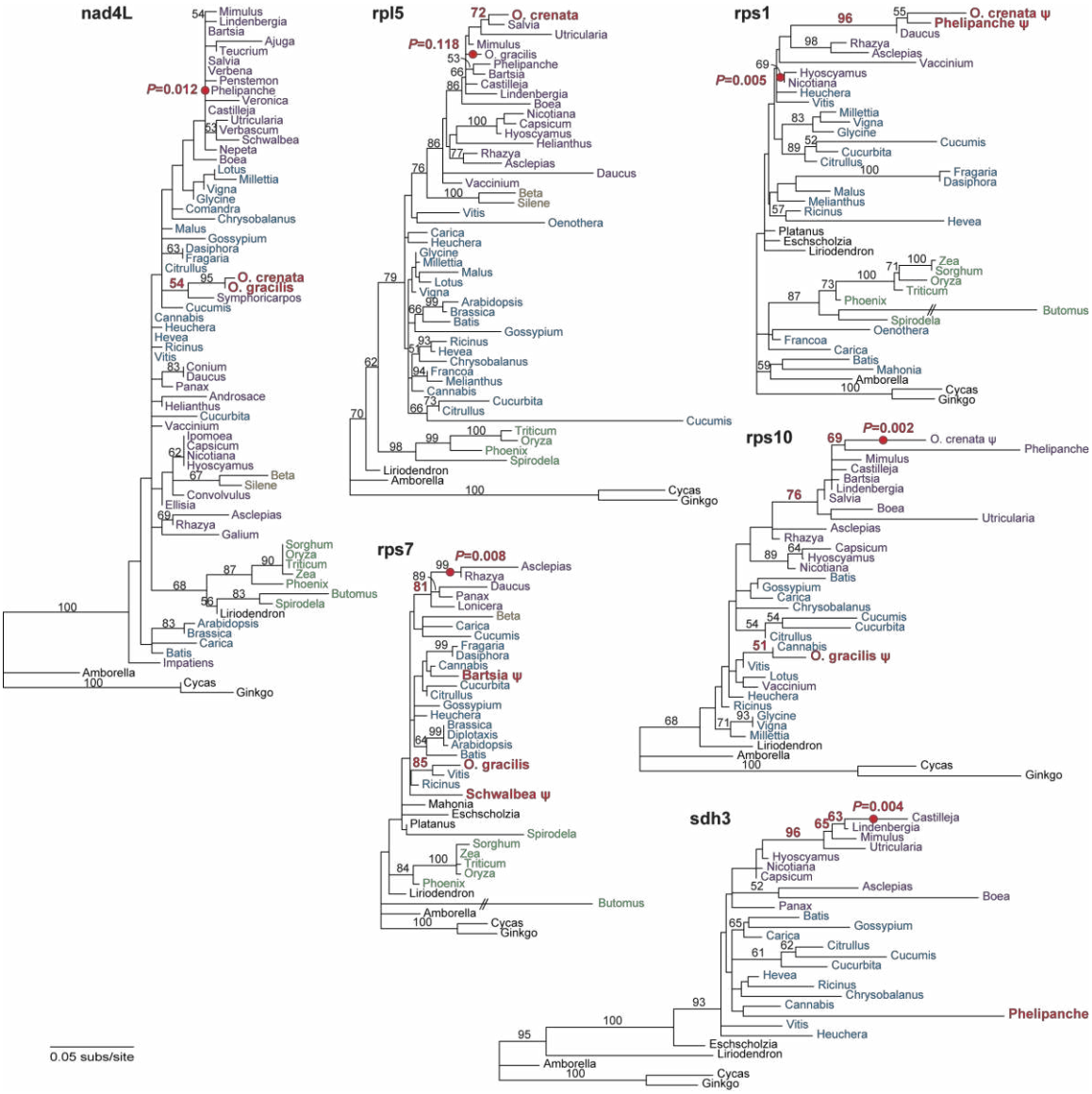

Figure 3-3. Phylogenetic evidence for horizontal gene transfer. Shown are the expanded reanalyses of the six genes identified from the initial genome survey for one or more Orobanchaceae species. The incongruently placed taxa and relevant bootstrap values supporting the incongruent placement are shown in bold red text, and their expected phylogenetic position under a scenario of vertical transfer is marked with a red dot. *P*-value results of an SH Test constraining the incongruently placed taxa at their expected vertical transfer position are shown in bold red text. Asterids are shown in purple,

caryophyllids in brown, rosids in blue, monocots in green, and other seed plants in black. Bootstrap values for all branches with ≥50% support are shown above the branch. Pseudogenes are marked with a "ψ". All trees were drawn to the same scale, shown at bottom left.

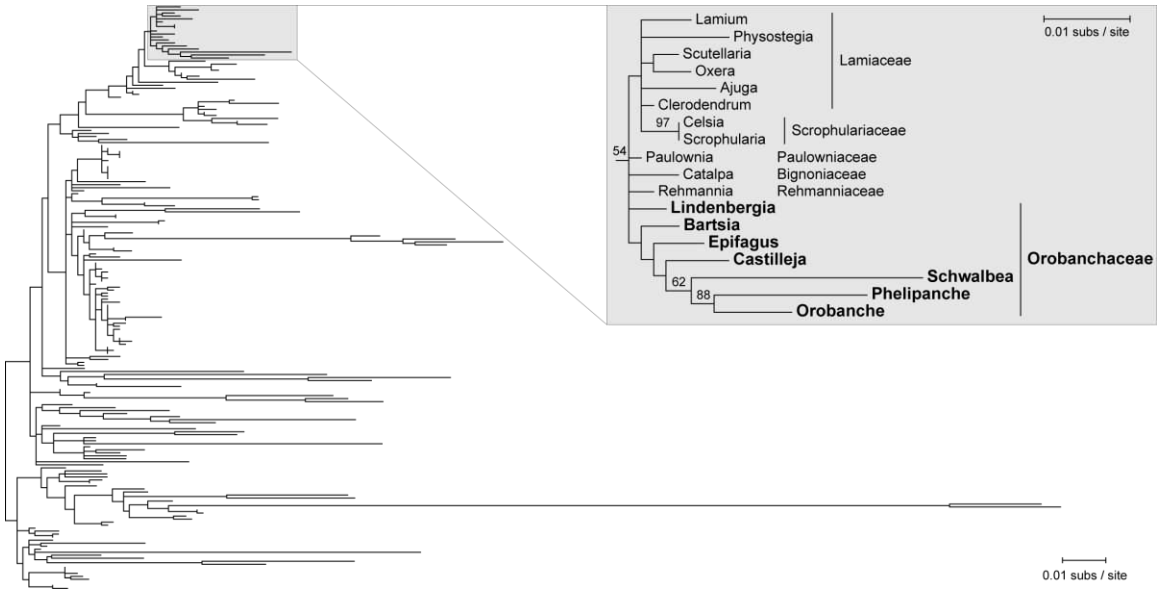

Figure 3-4. Phylogenetic analysis of the mobile *cox1* intron. The tree results from maximum likelihood evaluation of 194 intron sequences from diverse angiosperms. The expanded section of the tree depicts a clade of Lamiales sequences from the seven Orobanchaceae species (large, bold text) and closely related families. Family names are labeled to the right of the subtree. Bootstrap values ≥50% from 1000 replicates are shown on the branch. The subtree is drawn to a 2-fold expanded scale relative to the full tree; scale bars for the subtree and full tree are shown at top right and bottom right, respectively.

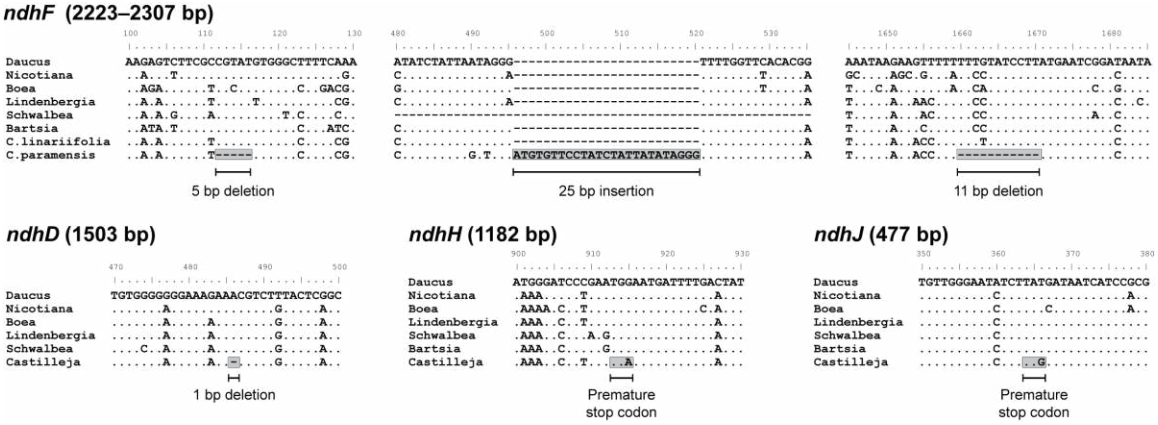

Figure 3-5. Evidence for pseudogenization of *Castilleja paramensis ndh* genes. Shown are sections of four *ndh* gene alignments with evidence of pseudogenization. The frameshifting indels and premature stop codons leading to loss of function are shaded in gray. The full length of functional versions of each gene is shown in parentheses next to each gene name. C. = *Castilleja*.

SUPPORTING INFORMATION

Table S1. GenBank accession numbers for mitochondrial sequences used in phylogenetic analysis

Species	Accn. No.
Whole and draft genomes (initial survey)	
Ajuga reptans	NC_023103
Amborella trichopoda	KF754800-KF754803
Arabidopsis thaliana	NC_001284
Asclepias syriaca	NC_022796
Bartsia pedicularioides	KP940485-KP940493
Batis maritima	NC_024429
Beta macrocarpa	NC_015994
Boea hygrometrica	NC_016741
Brassica napus	NC_008285
Butomus umbellatus	NC_021399
Capsicum annuum	NC_024624
Carica papaya	NC_012116
Castilleja paramensis	KT959112
Citrullus lanatus	NC_014043
Cucumis sativus	NC_016005
Cucurbita pepo	NC_014050
Cycas taitungensis	NC_010303
Daucus carota	NC_017855
Geranium maderense	NC_027000
Ginkgo biloba	NC_027976
Glycine max	NC_020455
Gossypium harknessii	NC_027407
Helianthus annuus	NC_023337
Hyoscyamus niger	NC_026515
Lindenbergia philippensis	SRX105023
Liriodendron tulipifera	NC_021152
Lotus japonicus	NC_016743
Malus domestica	NC_018554
Millettia pinnata	NC_016742
Mimulus guttatus	NC_018041
Nicotiana tabacum	NC_006581
Orobanche crenata	SRX105035
Orobanche gracilis	SRX105038
Oryza sativa	NC_007886
Panax ginseng	KF735063
Phelipanche ramosa	SRX105036
Phoenix dactylifera	NC_016740
Rhazya stricta	NC_024293
Ricinus communis	NC_015141
Salvia miltiorrhiza	NC_023209
Schwalbea americana	SRX105033
Silene latifolia	NC_014487
Sorghum bicolor	NC_008360
Spirodela polyrhiza	NC_017840
Triticum aestivum	NC_007579
Vaccinium macrocarpon	NC_023338
Vigna angularis	NC_021092
Vitis vinifera	NC_012119
Zea mays	NC_007982
Whole genomes (expanded analysis)	
Cannabis sativa	KR059940
Heuchera parviflora	KR559021
Hevea brasiliensis	AP014526

Species	Accn. No.
nad4L	
Androsace occidentalis	KT179529
California macrophylla	KP962982
Chrysobalanus icaco	KJ414531
Cicuta maculata	KT179500
Comandra umbellata	KT179514
Conium maculatum	KT179501
Convolvulus arvensis	KT179506
Dasiphora fruticosa	KC825144
Ellisia nyctelea	KT179518
Erodium texanum	KP962983
Evolvulus nuttallianus	KT179508
Fragaria vesca	KC825150
Galium aparine	KT179531
Impatiens capensis	KT179530
Ipomoea leptophylla	KT179507
Monarda fistulosa	KT179519
Monsonia emarginata	KP963001
Nepeta cataria	KT179521
Penstemon angustifolius	KT179522
Physalis heterophylla	KT179509
Plantago patagonica	KT179517
Symphoricarpos occidentalis	KT179498
Teucrium canadense	KT179527
Utricularia gibba	KC997784
Verbascum thapsus	KT179524
Verbena hastata	KT179525
Veronica americana	KT179526
rpl5	
Chrysobalanus icaco	KJ415044
Francoa sonchifolia	KP963153
Melianthus villosus	KP963154
Oenothera berteriana	X69553
Utricularia gibba	KC997780
rps1	
Dasiphora fruticosa	KC825081
Erodium texanum	KP963107
Eschscholzia californica	AY832261
Fragaria vesca	KC825087
Francoa sonchifolia	KP963108
Mahonia bealei	AY832256
Melianthus villosus	KP963127
Monsonia emarginata	KP963128
Oenothera berteriana	X78038
Platanus occidentalis	AY832259
rps7	
Dasiphora fruticosa	KC825054
Diplotaxis muralis	AB243571
Eschscholzia californica	AY832300
Fragaria vesca	KC825060
Lonicera sp.	AY832297
Mahonia bealei	AY832298
Platanus occidentalis	AY832306
rps10	
Chrysobalanus icaco	KJ414982
Utricularia gibba	KC997779
sdh3	
Chrysobalanus icaco	KJ415016
Eschscholzia californica	EU924198
Utricularia gibba	KC997787

