The Dissertation Committee for John Christensen Blazier Certifies that this is the approved version of the following dissertation:

Plastid Genome Rearrangement, Gene Loss, and Sequence Divergence in Geraniaceae, Passifloraceae, and Annonaceae.

Committee:

Robert K. Jansen, Supervisor

Z. Jeffrey Chen

David L. Herrin

C. Randal Linder

Claus O. Wilke

Plastid Genome Rearrangement, Gene Loss, and Sequence Divergence in Geraniaceae, Passifloraceae, and Annonaceae.

by

John Christensen Blazier, B.A.; M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2013

Plastid Genome Rearrangement, Gene Loss, and Sequence Divergence in Geraniaceae, Passifloraceae, and Annonaceae.

John Christensen Blazier, PhD

The University of Texas at Austin, 2013

Supervisor: Robert K. Jansen

Plastid genomes of flowering plants are largely identical in gene order and content, but a few lineages have been identified with many gene and intron losses, genomic rearrangements, and accelerated rates of nucleotide substitutions. These aberrant lineages present an opportunity to understand the modes of selection acting on these genomes as well as their long-term stability. My research has focused on two areas within plastid genome evolution in Geraniaceae: first, an investigation of the diversity of unusual plastid genomes in a single genus, *Erodium* (Geraniaceae) for chapters one and three. Chapter two focuses on the evolution of subunits of the plastid-encoded RNA polymerase (PEP). The first chapter described the loss of plastid-encoded NADPH dehydrogenase (*ndh*) genes from a clade of 13 *Erodium* species. Divergence time estimates indicate this clade is less than 5 million years old. This recent loss of *ndh* genes in *Erodium* presents an opportunity to investigate changes in photosynthetic function through comparative biochemistry between *Erodium* species with and without plastid-encoded *ndh* genes. Second, I examined the evolution of the gene encoding the alpha subunit (*rpoA*) of PEP in three disparate angiosperm lineages—*Pelargonium* (Geraniaceae), *Passiflora* (Passifloraceae), and Annonaceae—in which this gene has diverged so greatly that it is barely recognizable. PEP is conserved in the plastid genomes

of all photosynthetic angiosperms. I found multiple lines of evidence indicating that the genes remain functional despite retaining only ~30% sequence identity with *rpoA* genes from outgroups. The genomes containing these divergent *rpoA* genes have undergone significant rearrangement due to illegitimate recombination and gene conversion, and I hypothesized that these phenomena have also driven the divergence of *rpoA*. Third, I conducted a survey of plastid genome evolution in *Erodium* with the completion of 15 additional whole genomes. Except for *Erodium* and some legumes, all angiosperm plastid genomes share a quadripartite structure with large and small single copy regions (LSC, SSC) and two inverted repeats (IR). I discovered a species of *Erodium* that has re-formed a large inverted repeat. Demonstrating a precedent for loss and regain of the IR also impacts models of evolution for other highly rearranged plastid genomes.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Plastid genomes of flowering plants are largely identical in gene order and content, but a few isolated lineages have been identified with a high rate of gene and intron loss, genomic rearrangement, and accelerated rates of nucleotide substitutions. These aberrant lineages present an opportunity to understand the modes of selection acting on these asexually reproducing genomes as well as their long-term stability. The dissertation research has focused on two areas within plastid genome evolution in Geraniaceae: first, an investigation of the diversity of unusual plastid genomes in a single genus, *Erodium* (Geraniaceae) for chapters one and three. Second, chapter two focuses on the evolution of subunits of the plastid-encoded RNA polymerase (PEP) responsible for most photosynthetic gene expression in three unrelated angiosperm lineages in which subunits of PEP have diverged almost beyond recognition.