Table S2. GenBank accession numbers for all *cox1* intron sequences used in phylogenetic analysis

Family	Species	Accn. No.
Acanthaceae	Barleria prionitis	AJ247601
Acanthaceae	Justicia americana	AJ247602
Acanthaceae	Sanchezia nobilis	AJ223437
Acanthaceae	Thunbergia erecta	AJ247603
Anacardiaceae	Rhus glabra	EU281065
Apocynaceae	Alstonia plumosa	EU069541
Apocynaceae	Alyxia loesneriana	EU069542
Apocynaceae	Asclepias tuberosa	EU281054
Apocynaceae	Hoya sikkimensis	AJ247588
Apocynaceae	Neisosperma brevitubum	EU069543
Apocynaceae	Nerium oleander	AJ223421
Apocynaceae	Ochrosia elliptica	EU069544
Apocynaceae	Vinca rosea	AJ223423
Apodanthaceae	Pilostyles thurberi AZ	EU281092
Apodanthaceae	Pilostyles thurberi TX	EU281018
Aquifoliaceae	Ilex sp. Qiu 94038	AJ223429
Araceae	Alocasia cucullata	EF517193
Araceae	Alocasia gageana	EF517194
Araceae	Alocasia navicularis	EF517195
Araceae	Amorphophallus rivieri	AJ007548
Araceae	Ariopsis protanthera	EF517198
Araceae	Arisaema speciosum	EF517176
Araceae	Arisaema tortuosum	EF517177
Araceae	Arisaema triphyllum	AY009454
Araceae	Arum concinnatum	EF517179
Araceae	Arum dioscoridis	EF517180
Araceae	Arum italicum	EF517181
Araceae	Biarum davisii	EF517182
Araceae	Biarum tenuifolium	EF517183
Araceae	Caladium bicolor	EF517207
Araceae	Colocasia esculenta	EF517196
Araceae	Dracunculus canariensis	EF517184
Araceae	Dracunculus vulgaris	EF517185
Araceae	Eminium spiculatum	EF517186
Araceae	Helicodiceros muscivorus	EF517187
Araceae	Philodendron oxycardium	AJ223438
Araceae	Pinellia cordata	EF517175
Araceae	Pinellia ternata	EF517178
Araceae	Pistia stratiotes	AJ007546
Araceae	Remusatia vivipara	EF517197
Araceae	Steudnera cf. discolor	EF517200
Araceae	Steudnera discolor	EF517199
Araceae	Steudnera griffithii	EF517210
Araceae	Steudnera kerrii	EF517208
Araceae	Steudnera kerrii	EF517209
Araceae	Theriophonum infaustum	EF517202
Araceae	Typhonium albidinervum	EF517192
Araceae	Typhonium giganteum	EF517189
Araceae	Typhonium hirsutum	EF517190
Araceae	Typhonium trilobatum	EF517188
Araceae	Typhonium venosum	EF517191
Araceae	Xanthosoma mafaffa	AJ223807
Araceae	Xanthosoma sagittifolium	EF517206
Araceae	Zamioculcas zamiifolia	AJ007547
Araliaceae	Hydrocotyle rotundifolia	AJ223424
Aristolochiaceae	Aristolochia elegans	AY009431
Aristolochiaceae	Asimina triloba	AY009433
Balanophoraceae	Ombrophytum subterraneum	EU281081
Bignoniaceae	Catalpa fargesii	AJ223411
Burseraceae	Bursera simaruba	EU281030
Calceolariaceae	Calceolaria sp. IUGH	AJ247585
Celastraceae	Brexia madagascarensis	AJ223413
Clusiaceae	Montrouziera cauliflora	EU069550
Convolvulaceae	Cuscuta japonica	EU281077
Convolvulaceae	Ipomoea coccinea	EU281050
Cucurbitaceae	Citrullus lanatus	EU069546
Cucurbitaceae	Cucumis melo	EU069547
Cucurbitaceae	Cucumis metuliferus	EU069548
Cucurbitaceae	Cucumis sativus	AJ223416
Cucurbitaceae	Neoachmandra indica	EU069549
Cynomoriaceae	Cynomorium coccineum	EU281023
Cytinaceae	Cytinus ruber	EU281022
Dipterocarpaceae	Shorea talura	AJ247599
Droseraceae	Dionaea muscipula	AY600108
Ebenaceae	Diospyros virginiana	AJ223417
Ehretiaceae	Ehretia anacua	AJ247606
Ehretiaceae	Lennoa madreporoides	EU281080
Ehretiaceae	Pholisma arenarium	EU281083
Ericaceae	Pyrola secunda	AJ247582
Euphorbiaceae	Acalypha sp. Qiu95079	AJ247597
Euphorbiaceae	Codiaeum peltatum	EU069551
Euphorbiaceae	Croton alabamensis	EU281037
Euphorbiaceae	Croton sp. Qiu 94027	AJ247608
Euphorbiaceae	Euphorbia millii	AJ223418
Euphorbiaceae	Hevea brasiliensis	AJ223436
Euphorbiaceae	Hura crepitans	AJ247584
Gentianaceae	Frasera caroliniensis	EU281038
Gesneriaceae	Drymonia serrulata	AJ247579
Gesneriaceae	Nematanthus hirsutus	AJ247578
Heliotropiaceae	Heliotropium arborescens	AJ223425
Hydnoraceae	Hydnora africana	EU281079
Hydnoraceae	Prosopanche americana	EU281082
Lamiaceae	Ajuga reptans	AJ247595
Lamiaceae	Clerodendrum trichotomum	AJ223414
Lamiaceae	Lamium sp. Qiu 95015	AJ223428
Lamiaceae	Oxera sp. MVSP-2007	EU069545
Lamiaceae	Physostegia virginiana	AJ247594
Lamiaceae	Scutellaria mociniana	AJ247593
Lauraceae	Cassytha filiformis	EU281076
Lecythidaceae	Barringtonia asiatica	AJ247581
Lecythidaceae	Barringtonia racemosa	GU321958
Linaceae	Linum sp. Qiu96175	AJ247604
Loganiaceae	Strychnos spinosa	AJ247596
Loranthaceae	Dendrophthoe pentandra	EU281073
Malpighiaceae	Malpighia glabra	AJ223433
Marantaceae	Ctenanthe setosa	AY673019
Marantaceae	Maranta bicolor	AY673024
Marantaceae	Maranta leuconeura	AJ223432
Marantaceae	Monotagma laxum	AY673026
Marantaceae	Saranthe leptostachya	EU069559
Marantaceae	Saranthe sp. Kress 96-5737	AY673030
Meliaceae	Dysoxylum canalense	EU069558
Meliaceae	Melia toosendan	AJ223420
Mitrastemonaceae	Mitrastema yamamotoi	EU281021
Musaceae	Musa acuminata	AJ247609
Musaceae	Musella lasiocarpa	AY673040
Myristicaceae	Knema latericia	AJ223430
Myristicaceae	Myristica fragrans	AJ223434
Oleaceae	Jasminum floridum	EU281051
Oleaceae	Jasminum polyanthum	AJ247607

Table S2. Continued

Family	Species	Accn. No.
Opiliaceae	Lepionurus sylvestris	AJ223439
Orchidaceae	Chamorchis alpina	EF143191
Orobanchaceae	Bartsia pedicularioides	KP940490
Orobanchaceae	Castilleja paramensis	KT959112
Orobanchaceae	Epifagus virginiana	EU281078
Orobanchaceae	Lindenbergia philippensis	KT961690
Orobanchaceae	Orobanche crenata	KT961691
Orobanchaceae	Phelipanche ramosa	KT961692
Orobanchaceae	Schwalbea americana	KT961693
Paulowniaceae	Paulownia tomentosa	AJ247592
Pedaliaceae	Sesamum indicum	AJ247598
Phyllanthaceae	Breynia nivosa	AJ247605
Phyllanthaceae	Phyllanthus gneissicus	EU069552
Piperaceae	Peperomia cubensis	AF029783
Piperaceae	Peperomia griseoargentea	AF029781
Piperaceae	Peperomia obtusifolia	AF029782
Piperaceae	Peperomia polybotrya	X87336
Plantaginaceae	Aragoa abietina	EU069508
Plantaginaceae	Aragoa cundinamarcensis	EU069509
Plantaginaceae	Callitriche heterophylla	AJ247577
Plantaginaceae	Callitriche sp.	unpublished
Plantaginaceae	Digitalis purpurea	AJ223415
Plantaginaceae	Globularia punctata	EU156494
Plantaginaceae	Hebe subalpina	AJ223419
Plantaginaceae	Plantago arenaria	EU069513
Plantaginaceae	Plantago atrata	EU069536
Plantaginaceae	Sibthorpia peregrina	EU069540
Plantaginaceae	Veronica agrestis	AJ223427
Polygalaceae	Polygala sanguinea	EU281061
Polygalaceae	Polygala verticillata	unpublished
Rafflesiaceae	Rafflesia pricei	EU281020
Rafflesiaceae	Rhizanthes lowii	EU281019
Rehmanniaceae	Rehmannia glutinosa	AJ247589
Rhamnaceae	Bathiorhamnus cryptophorus	EU069557
Rhamnaceae	Frangula alnus	EU156521
Rhamnaceae	Frangula caroliniana	EU281063
Rhamnaceae	Hovenia dulcis	AJ247583
Rhamnaceae	Maesopsis eminii	EU069554
Rhamnaceae	Nesiota elliptica	EU156528
Rhamnaceae	Paliurus spina-christi	EU156525
Rhamnaceae	Phylica emirnensis	EU069555
Rhamnaceae	Rhamnella franguloides	EU156523
Rhamnaceae	Rhamnus alpina	unpublished
Rhamnaceae	Rhamnus cathartica	AJ223422
Rhamnaceae	Ziziphus ornata	EU156526
Rubiaceae	Calycosiphonia macrochlamys	EU156532
Rubiaceae	Coffea arabica	AJ247586
Rubiaceae	Ixora sp. Kew 21361	EU156535
Rubiaceae	Ixora sp. Qiu95051	AJ247587
Scrophulariaceae	Celsia arturus	AJ247590
Scrophulariaceae	Scrophularia nodosa	AJ247591
Solanaceae	Brunfelsia jamaicensis	JF966280
Solanaceae	Hyoscyamus aureus	JF966283
Solanaceae	Hyoscyamus boveanus	JF966294
Solanaceae	Hyoscyamus desertorum	JF966293
Solanaceae	Hyoscyamus muticus	JF966292
Solanaceae	Hyoscyamus niger	JF966290
Solanaceae	Hyoscyamus pusillus	JF966291
Solanaceae	Mandragora autumnalis	JF966297
Solanaceae	Mandragora officinarum	JF966295
Solanaceae	Mandragora sp. Kew23330	JF966296
Solanaceae	Physochlaina infundibularis	JF966285
Solanaceae	Physochlaina orientalis	JF966281
Solanaceae	Przewalskia tangutica	JF966284
Symplocaceae	Symplocus paniculata	AJ223435
Urticaceae	Pilea fontana	AJ247580
Violaceae	Hybanthus sp. IND-JM3091	EU069553
Violaceae	Viola sp. Qiu95018	AJ247600
Zingiberaceae	Boesenbergia rotunda	EU069561
Zingiberaceae	Gagnepainia godefroyi	EU069564
Zingiberaceae	Globba sessiliflora	EU069565
Zingiberaceae	Hedychium coronarium	AJ223426
Zingiberaceae	Kaempferia rotunda	EU069566
Zingiberaceae	Siphonochilus decorus	AY673043

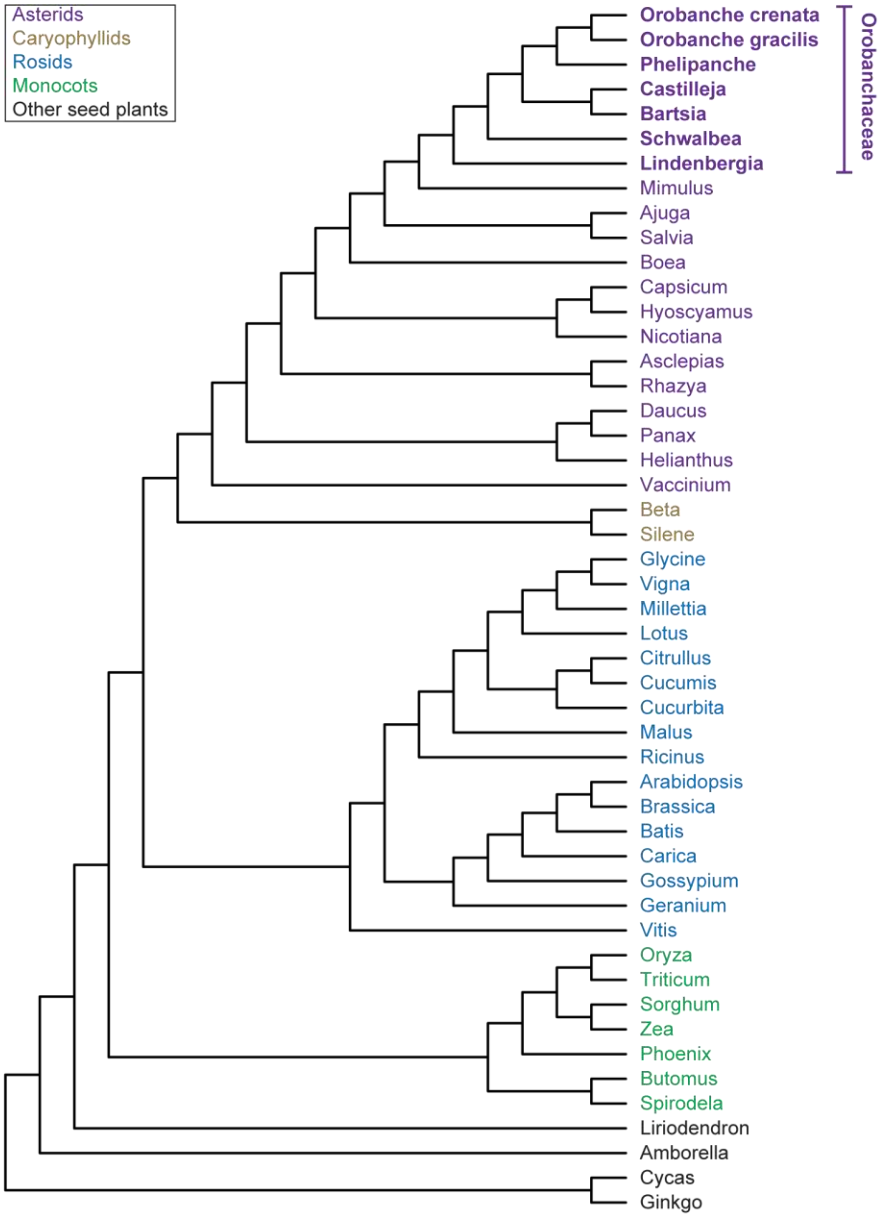

Figure S1. Mitogenome sampling for the initial genomic survey of horizontal gene transfer. Taxonomic groups are color coded according to the key at top left. Orobanchaceae species added for this study are shown in bold text. Species relationships were taken from version 13 of the Angiosperm Phylogeny Website (http://www.mobot.org/mobot/research/apweb/) and references therein.

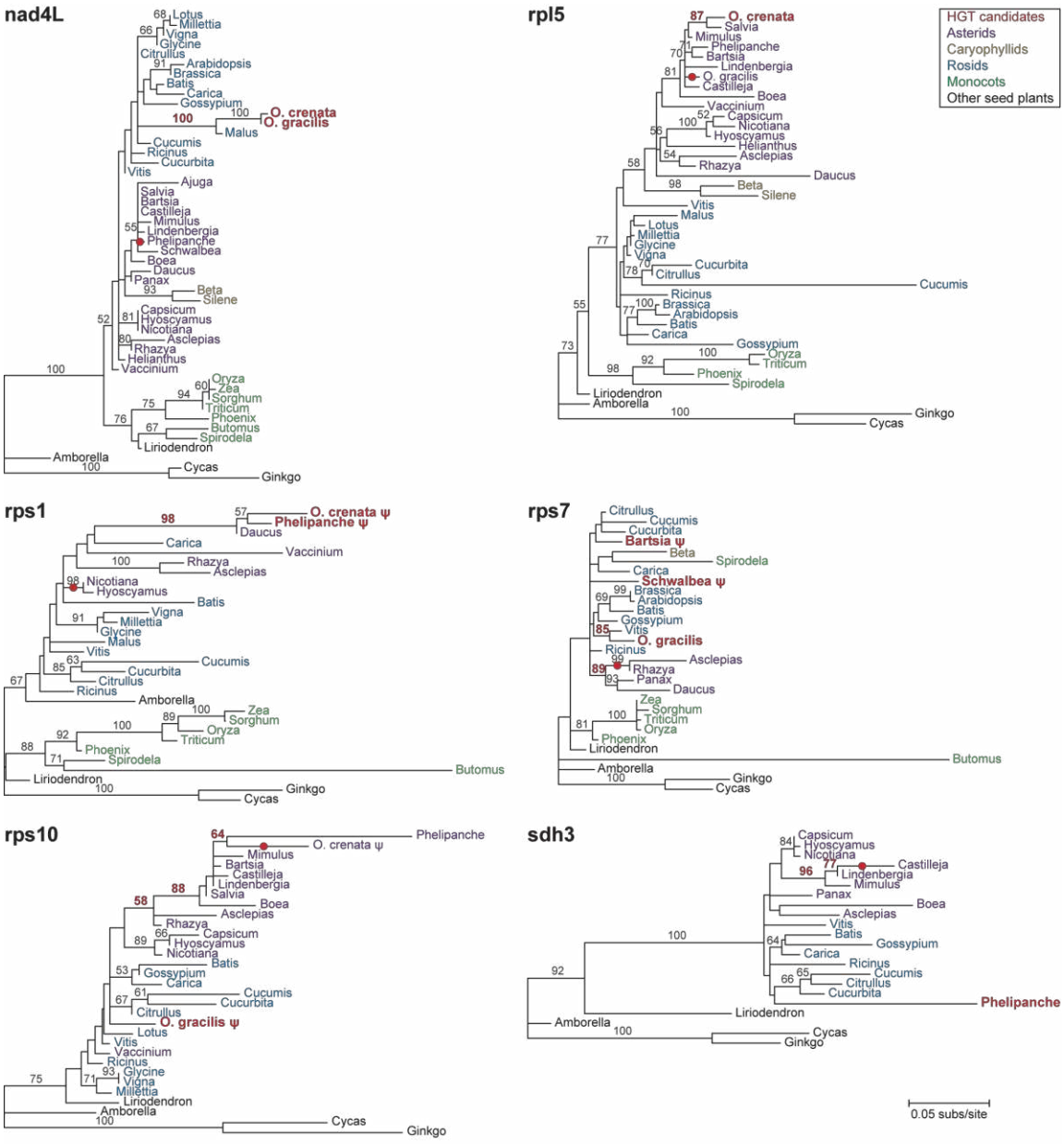

Figure S2. Phylogenetic evidence for horizontal gene transfer from the initial genomic survey. Shown are the six genes demonstrating phylogenetic incongruence, with strong (>80%) bootstrap support, for one or more Orobanchaceae species. The incongruently placed taxa and relevant bootstrap values are shown in bold red text, and their expected phylogenetic position under a scenario of vertical transfer is marked with a red circle. Taxonomic groups are color coded according to the key at top right. Bootstrap values for all branches with >50% support are shown above the branch. Pseudogenes are marked with a "ψ". All trees were drawn to the same scale, shown at bottom right.

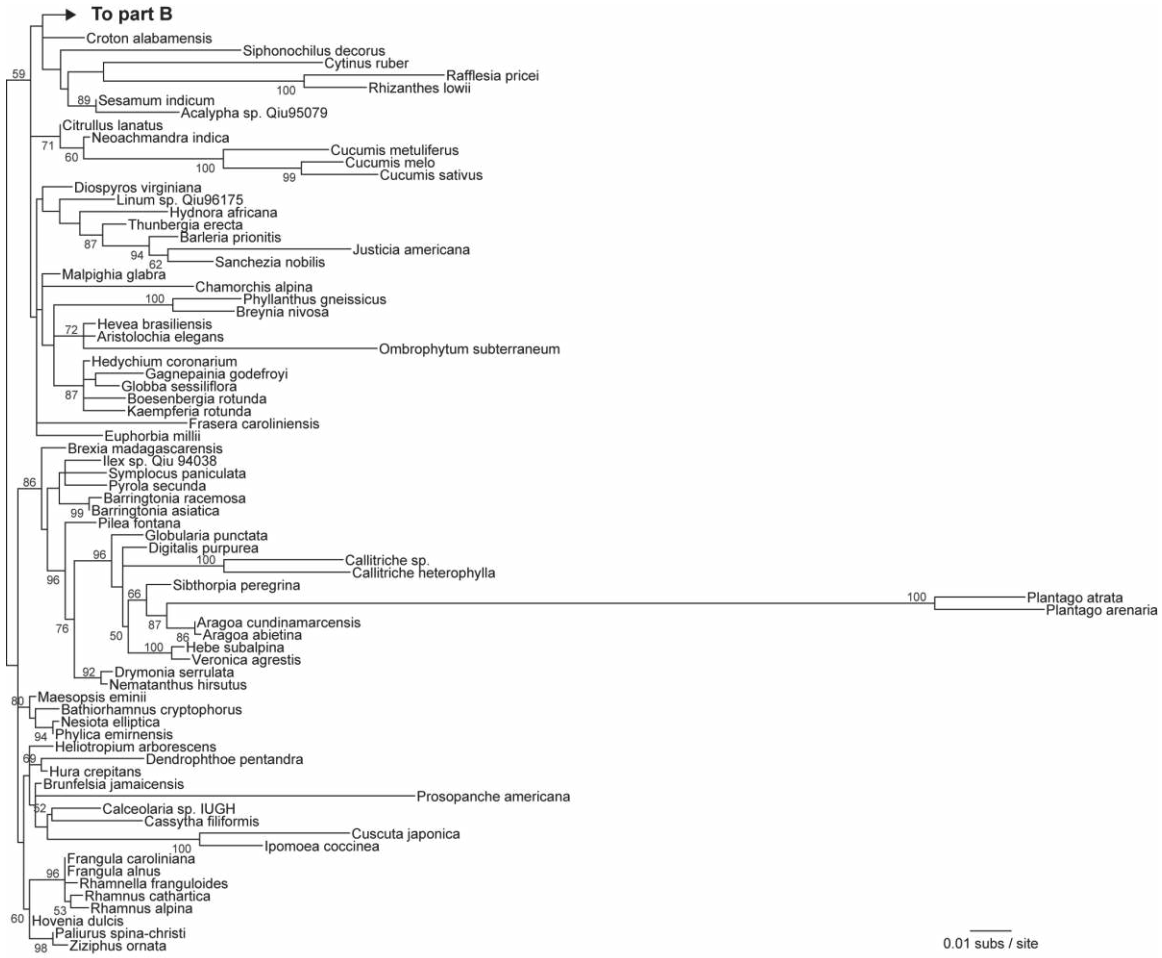

Figure S3. Labeled phylogenetic tree for the *cox1* intron. The depicted tree is identical to that in Figure 3, except that taxon names are included. The tree is split into two parts (A and B) to allow space for labeling of taxa. Orobanchaceae species are shown in bold blue text. Bootstrap values for all branches with >50% support are shown above, or in some cases below, the branch.


Lamium sp. Qiu 95015
Physostegia virginiana
Scutellaria mociniana
Oxera sp. MVSP-2007
Ajuga reptans
Clerodendrum trichotomum
Celsia arturus
Scrophularia nodosa
Paulownia tomentosa
Catalpa fargesii
Rehmannia glutinosa
Lindenbergia philippensis
Bartsia pedicularioides
Epifagus virginiana
Castilleja paramensis
Schwalbea americana
Phelipanche ramosa
Orobanche gracilis
Ixora sp. Qiu95051
Calycosiphonia macrochlamys
Coffea arabica
Ixora sp. Kew 21361
Nerium oleander
Hoya sikkimensis
Asclepias tuberosa
Alyxia loesneriana
Strychnos spinosa
Vinca rosea
Alstonia plumosa
Neisosperma brevitubum
Ochrosia elliptica
Maranta bicolor
Maranta leuconeura
Ctenanthe setosa
Monotagma laxum
Saranthe sp. Kress 96-5737
Saranthe leptostachya
Hydrocotyle rotundifolia
Polygala sanguinea
Polygala verticillata
Rhus glabra
Hybanthus sp. IND-JM3091
Shorea talura
Montrouziera cauliflora
Viola sp. Qiu95018
Dionaea muscipula
Hyoscyamus aureus
Hyoscyamus desertorum
Mandragora officinarum
Mandragora autumnalis
Mandragora sp. Kew23330
Przewalskia tangutica
Physochlaina infundibularis
Hyoscyamus boveanus
Hyoscyamus niger
Physochlaina orientalis
Hyoscyamus pusillus
Hyoscyamus muticus
Cynomorium coccineum
Mitrastema yamamotoi
Asimina triloba
Ehretia anacua
Lepionurus sylvestris
To part A
Pilostyles thurberi TX
Pilostyles thurberi AZ
Pholisma arenarium
Lennoa madreporoides
Musella lasiocarpa
Musa acuminata
Jasminum floridum
Jasminum polyanthum
Croton sp. Qiu 94027
Codiaeum peltatum
Melia toosendan
Philodendron oxycardium
Caladium bicolor
Peperomia polybotrya
Peperomia griseoargentea
Peperomia cubensis
Peperomia obtusifolia
Xanthosoma mafaffa
Xanthosoma sagittifolium
Amorphophallus rivieri
Zamioculcas zamiifolia
Pistia stratiotes
Steudnera discolor
Steudnera kerrii3
Colocasia esculenta
Steudnera griffithii
Remusatia vivipara
Steudnera kerrii4
Steudnera cf. discolor
Ariopsis protanthera
Typhonium albidinervum
Typhonium trilobatum
Pinellia ternata
Pinellia cordata
Typhonium giganteum
Typhonium hirsutum
Helicodiceros muscivorus
Dracunculus canariensis
Biarum davisii
Dracunculus vulgaris
Theriophonum infaustum
Biarum tenuifolium
Typhonium venosum
Eminium spiculatum
Arum dioscoridis
Arum concinnatum
Arum italicum
Arisaema triphyllum
Arisaema tortuosum
Arisaema speciosum
Alocasia navicularis
Alocasia cucullata
Alocasia gageana
Bursera simaruba
Dysoxylum canalense
Knema latericia
Myristica fragrans
0.01 subs / site