The first chapter describes the loss of the plastid-encoded NADPH dehydrogenase genes from a single clade of 13 *Erodium* species. Divergence time estimates indicate this clade is less than 5 million years old. The only other autotrophic angiosperms known to lack *ndh* genes are the orchids (Orchidaceae). Knockout experiments have demonstrated that their loss adversely affects plant growth under stressful conditions such as humidity stress and CO2 limitation. The NDH complex is known to participate in cyclic electron flow, but knowledge of its functions is incomplete. The discovery of a recent loss of *ndh* genes in *Erodium* presents an opportunity to investigate changes in photosynthetic function through comparative biochemistry between *Erodium* species with and without plastid-encoded *ndh* genes.

The second chapter focuses on the gene encoding the alpha subunit (*rpoA*) of the plastid-encoded RNA polymerase (PEP), which is conserved in the plastid genomes of all

photosynthetic angiosperms except for three unrelated lineages with unusually divergent open reading frames in the conserved location of the *rpoA* genes. We used sequence-based approaches to evaluate whether these genes are still functional. We find multiple lines of evidence indicating that the genes are functional despite retaining only 30% sequence identity with *rpoA* genes from outgroups. The ratio of non-synonymous substitutions to synonymous substitutions indicates that these genes are under purifying selection, and bioinformatic prediction of conserved domains indicates that conserved domains are preserved. The genomes containing these divergent *rpoA* genes have all undergone significant rearrangement due to illegitimate recombination and gene conversion, and we hypothesize that these phenomena have also driven the divergence of *rpoA*.

The third chapter surveys plastid genome evolution in *Erodium* with the completion of 15 additional whole genomes. Except for *Erodium* and some legumes, all angiosperm plastid genomes share a quadripartite structure with large and small single copy regions (LSC, SSC) and two inverted repeats (IR). Based on subsequent rearrangements in IR-lacking legume plastid genomes, it was hypothesized that the IR functioned to stabilize the plastid genome. *Erodium* plastid genomes do not support this hypothesis: the two major clades differ greatly in their degree of genomic rearrangement. Clade I contains highly rearranged plastid genomes, whereas Clade II contains plastid genomes with few, if any, unique rearrangements. Clade II plastid genomes are far less rearranged than those in the genus *Pelargonium*, in which the IR has greatly expanded. In addition, we find a species in Clade I that has re-formed a large inverted repeat. That evolution has converged back to quadripartite plastid genome architecture in this lineage raises further questions about the functional significance of the IR.

Chapter 2: Recent loss of plastid-encoded *ndh* genes within *Erodium* (Geraniaceae)

INTRODUCTION

With very few exceptions the plastid genomes (plastomes) of angiosperms typically contain 79 protein-coding genes, 30 tRNAs and four rRNA genes (reviewed in Raubeson and Jansen 2005; Bock 2007). Broadly, genes are involved in photosynthesis or housekeeping. Plastome organization is identical in the majority of angiosperms: two single copy regions, the large (~85kb; kilobase) and small (~15kb) single copy regions (LSC, SSC), are separated by two large (~25kb) inverted repeats (IR) designated IRa and IRb. The ribosomal operon lies within the IR. Plastid DNA appears to be present as isomers, differing in the orientation of the single copy regions (Palmer 1983). A few seed plant lineages have lost one copy of the IR, but the vast majority of angiosperms retain this quadripartite structure. Gene order is generally the same across angiosperms, most exceptions being inversions in the LSC as in the grasses, some legumes and sunflower. Very seldom do inversions interrupt conserved transcriptional units (Raubeson and Jansen 2005; Bock 2007).

Relatively few plastid genes have been functionally transferred to the nucleus subsequent to the diversification of angiosperms (Bock and Timmis 2008). Recent transfers or functional substitutions have been documented for only five genes thus far, but for a few of these, such as *infA*, there have been repeated, independent transfers (Millen et al, 2001). Certain ribosomal protein genes appear to have been lost and presumably functionally transferred independently across angiosperms (Jansen et al.