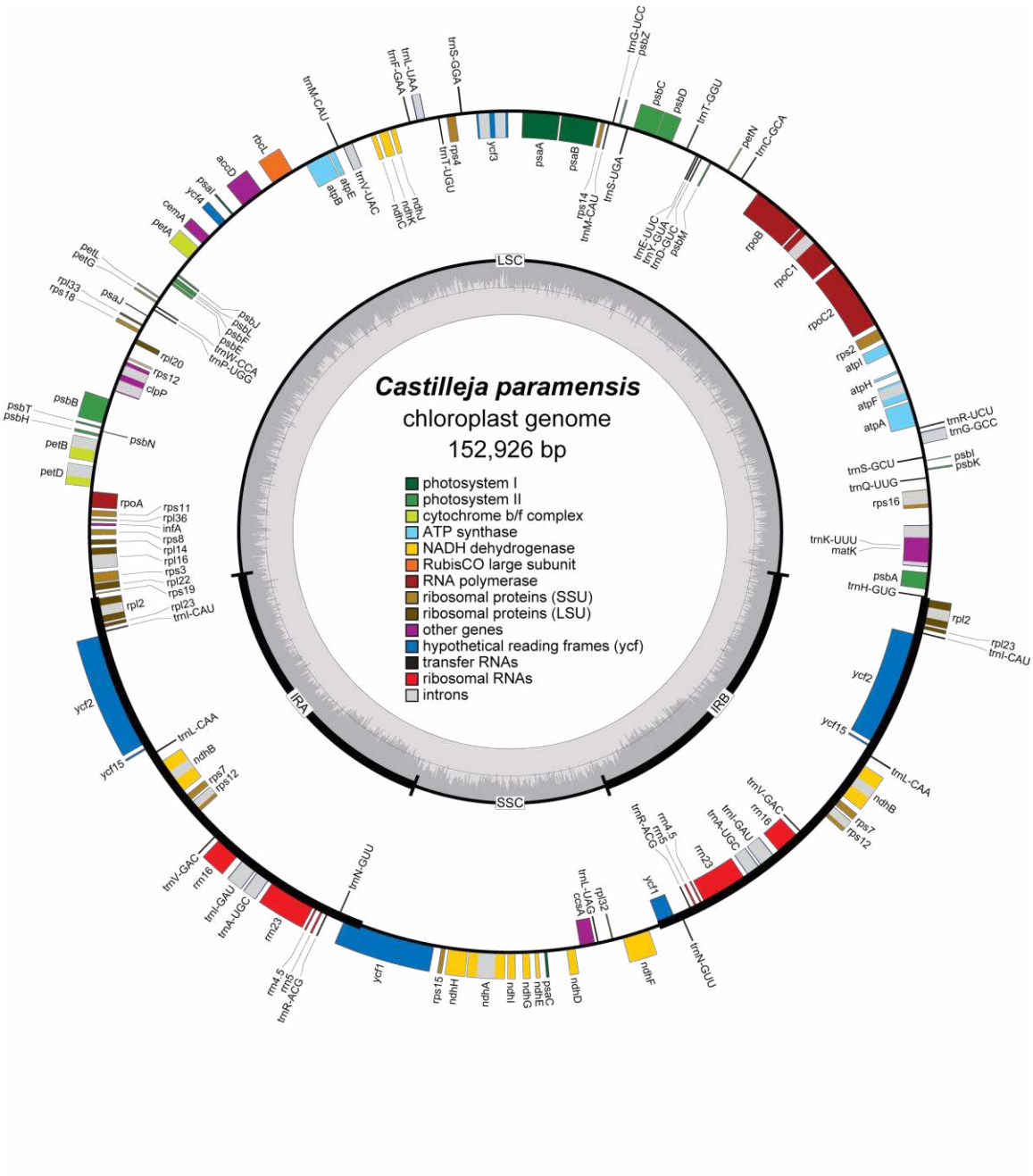

Figure S4. Circular genome map of the *Castilleja paramensis* plastome. Outer genes are transcribed counter-clockwise; inner genes are transcribed clockwise. Gene and intron colors correspond to the functional categories listed in the key. On the inner circle is shown GC content and the location of the inverted repeats (IRA, IRB) and single-copy regions (SSC, LSC). The map was drawn with OgDraw (http://ogdraw.mpimp-golm.mpg.de/).

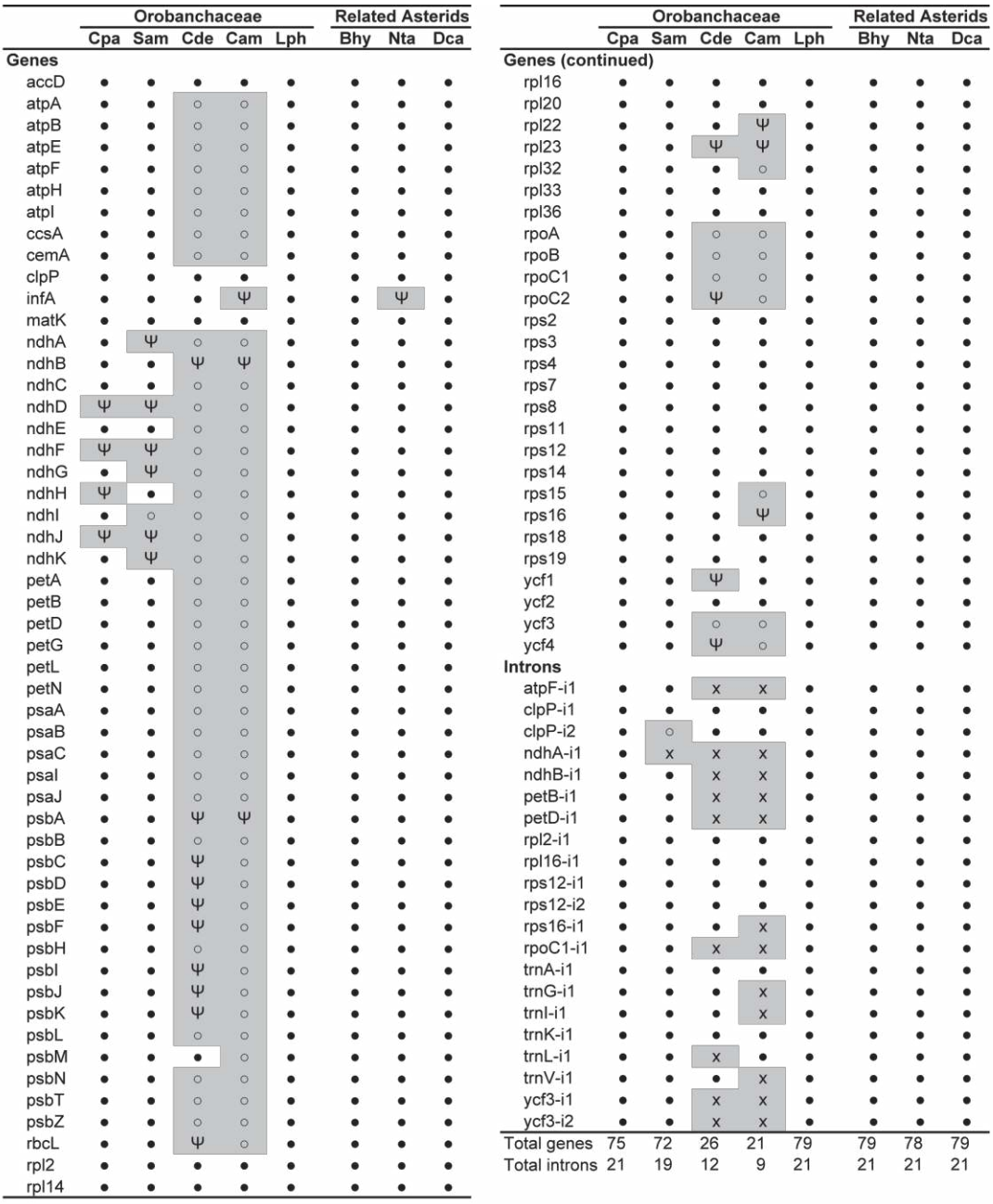

Figure S5. Plastome gene and intron content in plastomes from Orobanchaceae and selected asterids. Genes and introns present in each genome are marked with a filled circle ("●"). Lost genes and introns ("○"), pseudogenes ("ψ"), and missing introns due to loss or pseudogenization of the host gene ("x") are shaded gray. Cpa = *Castilleja paramensis*; Sam = *Schwalbea americana*; Cde = *Cistanche deserticola*; Cam = *Conophilis americana*; Lph = *Lindenbergia philippensis*; Bhy = *Boea hygrometrica*; Nta = *Nicotiana tabacum*; Dca = *Daucus carota*.

CHAPTER 4

Massive Loss of RNA Editing Sites from Mitochondrial Genes of *Welwitschia mirabilis*

Weishu Fan, Wenhu Guo, Lexis Funk and Jeffrey P. Mower

ABSTRACT

Analysis of mitochondrial genome diversity within gymnosperms has suggested an extensive loss of RNA editing sites in the xerophytic plant *Welwitschia mirabilis*. However, there is a lack of empirical data to confirm this loss of editing, and the mechanisms responsible for the loss of so many edit sites are still unclear. In order to gain insight on the abundance of mitochondrial RNA editing in *W. mirabilis* and examine potential mechanisms of edit site loss, we performed a comprehensive analysis of RNA editing by RT-PCR and cDNA sequencingof 29 protein-coding mitogenes. We found only 46 editing sites located in nine genes, which is substantially less than the 226 that were expected based on predicted data. Most of the editing sites were lost due either to genomic mutation or to gene conversion with a reverse-transcribed product (i.e. retroprocessing). The higher substitution rate in *Welwitschia* suggests that genomic mutation has led to some level of editing site loss. However, the nonrandom loss of editing in *ccmFc*, *mttB* and *nad7* and the correlated loss of editing sites and introns provide stronger evidence for the retroprocessing mechanism. Further studies will be required to determine the driving force of RNA editing loss in *Welwitschia*.

INTRODUCTION

In the mitochondrial genome (mitogenome) of land plants, a post-transcriptional process named RNA editing converts cytidines to uridines (C-to-U) at specific sites, and it could also involve in U-to-C editing in some species (Covello and Gray 1989, Hein et al. 2016, Hiesel et al. 1989, Maier et al. 1996, Shikanai 2006). RNA editing has been observed in all major groups of land plants (Chaw et al. 2008, Grewe et al. 2011, Oda et al. 1992, Unseld et al. 1997) suggesting a very early gain in land plant evolution (Malek et al. 1996), although the origins and the mechanism of this process are still unclear. Editing sites are preferentially located at first or second codon positions in protein-coding genes, whereas third position editing, which is silent because it doesn't alter the amino acid encoded by the codon, is much less frequent (Gray 2003). Some C-to-U RNA editing events can create start or stop codons, while U-to-C editing can remove the premature stop codons to restore internal codons (Brennicke et al. 1999, Carrillo et al. 2001). Thus, the identification of RNA editing is particularly important for the understanding of the genetic systems.

The frequency of RNA editing is very diverse among vascular plant, and several species have experienced a massive loss of editing sites. In lycophytes, editing was detected at over 2152 sites in the spike moss *Selaginella moellendorffii* (Hecht et al. 2011) and 1782 positions in the quillwort *Isoetes engelmannii* (Grewe et al. 2011),indicating a high frequency of RNA editing in this clade. However, another lycophyte, *Huperziasquarrosa*, was predicted to have only ~300-500 editing sites (Liu et al. 2012), suggesting a heavy reduction of editing in this species. In gymnosperms, based on the predicted 1214 sites in *Cycas taitungensis* (Chaw et al. 2008) and 1306 sites in *Ginkgo biloba* (Guo et al. 2016),

the ancestral number of RNA editing in this cluster was inferred to be around 1200-1300. In contrast, *Welwitschia mirabilis* was recently described as having extensive loss of RNA editing sites with only 226 predicted sites, which is less than 20% of the ancestral gymnosperm number (Guo et al. 2016). C-to-U RNA editing occurs less frequently in angiosperms, from 189 editing sites in *Silene noctiflora* (Sloan et al. 2010) to 835 sites in *Amborella trichopoda* (Rice et al. 2013). The ancestral count in angiosperms was inferred to be around 800 (Richardson et al. 2013), indicating that the 189 editing sites in *S. noctiflora* was due to a substantial loss of sites (Sloan et al. 2010).