2007), but most other classes of plastid genes are seldom or never lost from photosynthetic seed plants. For example, with few exceptions, genes involved in photosynthesis, electron transport, and ATP synthesis have not been found to be missing from the plastome of a photosynthetic seed plant (Bock 2007; Magee et al. 2010). The 11 plastid-encoded NADH dehydrogenase (*ndh*) genes are unique in being lost or retained only as a full set.

The 11 plastid-encoded *ndh* genes are only commonly lost in nonphotosynthetic plants (Martin and Sabater 2010). In fact, only two photosynthetic seed plant lineages have been documented to lack these genes, Pinaceae/Gnetales (Wakasugi et al. 1994; McCoy et al. 2008; Wu et al. 2009) and a large clade within the Orchidaceae (Neyland and Urbatsch 1996; Chang et al. 2005; Wu et al. 2010). In these two lineages, as in parasitic plants, the entire suite of 11 plastid-encoded *ndh* genes is lost together. Both of these losses of plastid-encoded *ndh* genes are relatively ancient—the divergence of Pinaceae has been estimated at approximately 140 million years ago (MYA) (Wang et al. 2000), and loss of *ndh* genes necessarily predates this divergence, especially if it represents a synapomorphy for Pinaceae and Gnetales (Braukmann et al. 2009). The sister relationship between Pinaceae and Gnetales—the so-called "gnepine" hypothesis—is still controversial (Zhong et al. 2010). However, whether *ndh* gene loss in these lineages represents a single or two separate events, the loss is relatively ancient. Loss of *ndh* genes in orchids is likely also ancient—fossil-calibrated molecular data suggests a crown radiation of Orchidaceae 76-84 MYA (Ramirez et al. 2007). The taxonomic group

confirmed to have lost *ndh* genes encompasses at least 70 orchid genera with over 1000 species (Wu et al. 2010), suggesting that the loss of *ndh* genes is not recent.

It is not known whether Ndh function has been lost or functionally replaced in these two lineages or the genes have been functionally transferred to the nucleus; however, experimental evidence suggests that the thylakoid-bound Ndh complex is a vital intermediary of linear/cyclic electron flow and is apparently indispensable to photosynthesis under a wide range of stress conditions (Endo et al. 1999; Martin et al. 2004; Rumeau et al. 2005; Casano et al. 2001). Unfortunately, neither the composition nor the functions of the Ndh complex have been fully characterized. At least three additional subunits are located in the nuclear genome of angiosperms (Rumeau et al. 2005).

A few angiosperm lineages have been identified with highly rearranged plastomes lacking a variety of genes and introns. One of these lineages, the geranium family (Geraniaceae), contains four major genera, each with distinctly rearranged plastomes (Chumley et al. 2006; Guisinger et al. 2010). All members of the family share the loss of *trnT-ggu* and of two introns (*rps16* and *rpl16*). At 217kb, the garden geranium (*Pelargonium x hortorum*) has the largest and most rearranged plastid genome among angiosperms. Most of the expansion in the *P. x hortorum* plastid genome results from massive expansion of the IR into the single copy regions and from proliferation of complex repeats. While *P. x hortorum* has an IR triple the normal size for an angiosperm at 76kb, the IR has been lost completely in *Erodium texanum* and reduced in *Geranium palmatum* (11kb) and *Monsonia speciosa* (7kb). *Monsonia speciosa* is the only land

plant plastome in which the IR does not include the entire rRNA operon. *Geranium*, *Monsonia*, and *Erodium* all share the loss of the two large ycfs, *ycf1* and *ycf2*, present in *Pelargonium* and in most angiosperm plastid genomes. Families of large dispersed repeats associated with rearrangement endpoints are found in the plastomes of all four Geraniaceae genera (Guisinger et al. 2010).

We chose to survey the genus *Erodium* (with 74 species) to improve our understanding of the extent and diversity of plastome abnormalities in a single lineage within Geraniaceae. The focus of this paper is to compare plastid genome organization of two species and to characterize a single divergent clade of 13 species that lack intact plastid-encoded *ndh* genes. This clade represents the most recent loss of *ndh* genes yet identified among photosynthetic seed plants. Because the subgenera and sections in *Erodium* have been found to be polyphyletic and have not yet been revised, this clade has no official taxonomic status (Fiz et al. 2006). Here we refer to the *ndh*-lacking clade as the long-branch clade (LBC), as it is separated from the rest of the genus by a long branch in phylogenetic reconstructions.

MATERIALS AND METHODS

Taxon sampling and sample preparation

Based on a molecular phylogeny of the genus (Fiz et al. 2006), *E. carvifolium* was chosen to represent Clade II—the published sequence of *E. texanum* represents Clade I (Guisinger et al. 2010)— and three additional taxa, *E. chrysanthum*, *E. guicciardii* and *E. gruinum*, were chosen from a group of species designated as the long-branch clade (LBC;

Fig. 1). Protocols for plastid isolation have been previously described (Jansen et al. 2005). Five to 10 g of fresh young leaf tissue from a single plant was used in each plastid isolation; plants were obtained from commercial sources (Geraniaceae.com; B&T World Seeds, Paguignan, France) and cultivated in the greenhouses at the Brackenridge Field Laboratory at University of Texas at Austin. Vouchers are deposited at TEX.

Genome amplification, sequencing, and finishing

Plastid DNA was amplified using rolling circle amplification (RCA; Qiagen GmbH, Hilden, Germany) utilizing bacteriophage Phi29 polymerase and random hexamer primers (Dean et al. 2001). An EcoRI restriction digest was performed and visualized with ethidium bromide on 1% agarose gel to verify the purity and quantity of plastid DNA. Amplified DNA was either sent to the DOE Joint Genome Institute (Walnut Creek, CA) for Sanger sequencing or the W. M. Keck Center for Comparative and Functional Genomics at University of Illinois for 454 pyrosequencing.

Genomic data was assembled into contigs de novo in the native 454 assembler (Newbler) under default settings, as well as with open-source assembler MIRA (Chevreux 2009) using the "accurate" setting. Putative gene identifications in each contig were performed in DOGMA (Wyman et al. 2004), and contigs were assembled in Geneious (Drummond et al. 2009). Final annotations were made using DOGMA. A circular map was created for *E. carvifolium* using GenomeVx (Conant and Wolfe 2008).

Extraction and analysis of *ndh* pseudogenes

Due to long complex repeats including the apparent duplication of the *atpB/E* transcriptional unit (data not shown), the plastomes of *E. chrysanthum*, *E. guicciardii*, and *E. gruinum* have not yet been assembled into a single contigs. To investigate the presence/absence of *ndh* genes, contigs for each species were converted to a custom BLAST database and queried with intact *ndh* genes from *E. texanum* and *E. carvifolium*.

Protein-coding genes were previously extracted from the *E. chrysanthum* genome assembly (Guisinger et al. 2008), suggesting that the assembly represents the complete plastid genome. The assembly is comprised of 3034 Sanger reads in three large contigs (37 kilobase (kb), 37 kb, and 32 kb) and three small contigs (6 kb, 3 kb, and 2 kb) for a concatenated length of 118,538 base pairs (bp). The *E. guiccardii* and *E. gruinum* datasets comprise of 27,620 and 43,067 454 Titanium pyrosequencing reads of 370 bp and 379 bp average length, respectively; the largest contigs obtained were 101 kb and 98 kb respectively. Both assemblies contain the same set of protein-coding genes previously found in *E. chrysanthum* (data not shown). Besides missing *ndh* genes, all three assemblies show the loss of the *rpoC1* intron, which is present in *E. texanum* and *E. carvifolium*.