Several mechanisms have been proposed for generating the variation in RNA editing frequency, particularly the massive loss of RNA editing, but they have not been fully clarified. First, mutation rate variation has been proposed as one explanation. High mutation pressure was postulated to drive lower editing rates due to the difficulty of maintaining the proper editing recognition sites under high mutation rates (Lynch et al. 2006). On the other hand, low mutation rates in plant mitochondria should exhibit more frequent RNA editing. In agreement with this model, species in *Pelargonium* (Geraniaceae) and *Silene* (Caryophyllaceae) have major increases in the mitochondrial synonymous substitution rate, and they have very few RNA editing sites (Parkinson et al. 2005, Sloan et al. 2010), while the extraordinarily slowly evolving species *Liriodendron tulipifera* is rich in RNA editing sites (Richardson et al. 2013). Also, in gymnosperms, the slowly evolving species *Ginkgo* and *Cycas* have high editing counts, while the faster evolving *Welwitschia* was predicted to have many fewer sites (Guo et al. 2016), suggesting that at least some of the edit loss in *Welwitschia* could be due to its increased substitution rate.

A second proposed mechanism for RNA editing loss is retroprocessing, which describes a process involving RNA-mediated gene conversion. Theoretically, this process could fully eliminate all edit sites and all introns from a gene. However, the available evidence suggests that this process usually affects only segments of a gene, driving the loss of a subset of introns together with the nearby editing sites (Itchoda et al. 2002, Sloan et al. 2010). The retroprocessing model predicts that multiple adjacent edit sites should frequently be lost together, but several statistical studies have provided little evidence that adjacent edit sites are more likely to be lost together than expected by random chance (Shields and Wolfe 1997, Sloan et al. 2010). Nevertheless, the clearest indications of retroprocessing come from the *rps3* gene in conifers (Ran et al. 2010) and the *cox2, nad1* and *nad4* genes in *Isoetes engelmannii* (Grewe et al. 2011), in which an intron and a large number of neighboring edit sites were eliminated. In addition, species in *Pelargonium* and *Welwitschia*, which have very few edit sites, also have a low number of introns (Guo et al. 2016, Park et al. 2015), consistent with expectations of retroprocessing. Overall, the generality of this process in driving edit site loss is unclear.

A third mechanism for edit site loss is the loss of recognition of an edit site by the RNA editing machinery. Editing sites recognition refers to the selection of the edited cytidines involving in the conversion from cytidines to uridines, which is highly specific (Hermann and Bock 1999). This recognition is substantially dependent on the 5' flanking RNA sequence (Mulligan et al. 1999). Analysis of the *rps12* and *rps3* transcript editing status suggested that the sufficient editing site recognition is critical to editing sites (Williams et al. 1998). Study in cauliflower also suggested that the nucleotide identity is important (Neuwirt et al. 2005). Although the editing complex is not fully defined, it is well

established that pentatricopeptide repeat (PPR) proteins are needed to site-specifically identify the sites to be edited (Barkan and Small 2014). Thus, loss of a particular PPR protein could result in the loss of recognition of an edit site although it has not been approved in *Welwitschia*.

Welwitschia mirabilis is the only species in the family of Welwitchiaceae (so the following text will use *Welwitschia* to represent *W. mirabilis*), and it is one of the three reported species in gymnosperm with a completed mitogenomic sequence (Chaw et al. 2008, Guo et al. 2016). Comparative analysis from the gymnosperm mitogenomes has shown some unusual features in *Welwitschia*: expanded size, numerous losses of protein and tRNA genes as well as introns, massive loss of RNA editing sites and higher substitution rate (Guo et al. 2016). However, the massive editing loss in *Welwitschia* was inferred using only predicted data, so experimental data is needed to confirm the editing sites loss. In this study, I have comprehensively examined the frequency of RNA editing in order to determine whether the massive editing loss is true, and if so, to explore whether the loss is due to accelerated substitution rates, retroprocessing, or a loss of editing recognition.

MATERIALS AND METHODS

Plant material and RNA extraction

Welwitschia was grown in the Beadle Center greenhouse at the University of Nebraska-Lincoln. Fresh leaf tissue was collected for RNA extraction. Total RNA was isolated using the TRIzol reagent (Life Technologies Corporation, U.S.A) according to the

suggested procedures provided by the manufacturer. To remove potential genomic DNA contamination, the isolated RNA was incubated with RNase-free DNase I (Thermo Fisher Scientific Inc. U.S.A) for 30 min at 37 °C. The reaction was terminated by adding EDTA (0.5M).

Reverse transcription, PCR amplification and sequencing

First-strand cDNA was generated from the isolated RNA by reverse transcription using random hexamers and M-MLV reverse transcriptase (Promega Corporation, U.S.A) in accordance with manufacturer's recommendations. A negative control sample, using the same amount of nuclease-free water instead of reverse transcriptase, was also prepared and used in cDNA construction. This control was used to test for potential DNA contamination in later analyses.

Reverse-transcription PCR (RT-PCR) assays were performed using the first-strand cDNA as template and degenerate primers. RT-PCR primers were designed to amplify all the protein-coding genes of *Welwitschia* (Table S1), taking care to exclude any predicted editing sites from the primer sequences to avoid enriching for unedited or partially edited transcripts (Mower and Palmer 2006). Two primers were designed and used for the upstream sequence of *mttB*, because no single best sequence was identified. Additional internal primer sets were designed for genes longer than 1 kb (*ccmFn, matR, rps3*) to assist in sequencing. The RT-PCR program settings were used as described before (Hepburn et al. 2012) except the annealing temperature for each reaction was set to 5°C below the lowest T_m of the particular pair of primers used.

RT-PCR products were directly sequenced on both strands at GenScript (NJ, GenScript USA Inc.). Sequences were assembled with CodonCode Aligner version 5.0.1 (CodonCode Corporation). Newly generated sequences will be deposited in GenBank, and additional sequences used in this study were extracted from GenBank (Table S2).

Identification of RNA editing sites

Both empirical and computational methods were used for identification of RNA editing sites in *Welwitschia*. RNA editing sites were empirically determined by comparing the aligned cDNA sequences with the DNA sequences (Hepburn et al. 2012, Rice et al. 2013, Richardson et al. 2013). Computational prediction data of RNA editing sites was from the PREP-Mt online server (Mower 2009), with a cutoff value of 0.2. For the *nad1* and *nad7* gene sequences, an expanded analysis of RNA editing used experimentally determined (*Ginkgo* and *Welwitschia*) and predicted (*Araucaria*, *Cycas*, *Gnetum* and *Pinus*) edit sites that were collected from six selected seed plants (Table S2).

Sequencing analysis

All *Welwitschia* mitochondrial coding genes were aligned either with its cDNA sequence (for identification of RNA editing sites) or with gene sequences from other species (for determination of RNA editing loss) using MUSCLE version 3.8.31 (Edgar 2004) with default parameters. When necessary, alignments were adjusted manually. To better understand whether the intron loss is associated with RNA editing site loss, the intron positions were also manually inserted into the alignments of intron-containing genes using BioEdit (Hall 1999). To calculate the different evolutionary transitions that result in editing site loss, the nucleotide status of each species was counted for every position in

the alignment that was edited in at least one species. This analysis was based on alignments for the 21 protein-coding genes that were present in all species and could be unambiguously aligned. The other genes were excluded because data for some species were missing or the genes were too divergent to be aligned.

In order to test whether the loss of editing sites is random (as expected in a mutation rate model) or clustered (as expected in a retroprocessing model), the probability of editing sites in specific group (every 250 bp was designed as a group) was calculated. The editing site variance caused by intron loss or retention was also estimated by divided all the editing sites into two groups: intron effect region (intron sites ± 100 bp) and non-effect region. A chi-square test was then applied based on the expected number and the observed number.

All topologies used in this study were constructed according to relationships defined on the Angiosperm Phylogeny Website, version 12 (http://www.mobot.org/MOBOT/research/APweb/).

RESULTS

Low levels of RNA editing in *Welwitschia mirabilis* mitochondrial genes

By comparison of mitochondrial cDNA and genomic sequences, a total number of 46 C-to-U RNA editing sites were identified in nine genes (Table 4-1, Table S3) out of the 29 protein-coding genes in the *Welwitschia* mitogenome (Table 4-2), and no U-to-C edited sites were detected. For each editing site, we summarized the gene location, the codon position, the codon sequence, and encoded amino acid (Table 4-1, Table S3). The editing

events were predominantly located in *ccm* genes (45.7%) and *nad* genes (23.9%) in *Welwitschia*. The maturase gene (*matR*) and membrane transport protein gene (*mttB*) also exhibited a relatively high editing count, representing 10.9% of the total editing sites. In terms of the editing locations, the second codon position possessed a very large portion (76.1%), whereas third position edits were very infrequent (4.3%). The RNA editing events did not create any start or stop codons.

Prediction of C-to-U editing sites was also conducted using PREP-Mt online tool (Mower 2009) with a cutoff of 0.2. Surprisingly, the number of editing sites determined from experimental data was much less than from the prediction. In total, 226 edit sites were predicted (Guo et al. 2016). Of these, 171 editing sites were located in the regions empirically examined by amplified RT-PCR products, whereas only 46 sited were experimentally identified (Table 4-2). The 73% difference between predicted and observed editing sites is affecting almost all genes, where empirical editing counts are consistently lower than predicted counts. Specifically, no editing sites were detected for all five *atp* genes, including *atp4* which was predicted to contain 11 editing sites. Similarly, the *ccmFn* and *rps3* genes were predicted to have 15 and 13 editing sites, respectively, but no edit sites were detected by RT-PCR. The total number of editing sites in *Welwitschia* is lower in comparison with all other vascular plants that have been examined to date.

Low editing levels are due to extensive loss of RNA editing sites from *Welwitschia* mitogenes

In a previous study, the low level of editing in *Welwitschia* was suggested to be due to a massive loss of editing relative to the common ancestor of gymnosperms, which was inferred to be rich in introns and edit sites (Guo et al. 2016). We compared the number of RNA editing sites in *Welwitschia* with the other cDNA sequences in gymnosperm (*Cycas* and/or *Ginkgo*), which confirmed the massive RNA editing loss in this lineage (Figure 4-1). Among all 29 mitochondrial genes in *Welwitschia*, only nine genes exhibit RNA editing (Figure S1). This is much lower than in homologous genes from other species in gymnosperms. For *ccmB*, which has relatively more editing sites (10 sites in the examined region) compared with other *Welwitschia* genes, *Cycas* still contains over three times more editing sites. In fact, for every *Welwitschia* gene, empirical editing counts are lower compared with *Cycas* or *Ginkgo*.

To further assess the evolution of editing in *Welwitschia* compared with other gymnosperms, the status of editing in the *nad1* and *nad7* genes were shown in a phylogenetic context (Figure 4-2). The editing sites in *Cycas, Ginkgo*, and *Pinus* are all abundant and largely shared, indicating that the ancestral state of gymnosperm was probably rich in editing sites (Guo et al. 2016). In contrast, these two genes show that the massive RNA editing loss occurred in the *Welwitschia* lineage. Taking into account the topology, *Gnetum* also displayed less editing sites than the other taxa although not as low as *Welwitschia*, suggesting that some of this RNA editing loss began in the common ancestor of *Gnetum* and *Welwitschia*. However, since no empirical data available for *Gnetum*, this hypothesis need to be further tested.

Mechanisms of editing loss in *Welwitschia*

To further investigate the mechanism of editing sites loss, the pattern of edit site loss was examined in order to distinguish among the three possibilities: (1) Loss of editing site recognition by the machinery that performs RNA editing, (2) Genomic mutation, and (3) retroprocessing. If the recognition of the edit site is lost, we would observe a change in the gene sequence from an edited C to an unedited C (Table 4-3, E>C). A genomic mutation would appear as a change from an edited C to another nucleotide (adenine (A), guanine (G), or T). These mutations would have a random distribution in the gene, and would have no correlation to the positions of introns (E>R or E>T). Retroprocessing would look like a change from an edited C to a T in the gene (E>T). Because retroprocessing is a gene conversion process, it should tend to remove edit sites and introns nonrandomly. That is, adjacent edit sites and introns should be preferentially eliminated, leaving behind clusters of retained edit sites and introns.

The loss of editing recognition was identified by calculating the change as E>C. The overall percentage of E>C in *Welwitschia* is 0.16 which is the lowest proportion comparing with other selected seed plants (Table 4-3). In fact, this process is about two-fold lower when compared to the other gymnosperms *Cycas* and *Ginkgo*. These results indicate that the loss of RNA editing recognition has occurred for some sites, but it happens less frequently in *Welwitschia* than in other seed plants. In other words, this does not appear to be the main mechanism for editing site loss in *Welwitschia*.