Putative *ndh* pseudogenes were aligned against intact genes from *E. texanum* and *E. carvifolium* in Geneious under default settings; indels relative to the intact genes were characterized. Only indels > 3 bp were tabulated in order to avoid counting as indels any sequencing errors due to homopolymer runs. For degraded, unalignable pseudogenes the sizes of BLAST hits from *E. texanum ndh* genes were tabulated.

Primers were designed to amplify an approximately 600 bp region of *ψndhD* containing four deletions shared among the three LBC genome taxa (*ψndhD*-Forward TCCGCAGGTTCCTTCATTTGT; *ψndhD*-Reverse TCTCCGCGAGTGTCTGGTAAC). This *ψndhD* region was amplified for all 13 LBC taxa to confirm the presence of pseudogenes with shared indels throughout the clade. LBC DNAs were provided by J. J. Aldasoro. Sanger sequencing of PCR products was performed on an ABI 3730 platform at The University of Texas at Austin.

Two non-*ndh* genes formerly transcribed with *ndh* genes, *rps15* and *psaC*, were also extracted from LBC genome assemblies using BLAST in the manner described above for *ndh* pseudogenes. The *rps15* and *psaC* genes were aligned with those from *E carvifolium* and *E. texanum* using MUSCLE (Edgar 2004) under default settings as implemented in Geneious.

RESULTS

Genome organization in *Erodium*

The plastome of *E. carvifolium* is 116,934 bp and contains 75 protein-coding genes, 28 tRNA genes and the standard four rRNA genes (Fig. 2). Gene content of *E. carvifolium* is nearly identical to that of *E. texanum* (Table 1), the only differences being the putative loss of *trnK-uuu* in the latter (Guisinger et al. 2010) and the presence of *trnG-gcc* in *E. carvifolium*. Both exons of *trnK-uuu* appear intact in *E. carvifolium*. Relative to *Arabidopsis* the first and second exons of *trnK-uuu* in *E. carvifolium* both have 94%

identity, compared to 83.8% and 63.9% for the *E. texanum* exons, respectively. The losses of the *trnT-ggu* gene and the introns in *rpl16* and *rps16* were previously described for the family Geraniaceae, and we confirm these losses in *E. carvifolium*. The genes *ycf1* and *ycf2* were shown to be lost in three of the four major Geraniaceae genera (*Erodium*, *Geranium*, and *Monsonia*) and these genes also appear to be lost in *E. carvifolium*. The gene *accD* is absent from *E. texanum*, and we did not detect the gene in *E. carvifolium*.

Given the high degree of genomic rearrangement found in representatives of all major Geraniaceae genera (Chumley et al. 2006; Guisinger et al. 2010), the plastome of *E. carvifolium* is remarkably unrearranged, displaying just a single unique inversion. Two additional inversions separate *E. carvifolium* from the ancestral angiosperm gene order typified by tobacco (Raubeson and Jansen 2005), and both inversions are shared among all Geraniaceae plastomes. Among photosynthetic angiosperms *E. carvifolium* has the smallest reported plastid genome at 116,934 bp. Extensive rearrangement in *E. texanum* is associated with the proliferation of long, complex repeats in intergenic regions, contributing to the 14 kb difference in genome size despite nearly identical gene content.

The IR adjacent to the *trnH-gug—psbA* LSC junction has been cleanly deleted in *E. carvifolium*, leaving only 233 bp between *trnH-gug* and *trnL-uag*, formerly LSC and SSC junctions, respectively. Comparison of *E. texanum* and *E. carvifolium* reveals that movement of the IR boundary prior to its loss is not the cause of rearrangement in *Erodium*. Two inversions present in *E. carvifolium* are present in genomes of all four

major genera of Geraniaceae, and the endpoints are associated with the loss of *trnT-ggu* (Guisinger et al. 2010). The tandem duplication of *trnfM-cau* between the rearranged *psaA/B psbC/D* units is also found in *Geranium*. The three additional duplications of *trnfM-cau* found in *E. texanum* are absent from *E. carvifolium*.

The only repetitive region in *E. carvifolium* is located between the rearranged *psbC-D* genes and a conserved cluster of three tRNAs (*trnE-ucc*/*trnY-gua*/*trnD-guc*). This intergenic region contains a pseudogene of *rps18* flanked by two intact copies of *trnG-gcc*. The *rps18* pseudogene shows 76% pairwise identity with the intact gene and contains eight premature stop codons. The two copies of trnG-ucc are located on opposite strands. Duplication of *rps18* and trnG-gcc are not associated with further genomic rearrangement.

Loss of plastid-encoded *ndh* genes in *Erodium*

The complete plastomes of *E. texanum* and *E. carvifolium*, representing the two major clades of *Erodium*, both contain intact copies of all 11 plastid-encoded *ndh* genes (Fig. 1). The draft genome of *E. chrysanthum*, however, was suggested to lack intact copies of these genes (Guisinger et al. 2008). To verify this unusual loss and determine its phylogenetic distribution, three species were chosen for plastome sequencing from the LBC (Fig. 1): *E. chrysanthum*, *E. guicciardii* and *E. gruinum*.

No intact open reading frames (ORFs) were found for *ndh* genes in *E. chrysanthum*, *E. guicciardii*, or *E. gruinum*. For larger *ndh* subunits (e.g., *ndhF*, *ndhD*) pseudogenes

(sequences to be submitted to GenBank) were found by BLAST searches, using intact *ndh* genes from *E. texanum*. These larger subunits could be aligned with confidence against intact genes from *E. texanum* and *E. carvifolium*, revealing that *ndh* pseudogenes retain high sequence identity to intact genes (~90% pairwise identity) despite disruption by copious indels inducing frameshifts (Table 2). For other, primarily shorter *ndh* genes, pseudogenes could not be aligned with confidence with intact *ndh* genes. Only short BLAST hits (< 50bp) were obtained when queried with intact genes from *E. texanum* (Table 3). With few exceptions, the same genes are unalignable or completely degraded in the three draft genome assemblies of *E. chrysanthum*, *E. guicciardii* and *E. gruinum*. *Erodium gruinum* contains longer pseudogenes for a few genes degraded in *E. chrysanthum* and *E. guicciardii* (Table 3); this is consistent with the sister relationship between *E. chrysanthum* and *E. guicciardii* (Fig. 1).

Analysis of indel events in *ndh* pseudogenes reveals a remarkably consistent pattern (Table 2): the great majority of indel events are deletions (86%), and a large proportion of these events (74%) is shared among all three taxa. To determine the extent of ndh gene loss in *Erodium*, we sequenced a disrupted, deletion-rich region of *ψndhD* for all 13 species in the LBC (Fig. 1). The sequenced region contained the same four deletions, three of which cause frameshifts, in all 13 taxa (Figs. 3a and 3b), indicating that the dispersed deletions associated with *ndh* gene loss are shared by all 13 members of this clade.

Two genes formerly co-transcribed with ndh genes in an operon, *rps15* and *psaC*, are intact in Erodium genomes lacking *ndh* genes. *psaC* encodes a subunit of photosystem I

and *rps15* encodes a ribosomal protein. The 5' end of *rps15* is variable in *Erodium*, and there are two indels unique to those taxa also lacking *ndh* genes (accession numbers EU922083.1, HQ730922.1, and HQ730923.1). However, *psaC* is highly conserved and shows no indels or nonsynonymous substitutions in any of the five *Erodium* taxa examined (accession numbers EU922048.1, HQ730924.1 and HQ730925.1).