The evidence of editing sites loss involving the genomic mutation model was examined by calculating the nucleotide changes from edited C to A or G (E>R).

*Welwitschia*possesses the highest percentage of loss (0.07) compared with other species (Table 4-3). The higher frequency of E>R changes in *Welwitschia* suggests that loss of RNA editing sites due to genomic mutation may be slightly higher in *Welwitschia* compared with other species. However, in all species, E>R mutations make up only a small fraction of all lost edit sites.

The most common type of substitution frequency resulting in a loss of editing is the change from an edited C to T (E>T). This type of substitution could be due to either retroprocessing or genomic mutation. Compared with other gymnosperms, *Welwitschia* has a substantially higher value (0.77). The *Welwitschia* frequency is more similar to the values observed in angiosperms. Because angiosperms have a lower frequency of editing compared to other vascular plants, it is possible that the mechanisms reducing editing counts in *Welwitschia* are similar to angiosperms.

In order to distinguish genomic mutation from retroprocessing, the distribution of the lost edit sites were examined. Loss of editing sites by genomic mutation should be random, which means that the edit sites retained in a gene should also have a random distribution. In contrast, retroprocessing should remove clusters of edit sites, which should also result in a nonrandom clustering of sites that were retained. For several genes, such as *ccmFc, mttB* and *nad7*, the pattern of editing loss appears to be nonrandom (Figure 4-1, Figure S1). Three of the retained sites in *ccmFc* are located at the 3' end of the sequence, and the probability of three retained sites clustering together is significantly lower than expected for a random distribution (P=0.0046). Similarly, the five retained editing sites in *mttB* are nonrandomly clustered at the 5' end of the gene while the rest of the gene lost all editing sites (P=0.0041). Likewise, the editing site distribution of the *nad7* gene is also

nonrandomly clustered (P=0.0002) at the 3' end. All of these genes with clustered edit sites are consistent with retroprocessing but not expected for the other loss models. Because this test requires the retention of several edit sites, most of the other *Welwitschia* genes cannot be tested due to their loss of most or all edit sites.

Retroprocessing is also predicted to lose introns together with large scale editing sites. Thus, a correlated loss of editing and introns provides evidence of the retroprocessing mechanism, whereas genomic mutation should show no such correlation. For those intron-containing genes, *Welwitschia* has lost most if not all editing sites in combination with a loss of several introns. In *cox2*, *nad2*, *nad4*, and *rps3*, all edited sites were lost together with some level of intron loss (Figure 4-1, Figure S1). For the other four intron-containing genes (*ccmFc*, *nad1*, *nad5* and *nad7*) a few editing sites were retained even though they also had the highly reduced number of editing, and they also experienced one or two intron losses (Figure 4-1, Figure 4-2). A similar pattern of intron and editing loss was also observed for *Araucaria* and *Gnetum nad7* based on the evolutionary view of editing loss (Figure 4-2). However, a chi-square test for the correlation between intron loss and edit site loss in three genes (*ccmFc*, *nad1* and *nad7*) of *Welwitschia* did not show significant difference with the P value of 0.3102, 0.5247 and 1, respectively.

DISCUSSION

Massive loss of mitochondrial RNA editing in *Welwitschia*

Our previous study on the mitogenomes of gymnosperms indicated a dramatic loss of RNA editing sites in *Welwitschia* relative to the ancestral high level of editing in

gymnosperms. In this study, empirical data confirmed that RNA editing is very low in *Welwitschia*, and surprisingly, even lower than the predicted number. Within the 29 functional protein-coding genes in *Welwitschia* mitogenome, RNA editing sites were detected from only nine of them. Three of the genes were predicted to have no editing sites (*atp6*, *atp9* and *nad9*), and two of them were confirmed by cDNA sequencing results (the third gene, *atp6*, could not be amplified by RT-PCR). Another 17 genes, although they were predicted to have some level of editing, had no detectable edit sites by the reverse transcription sequencing approach. RNA editing loss from a single gene has been reported in seed plants; for example, the *cox3* and *rps13* lacked RNA editing in Iridaceae and Amaryllidaceae (Lopez et al. 2007). However, such a massive RNA editing loss from most of the genes in a seed plant mitogenome has not been reported yet.

The discrepancy between empirical results and predicted results was surprising. PREP-Mt was used to predict the RNA editing sites, which has been shown to be accurate in previous studies (Guo et al. 2016, Mower 2009). The reason this program did not perform well for *Welwitschia* is unclear, but it could be related to the high substitution rate or to the highly lineage-specific loss of editing in this species. In particular, genes such as *ccmFc*, *ccmFn*, *matR*,and *rps3* were very difficult to align to other homologs due to their high level of divergence, and they had some of the worst prediction results, suggesting a negative correlation between substitution rate and editing prediction accuracy.

In *Gnetum*, the *nad1* and *nad7* aligned results in this study also suggested the RNA editing loss. Prediction analysis was performed for all genes in this species, suggesting a decrease in editing sites, although not to the same degree as in *Welwitschia* (data not shown). Given the inaccuracy for *Welwitschia* prediction results, the lack of cDNA or

transcriptome data from *Gnetum* makes it difficult to evaluate whether the prediction is accurate in this species. It may also prove to have fewer edit sites than suggested by the prediction. Accordingly, the loss of RNA editing in both *Gnetum* and *Welwitschia* suggests that these lineage-specific losses of editing are problematic for PREP-Mt accuracy.

Retroprocessing model for the loss of RNA editing sites

Several different evolutionary models have been proposed for the loss of RNA editing in plants, including the loss of editing site recognition, increased substitution rates, and RNA-mediated gene conversion (retroprocessing). A retroprocessing model is most often cited in the literature to explain some of the RNA editing loss (Grewe et al. 2011, Ran et al. 2010, Sloan et al. 2010), resulting in preferential C-to-T substitution at RNA editing sites. Generally, retroprocessing is mentioned because one or more intron losses occurred along with the surrounding editing sites loss, which is most easily attributable to the effect of retroprocessing. The other two models do not predict any correlation between intron loss and editing loss. The retroprocessing model also predicts a clustered loss of editing sites, although this expectation has previously not been supported with statistical analyses (Shields and Wolfe 1997, Sloan et al. 2010).

In *Welwitschia*, there are three genes (*ccmFc*, *mttB*, *nad7*) that showed a significantly nonrandom distribution of edit sites, consistent with retroprocessing (Figure 4-1, Figure S1). In addition, there are four genes, including *ccmFc*, *nad1*, *nad5* and *nad7*, that have lost introns and surrounding edit sites, which could be best explained by retroprocessing. For example, the phylogenetic distribution of introns and editing sites in *nad7* (Figure 4-

2B) shows that the first intron was lost in three lineages (*Araucaria*, *Gnetum* and *Welwitschia*), together with all adjacent editing sites. In addition, the third intron is lacking in *Araucaria* and *Welwitschia*, and the region around this intron has also experienced dramatic editing loss in these two species, from 5 (compared with *Cycas*) to 11 (compared with *Pinus*). Thus, the retroprocessing mechanism seems to be a good explanation for editing sites loss from the *nad7* gene of all three species due to the simultaneous loss of introns and flanking editing sites. However, a second statistical analysis looking for a significant association between intron loss and editing sites was not significant for these genes in *Welwitschia*. It may be due to the very few edit sites remaining in these genes, which may be too low for the chi-square test, which generally requires at least five data points. .Thus, RNA editing loss in *Welwitschia* seems to be attributable to retroprocessing, although this conclusion is tentative until more data becomes available.

In contrast, there is little unambiguous support for the other models. E to C rates are lower in *Welwitschia* than all other seed plants examined, arguing against a major role for the loss of editing recognition. E to R rates are slightly higher in *Welwitschia* compared with other species. This suggests that the higher substitution rate in *Welwitschia* is leading to an increased rate of genomic mutation to eliminate edit sites. Unfortunately, for most genes, the loss of editing is so extensive that is not possible to distinguish between expected patterns for genomic mutation and retroprocessing. Further studies of RNA editing patterns in other gneotphyte species (*Gnetum* and *Ephedra*) are required to help determine whether the editing loss is more strongly correlated with intron loss or substitution rates.

ACKNOWLEDGMENTS

We thank Qian Du for the enjoyable discussion on the data analysis. We also thank Amy Hilske and Samantha Link for care of *Welwitschia* plants in the Beadle Center greenhouse. This research was supported by the National Science Foundation (to JPM), a fellowship from China Scholarship Council (to WF) and by the UCARE program (to LF).

REFERENCES

Barkan A, Small I. 2014. Pentatricopeptide repeat proteins in plants. Annu. Rev. Plant Biol 65: 415-442.

Brennicke A, Marchfelder A, Binder S. 1999. RNA editing. FEMS Microbiol. Rev. 23: 297-316.

Carrillo C, Chapdelaine Y, Bonen L. 2001. Variation in sequence and RNA editing within core domains of mitochondrial group II introns among plants. Mol Gen Genet 264: 595-603.

Chaw S-M, Shih AC-C, Wang D, Wu Y-W, Liu S-M. 2008. The mitochondrial genome of the gymnosperm Cycas taitungensis contains a novel family of short interspersed elements, Bpu sequences, and abundant RNA editing sites. Mol Biol Evol 25: 603-615.

Covello PS, Gray MW. 1989. RNA editing in plant mitochondria.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.

Gray M. 2003. Diversity and evolution of mitochondrial RNA editing systems. IUBMB life 55: 227-233.

Grewe F, Herres S, Viehover P, Polsakiewicz M, Weisshaar B, Knoop V. 2011. A unique transcriptome: 1782 positions of RNA editing alter 1406 codon identities in mitochondrial mRNAs of the lycophyte Isoetes engelmannii. Nucleic Acids Res 39: 2890-2902.

Guo W, Grewe F, Fan W, Young GJ, Knoop V, Palmer JD, Mower JP. 2016. Ginkgo and Welwitschia Mitogenomes Reveal Extreme Contrasts in Gymnosperm Mitochondrial Evolution. Mol Biol Evol 33: 1448-1460.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Pages 95-98. Nucleic acids symposium series.

Hecht J, Grewe F, Knoop V. 2011. Extreme RNA editing in coding islands and abundant microsatellites in repeat sequences of Selaginella moellendorffii mitochondria: the root of

frequent plant mtDNA recombination in early tracheophytes. Genome Biol. Evol. 3: 344-358.

Hein A, Polsakiewicz M, Knoop V. 2016. Frequent chloroplast RNA editing in early-branching flowering plants: pilot studies on angiosperm-wide coexistence of editing sites and their nuclear specificity factors. BMC Evol. Biol 16: 1.

Hepburn NJ, Schmidt DW, Mower JP. 2012. Loss of two introns from the Magnolia tripetala mitochondrial cox2 gene implicates horizontal gene transfer and gene conversion as a novel mechanism of intron loss. Mol Biol Evol 29: 3111-3120.

Hermann M, Bock R. 1999. Transfer of plastid RNA-editing activity to novel sites suggests a critical role for spacing in editing-site recognition. Proc Natl Acad Sci USA 96: 4856-4861.

Hiesel R, Wissinger B, Schuster W, Brennicke A. 1989. RNA editing in plant mitochondria. Science 246: 1632-1634.

Itchoda N, Nishizawa S, Nagano H, Kubo T, Mikami T. 2002. The sugar beet mitochondrial nad4 gene: an intron loss and its phylogenetic implication in the Caryophyllales. Theor Appl Genet 104: 209-213.

Liu Y, Wang B, Cui P, Li L, Xue J-Y, Yu J, Qiu Y-L. 2012. The mitochondrial genome of the lycophyte Huperzia squarrosa: the most archaic form in vascular plants. PLoS One 7: e35168.

Lopez L, Picardi E, Quagliariello C. 2007. RNA editing has been lost in the mitochondrial cox3 and rps13 mRNAs in Asparagales. Biochimie 89: 159-167.

Lynch M, Koskella B, Schaack S. 2006. Mutation pressure and the evolution of organelle genomic architecture. Science 311: 1727-1730.

Maier RM, Zeitz P, Kössel H, Bonnard G, Gualberto JM, Grienenberger JM. 1996. RNA editing in plant mitochondria and chloroplasts. Pages 343-365. Post-Transcriptional Control of Gene Expression in Plants.

Malek O, Lättig K, Hiesel R, Brennicke A, Knoop V. 1996. RNA editing in bryophytes and a molecular phylogeny of land plants. The EMBO journal 15: 1403.

Mower JP. 2009. The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. Nucleic Acids Res 37: W253-W259.

Mower JP, Palmer JD. 2006. Patterns of partial RNA editing in mitochondrial genes of Beta vulgaris. Mol Genet Genomics 276: 285-293.

Mulligan R, Williams M, Shanahan M. 1999. RNA editing site recognition in higher plant mitochondria. J. Hered 90: 338-344.

Neuwirt J, Takenaka M, van der Merwe JA, Brennicke A. 2005. An in vitro RNA editing system from cauliflower mitochondria: editing site recognition parameters can vary in different plant species. RNA 11: 1563-1570.

Oda K, Yamato K, Ohta E, Nakamura Y, Takemura M, Nozato N, Akashi K, Kanegae T, Ogura Y, Kohchi T. 1992. Gene organization deduced from the complete sequence of liverwort Marchantia polymorpha mitochondrial DNA: a primitive form of plant mitochondrial genome. J. Mol. Biol 223: 1-7.

Park S, Grewe F, Zhu A, Ruhlman TA, Sabir J, Mower JP, Jansen RK. 2015. Dynamic evolution of Geranium mitochondrial genomes through multiple horizontal and intracellular gene transfers. New Phytol 208: 570-583.

Parkinson CL, Mower JP, Qiu YL, Shirk AJ, Song K, Young ND, DePamphilis CW, Palmer JD. 2005. Multiple major increases and decreases in mitochondrial substitution rates in the plant family Geraniaceae. BMC Evol Biol 5: 73.

Ran J-H, Gao H, Wang X-Q. 2010. Fast evolution of the retroprocessed mitochondrial rps3 gene in Conifer II and further evidence for the phylogeny of gymnosperms. Mol. Phylogenet. Evol 54: 136-149.

Rice DW, et al. 2013. Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm Amborella. Science 342: 1468-1473.

Richardson, Rice DW, Young GJ, Alverson AJ, Palmer JD. 2013. The "fossilized" mitochondrial genome of Liriodendron tulipifera: ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. BMC Biol 11: 29.

Shields DC, Wolfe KH. 1997. Accelerated evolution of sites undergoing mRNA editing in plant mitochondria and chloroplasts. Mol Biol Evol 14: 344-349.

Shikanai T. 2006. RNA editing in plant organelles: machinery, physiological function and evolution. Cell Mol Life Sci 63: 698-708.

Sloan DB, MacQueen AH, Alverson AJ, Palmer JD, Taylor DR. 2010. Extensive loss of RNA editing sites in rapidly evolving Silene mitochondrial genomes: selection vs. retroprocessing as the driving force. Genetics 185: 1369-1380.

Unseld M, Marienfeld JR, Brandt P, Brennicke A. 1997. The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924. Nat genet 15: 57-61.

Williams MA, Kutcher BM, Mulligan RM. 1998. Editing site recognition in plant mitochondria: the importance of 5'-flanking sequences. Plant Mol Biol 36: 229-237.

TABLES AND FIGURES

Table 4-1. Summary of RNA editing sites in *Welwitschia mirabilis* mitochondrial genes

		Number	Percentage
Total C-to-U		46	100.0
	ccmB	14	30.4
	ccmC	3	6.5
	ccmFc	4	8.7
	matR	5	10.9
	mttB	5	10.9
	nad1	3	6.5
	nad5	1	2.2
	nad7	7	15.2
	rps4	4	8.7
Coding			
	1st	9	19.6
	2nd	35	76.1
	3rd	2	4.4

Table 4-2. Accuracy of edit site prediction in *Welwitschia mirabilis*

Gene	Predicted*	Observed	O-P	Gene	Predicted*	Observed	O-P
atp1	2	0	-2	*nad1*	4	3	-1
atp4	8	0	-8	*nad2*	2	0	-2
atp6	0	-	0	*nad3*	2	0	-2
atp8	6	0	-6	*nad4*	1	0	-1
atp9	0	0	0	*nad4L*	4	0	-4
ccmB	19	14	-5	*nad5*	4	1	-3
ccmC	6	3	-3	*nad6*	12	0	-12
ccmFc	11	4	-7	*nad7*	8	7	-1
ccmFn	15	0	-15	*nad9*	0	0	0
cob	3	0	-3	*rpl10*	6	0	-6
cox1	2	0	-2	*rps3*	13	0	-13
cox2	2	0	-2	*rps4*	9	4	-5
cox3	1	0	-1	*rps12*	0	0	0
matR	17	5	-12	*sdh4*	2	0	-2
mttB	12	5	-7	**Total**	171	46	

* Predicted counts are taken from the regions that were amplified for empirical analysis

Table 4-3. Substitution frequencies at edited sites in mitochondrial genes of seed plants

		Angiosperms					Gymnosperms		
Process	Change	*Amborella*	*Arabidopsis*	*Beta*	*Liriodendron*	*Nicotiana*	*Cycas*	*Ginkgo*	*Welwitschia*
Loss of edit site recognition	E>C	0.24	0.19	0.18	0.27	0.20	0.31	0.34	0.16
Genomic mutation or retroprocessing	E>T	0.72	0.78	0.79	0.70	0.78	0.64	0.61	0.77
Genomic mutation	E>R	0.04	0.02	0.02	0.03	0.02	0.05	0.05	0.07

E refers to the editing sites

R refers to either adenine (A) or guanine (G)

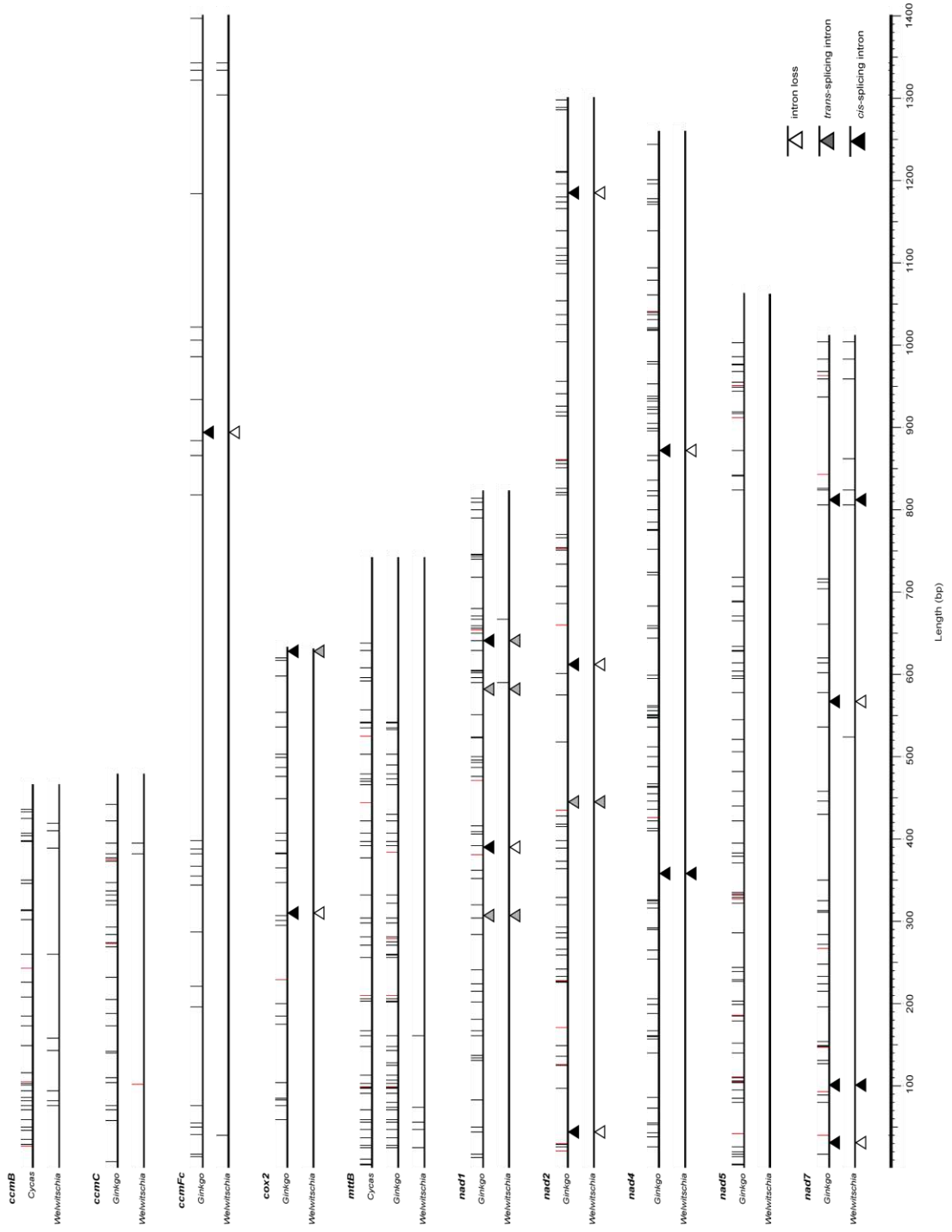
ccmB
Cycas
Welwitschia
ccmC
Ginkgo
Welwitschia
ccmFc
Ginkgo
Welwitschia
cox2
Ginkgo
Welwitschia
mttB
Cycas
Ginkgo
Welwitschia
nad1
Ginkgo
Welwitschia
nad2
Ginkgo
Welwitschia
nad4
Ginkgo
Welwitschia
nad5
Ginkgo
Welwitschia
nad7
Ginkgo
Welwitschia
intron loss
trans-splicing intron
cis-splicing intron
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
Length (bp)

Figure 4-1. Loss of RNA editing sites and introns in selected *Welwitschia* mitochondrial genes. Vertical lines indicate RNA editing sites. Non-silent and silent C-to-U editing are displayed in black and red, respectively. Selected gene names and species names are shown on the left. For each gene, only the amplified shared region was displayed and the length bar show at the bottom. This figure was generated by PREPACT with default graphic tool option.

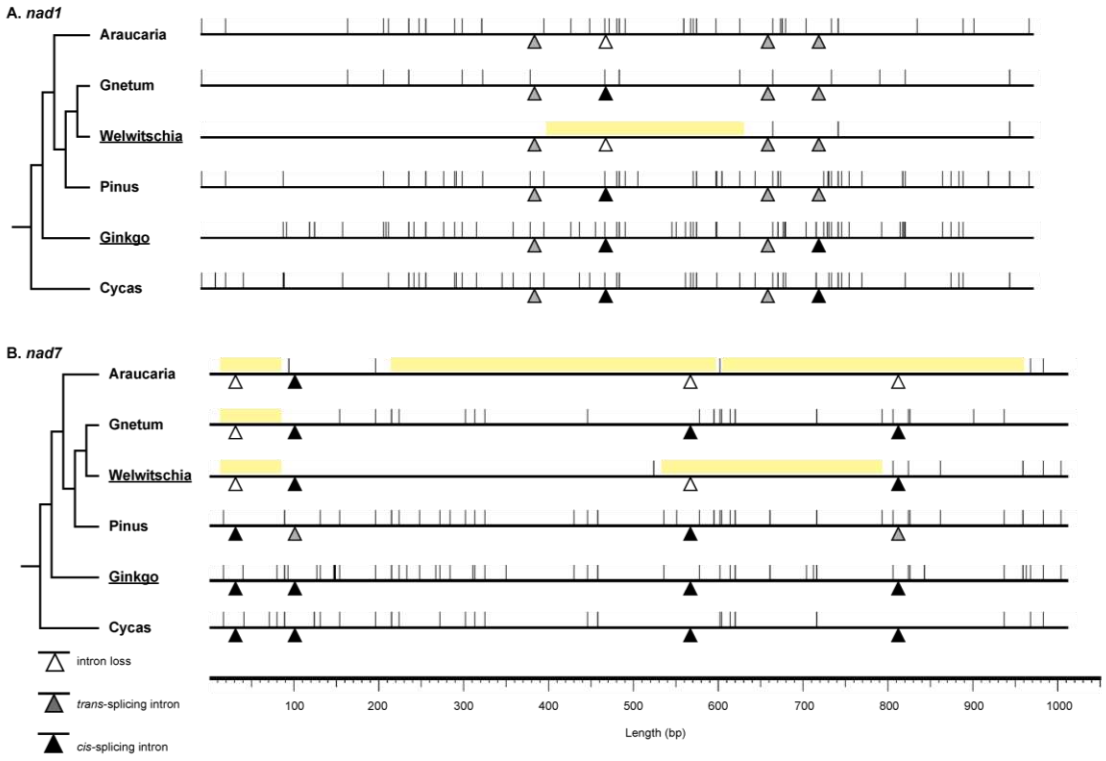

Figure 4-2. Phylogenetic distribution of *nad1* and *nad7* introns and RNA editing sites in selected seed plants. The editing sites are indicated by vertical lines. Empirical RNA editing (underlined) was used for two species and other species used predicted data by PREP-mt (Mower 2009). The yellow shadow regions are intron absent, coinciding with regions lacking RNA editing sites. The scale bar was drawn at the bottom. The phylogenetic relationships are based on angiosperm phylogeny website (http://www.mobot.org/mobot/research/apweb/).