DISCUSSION

Plastome organization in *Erodium*

Despite encoding almost the same number of genes, genome organization is strikingly different in *E. texanum* (clade I) and *E. carvifolium* (clade II). The four published Geraniaceae plastomes representing each major genus, *P. x hortorum*, *M. speciosa*, *G. palmatum*, and *E. texanum* are all highly rearranged and show extreme variation in IR size, from 76kb in *Pelargonium* to IR loss in *Erodium*. In addition to inversions, expansion and contraction of the IR has been proposed as a mechanism of genomic rearrangement (Chumley et al. 2006). However, comparison of *E. carvifolium* and the highly rearranged *E. texanum*, both lacking the IR, suggests that the deletion of the IR occurred before the divergence of the two major clades, and that no rearrangements occurred concurrent with the deletion event. In *E. carvifolium*, the IR deletion site lies between two tRNAs, leaving less than 250 bp between former single copy region junctions. The *E. carvifolium* genome, with a 'clean' deletion of the IR but no other associated rearrangements (despite multiple rearrangements in related species) is

reminiscent of the situation in legumes, in *Medicago* relative to *Pisum* (Palmer et al. 1988).

The paucity of rearrangements in *E. carvifolium* is striking given the extensive rearrangement evident in other published Geraniaceae plastid genomes (Chumley et al. 2006; Guisinger et al. 2010). Two inversions that distinguish *E. carvifolium* from the ancestral angiosperm gene order are common to all four major genera and thus a synapomorphy of Geraniaceae. Associated with these inversions are the loss of *trnT-ggu* and the duplication of *trnfM-cau*, raising the possibility of illegitimate recombination between tRNAs, a mechanism that has been suggested in at least two other angiosperm groups (Hiratsuka et al. 1989; Haberle et al. 2008). Although the loss of an IR is a derived character in *Erodium*, the lack of other unique rearrangements allows us to hypothesize an ancestral gene order for Geraniaceae that resembles *E. carvifolium* prior to the deletion of the IR. Having a model for the organization of the simplest possible Geraniaceae plastome should greatly assist in reconstructing the rearrangement history of each genus. This model suggests that the rearrangement in each genus has occurred independently, with the exception of two inversions shared by all genera. It also suggests that expansion and contraction of the IR was not a force in genomic rearrangement in all genera, because loss of the IR preceded any unique genomic rearrangements in *Erodium*. Finally, the conservation of the S10 operon in *E. carvifolium* suggests that its fragmentation in *E. texanum* and *G. palmatum* represents two independent events (Guisinger et al. 2010).

Phylogenetic distribution and timing of *ndh* gene loss

The loss of *ndh* genes in the long-branch clade of *Erodium* (Fig. 2) is the most recent and phylogenetically restricted among photosynthetic seed plants. The presence of intact *ndh* genes in both major clades of *Erodium* has been established with the publication of *E. texanum* (Guisinger et al. 2010) and in *E. carvifolium* in this paper; the absence of intact *ndh* genes has been confirmed in the long-branch clade for all 13 species by plastid genome sequencing (3 taxa) or targeted PCR (10 taxa) (Figs. 2a and 2b). Except for a difference in chromosome number—n=8 or 9 in the LBC compared with n=10 in the rest of the genus—no morphology, life history or other trait appears to distinguish this clade from other members of the genus (Fiz et al. 2006). Indeed, homoplasy in morphology caused the LBC to be grouped into the polyphyletic *section Absinthoidea* (Guittonneau 1990; Fiz et al. 2006). Alignment of pseudogenes from three LBC taxa with intact *ndh* genes from *E. texanum* and *E. carvifolium* reveals that the great majority of indels present in pseudogenes are shared by all three LBC taxa (74%). This consistency is a strong indication that loss of plastid-encoded *ndh* genes occurred prior to the diversification of the clade.