SUPPORTING INFORMATION

Table S1. Gene specific primers for *Welwitschia*

Name	Sequence
atp1_Wel_F	ATGGTGAACAACTTGGAGAGC
atp1_Wel_R	GGTCCTTACTAATGGTGTCAC
atp4_Wel_F	GCTGCTATTCTCTTGATAAGTG
atp4_Wel_R	ATAGAGCCATTCGTGATACC
atp8_Wel_F	TTCTTCTGGTTATGCCTG
atp8_Wel_R	TTACGGTACGATGTGAAC
atp9_Wel_F	GCTGCTGTTGGTATTGGAAACG
atp9_Wel_R	TCAGAAACCCCATCATTAAGG
ccmB_Wel_F	GTATATCGTAGTAACGCCCTT
ccmB_Wel_R	GCCATTTTTTCTTGTAAACC
ccmC_Wel_F	CCCAACCTTCTGTATTTATG
ccmC_Wel_R	GTAAAGGAATAGGAAGTTAGCC
ccmFc_Wel_F	AGCACCCGTACTCTTGAAATGG
ccmFc_Wel_R	ACAAACTACACAAGCCTCC
ccmFn_Wel_F1	ATGGAAGGGTGTTTTCGTTGC
ccmFn_Wel_F2	CGCTCCCGAAACGAAGGTTC
ccmFn_Wel_F3	GCTACGCAATAGAAGCCT
ccmFn_Wel_R1	GGAGCCTTCCCTCATTTTGAA
ccmFn_Wel_R2	AAACATATCGGCTTCGTAGTGG
ccmFn_Wel_R3	CTGAGTTACCCCTTCTGCTCT
cob_Wel_F	TCTATTCTCAAACAACCC
cob_Wel_R	TCCAACTCGTCCCAGAAT
cox1_Wel_F	CCACTAACCACAAGGATATAGG
cox1_Wel_R	CTATGGATTGATAGAGGTCTCC
cox2_Wel_F	ACTTGCTGCTTACTCTCC
cox2_Wel_R	GACGAGTTTATTGGATACCC
cox3_Wel_F	ATGGTAGCAGAACAGAAGAGGC
cox3_Wel_R	TCATAGACCTCCCCACCAATAG
matR_Wel_F1	GGTTGAAGTTTAGACCGCTAAC
matR_Wel_F2	GCTCCGCAGGATCAACAA
matR_Wel_R1	GTGGGGAACGACTTCTAC
matR_Wel_R2	CTGATATAGGGTCTTGTACGC
mttB_Wel_F1	TTGCATTGAAAACTCCTC
mttB_Wel_F2	GGTCTTAGTTTGACATGGTT
mttB_Wel_R	ATAGTTCGCACCCGAGGC

Name	Sequence
nad1_Wel_F	GGACATTGCTATATACGCCTC
nad1_Wel_R	CGGTATTACTTGAAAGGTGACC
nad2_Wel_F	GGAGTTGTATTTAGTACCTCG
nad2_Wel_R	TCACAGATAGAAACTTAGTGCC
nad3_Wel_F	GTTTGCTAGTTTCTTTGATC
nad3_Wel_R	AGCACCCTTCTTCCATTC
nad4_Wel_F	CGACTTATGTCAGAATGCTC
nad4_Wel_R	GCACTAAGTTACCTACGGA
nad4L_Wel_F	GCTCACAACTACAATGAAGGC
nad4L_Wel_R	TCTACTGCGATGGTCCCT
nad5_Wel_F	GGAACCGCCATAGTAACCACC
nad5_Wel_R	AGCACCTTTGTCTAATATGACG
nad6_Wel_F	CCTATTTATGGTCTCTGGGTTG
nad6_Wel_R	TTCGCATATCGCCATGTC
nad7_Wel_F	ACCTCAACATCCTGCTGC
nad7_Wel_R	CTATCTACCTCTCCAAACAC
nad9_Wel_F	CTGAGAGATACTTTACCCAA
nad9_Wel_R	GACCGCCATAAAGTACCA
rpl10_Wel_F	GAGTCAAACCCAGGACAC
rpl10_Wel_R	CTATCGTTCGTTCCCGCTTC
rps12_Wel_F	CTTTGGAGAAATGTCCTCAG
rps12_Wel_R	GGATCTGCTACTTTTTCG
rps3_Wel_F1	CAAGATGTGAATCAACGAG
rps3_Wel_F2	CGATAAGCGAAGCAACAAT
rps3_Wel_F3	GAGCGGCGATCATATCAAGC
rps3_Wel_R1	GCCAATATGCTTCTGTTCAA
rps3_Wel_R2	AACACTGGACCGCTGGAA
rps3_Wel_R3	CACGCTGCTATATGAGATCCAC
rps4_Wel_F	CTGACAAGAATACAACGCCGCA
rps4_Wel_R	CTCTGAGTGACGCTGCTCTC
sdh4_Wel_F	CTATTGGTGGGGAGGTCTATG
sdh4_Wel_R	CCCAGATAGATGAAAATCCAGC

Table S2. GenBank accession numbers for mitochondrial sequences used in the study

Species/Genes	Accn. No.
Araucaria heterophylla/nad1	KM672352
Araucaria heterophylla/nad7	KM672359
Cycas taitungensis	AP009381
Ginkgo biloba	NC_027976
Gnetum gnemon/nad1	KM672389
Gnetum gnemon/nad7	KM672396
Pinus strobus/nad1	KM672422
Pinus strobus/nad7	KM672429
Welwitschia mirabilis	NC_029130

Table S3. *Welwitschia mirabilis* mitochondrial editing locations

Gene	Nucleotide	Codon	Genomic	cDNA	Genomic	cDNA
ccmB	157	1	R	W	CGG	TGG
ccmB	163	1	P	S	CCT	TCT
ccmB	175	1	P	S	CCG	TCG
ccmB	224	2	P	L	CCC	CTC
ccmB	239	2	S	L	TCA	TTA
ccmB	341	2	S	L	TCG	TTG
ccmB	470	2	S	L	TCG	TTG
ccmB	491	2	P	L	CCG	CTG
ccmB	500	2	S	L	TCG	TTG
ccmB	554	2	S	L	TCA	TTA
ccmB	557	2	S	L	TCG	TTG
ccmB	569	2	S	F	TCC	TTT
ccmB	570	3	S	F	TCC	TTT
ccmB	575	2	P	L	CCG	CTG
ccmC	258	3	F	F	TTC	TTT
ccmC	538	1	R	W	CGG	TGG
ccmC	551	2	S	L	TCG	TTG
ccmFc	145	1	L	F	CTT	TTT
ccmFc	1370	2	A	V	GCG	GTG
ccmFc	1400	2	P	L	CCA	CTA
ccmFc	1409	2	P	L	CCG	CTG
matR	74	2	P	L	CCC	CTC
matR	1088	2	P	L	CCA	CTA
matR	1163	2	P	L	CCG	CTG
matR	1223	2	S	F	TCC	TTC
matR	1226	2	P	L	CCC	CTC
mttB	88	1	R	C	CGT	TGT
mttB	110	2	S	L	TCA	TTA
mttB	119	2	P	L	CCA	CTA
mttB	137	2	P	L	CCG	CTG
mttB	224	2	S	F	TCC	TTC
nad1	674	2	S	F	TCT	TTT
nad1	751	1	R	W	CGG	TGG
nad1	953	2	P	L	CCG	CTG
nad5	1483	1	R	W	CGG	TGG
nad7	644	2	S	L	TCG	TTG
nad7	926	2	S	L	TCA	TTA
nad7	944	2	P	L	CCC	CTC
nad7	982	1	H	Y	CAT	TAT
nad7	1079	2	S	F	TCT	TTT
nad7	1103	2	S	F	TCT	TTT
nad7	1124	2	P	L	CCA	CTA
rps4	269	2	S	L	TCG	TTG
rps4	281	2	P	L	CCG	CTG
rps4	326	2	P	L	CCG	CTG
rps4	350	2	A	V	GCG	GTG

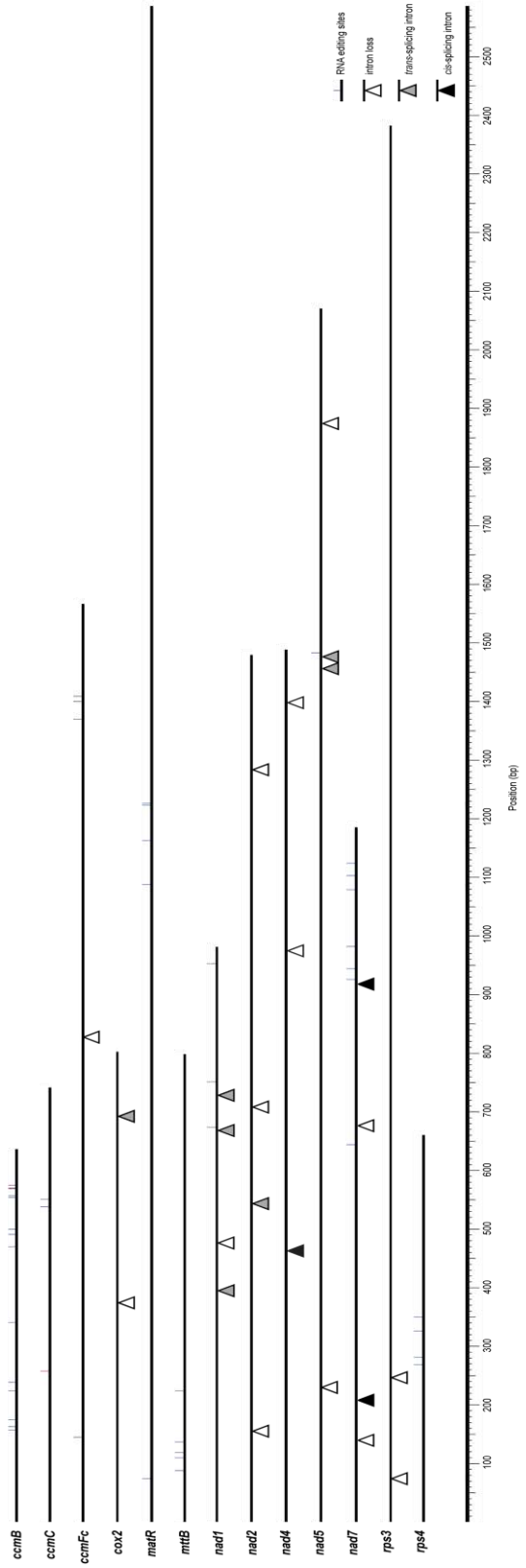
ccmB
ccmC
ccmFc
cox2
matR
mttB
nad1
nad2
nad4
nad5
nad7
rps3
rps4
RNA editing sites
intron loss
trans-splicing intron
cis-splicing intron
Position (bp)
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
2400
2500

Figure S1. Introns and RNA editing sites distribution in *Welwitschia* mitogenes. The genes either were detected editing sites or contained introns were displayed. Gene name were listed at left and the position scale bar was shown at bottom.

CHAPTER 5

Conclusions

In this study, comparative genomics was used to assess the effects of different lifestyles (parasitism, endosymbiosis, xerophytism) in shaping the architecture of organellar genome in green plants.

In the three endosymbiotic green algae selected for study from Chlorellaceae, the organellar genomes did not exhibit genome reduction, in contrast to the extensive genome reduction that has been previously observed in endosymbiotic bacteria and some parasitic algae. Instead, the algal organellar genomes show relatively larger genome size and more introns. Because this endosymbiont lifestyle evolved independently in the three algal species examined, this lack of genomic reduction seems to be a common evolutionary outcome for endosymbionts in the Chlorellaceae. Whether these features are representative of all endosymbiotic green algae are still unclear. Further investigations of other endosymbiont lineages, not only in *Chlorella* and *Micractinium* but also in other diverse groups, will be important to corroborate how the endosymbiotic lifestyle impacts the plant genome.

In the study of parasitic plants, hemiparasites from Orobanchaceae show no evidence of mitogenome degradation, which completely contrasts with results from some other reported parasitic plant mitogenomes. The Orobanchaceae mitogenomes also showed some evidence for horizontal transfer of mitochondrial genes, which is consistent with observations that the parasitic lifestyle increases the propensity for the transmission of foreign DNA between different plant species. In the Orobanchaceae plastomes, the detection of several *ndh* pseudogenes provides strong support that the degradation of the NAD(P)H complex is the initial stage of the transition from fully functional to degraded plastome, in agreement with previous suggestions. However, the hemiparasitic species in

Orobanchaceae may not be representative of all hemiparasites in the green plants. Denser sampling of genomes and transcriptomes of hemiparasites in other families is needed to provide a more comprehensive view of plastome and mitogenome diversity in parasitic plants. Additionally, it will be worthwhile to assess the extent to which parasitic plants show the propensity of HGT in their mitogenome.

In the xerophytic plant *Welwitschia*, a gymnosperm lineage, the massive loss of introns and RNA editing sites from its mitogenome was peculiar in comparison with other plant mitogenomes. However, current understanding of the mechanisms leading to RNA editing variation is limited and biased towards angiosperm species. In *Welwitschia*, the loss of editing sites could be attributed to genomic mutation and retroprocessing, but there are so few edit sites remaining in most genes that statistical analysis was limited. A deeper survey in land plants is needed to reveal the origin and evolution of RNA editing. The study of editing loss mechanisms could also determine whether lineage-specific loss of RNA editing is associated with different living styles.

In a broad sense, these findings and results have contributed to a greater understanding of evolutionary diversity in organellar genome across green plants. With the fast developing sequencing technologies, future studies could answer more comprehensive questions: What is the evolutionary driving force of the organelle genome diversity? To what extent does the photoautotrophic lifestyle affect the conservation of the plastid genome? What is the evolutionary trend of the mitogenome of parasitic plants and what are the key factors that drive these changes? I anticipate that the fantastic and mysterious world of plant organellar genomes will bring even more exciting insights.