Parkinson et al. (2005) used the *cox1* mitochondrial marker to date the divergence of the major clades in Geraniaceae. The inferred divergence times for *Erodium* (from *Geranium*) and the LBC (from the rest of *Erodium* clade I) are at approximately 22 MYA and 7 MYA, respectively. Another estimate using plastid markers indicated slightly earlier divergence times for *Erodium* and the LBC (from the rest of clade I) at approximately 15.45 MYA and 5 MYA respectively (Fiz et al. 2008). These divergence

time estimates are extremely recent compared with the age of the other two known losses of *ndh* genes from photosynthetic seed plants. The loss of *ndh* genes in Pinaceae/Gnetales, if indeed this represents a single loss and a synapomorphy for the clade proposed under the "gnepine" hypothesis (Bowe et al. 2000; Chaw et al. 2000), is ancient. Within gymnosperms, divergence of Pinaceae has been estimated at 140 MYA (Wang et al. 2000), and loss of *ndh* genes necessarily predates this divergence estimate if it represents a synapomorphy for Pinaceae and Gnetales (Braukmann et al. 2009). Loss of *ndh* genes in orchids is likely also ancient compared to that in *Erodium*. An estimate based on fossil-calibrated molecular data suggests a crown radiation of Orchidaceae 76-84 MYA (Ramirez et al. 2007). The taxonomic group recently confirmed to have lost *ndh* genes encompasses 70 orchid genera and over 1000 species, suggesting that the loss of *ndh* genes is not recent (Wu et al. 2010). Orchidaceae is among the largest angiosperm families with ~30,000 species (Atwood 1986), and the full extent of *ndh* gene loss in Orchidaceae has not been determined. For example, many members of the subfamily Epidendroideae (*Brassavola*, *Meiracyllium*, *Cattlea*, *Epidendrum*, *Encyclia*, *Stanhopea*, *Phalaenopsis*, and *Oncidium*), the largest subfamily with more than 10,000 species, have deletions in *ndhF* inducing frameshifts. These data suggest that the loss of plastid-encoded *ndh* genes may be a synapomorphy for much of Orchidaceae (Neyland and Urbatsch 1996; Chang et al. 2006; Wu et al. 2010).

Fate of *ndh* genes: loss, replacement, or transfer?

It is unknown whether Ndh function has been lost entirely in these three seed plant lineages—in gymnosperms, orchids, and Geraniaceae—or whether the genes have been functionally transferred to the nucleus or otherwise replaced. No common features such as parasitism, mycotrophism, or epiphytic lifestyle are associated with *ndh* gene loss in these lineages. This discovery of a recent loss of *ndh* genes in *Erodium* presents an opportunity to investigate changes in photosynthetic function through comparative biochemistry between *Erodium* species with and without plastid-encoded *ndh* genes. The chlorophyll fluorescence assay used to detect *ndh* mutants in tobacco (Horváth et al. 2000) can be used as a first step to assess the presence or absence of Ndh function in LBC species. Detection of Ndh function in LBC taxa could indicate an unprecedented recent burst of gene transfer from the plastid genome to the nuclear genome. Functional replacement of the Ndh complex by a nuclear gene or genes is another possibility; for example, in grasses, the multisubunit acetyl-CoA carboxylase complex encoded by the plastid gene *accD* and nuclear-encoded subunits was replaced by a single-subunit enzyme (Konishi et al. 1996). A failure to detect Ndh function in LBC taxa will suggest that regulation of electron flow between the two photosystems occurs through an alternative mechanism. Experimental data indicates that the Ndh complex is indispensable to the photosynthetic stress response (Endo et al. 1999; Martin et al. 2004; Rumeau et al. 2007; Casano et al. 2001). However, the consistent pattern of deletions found among LBC *ndh* pseudogenes suggests that loss of gene function occurred prior to the diversification of the clade >5MYA, and the subsequent diversification and persistence of this clade

suggest that photosynthetic stress response has not been greatly compromised. Because the electron donor and functions of the Ndh complex are still unresolved it is difficult to evaluate the relative likelihood of gene loss, functional replacement, and functional gene transfer of *ndh* genes in *Erodium*.

CONCLUSIONS

The causal relationships among the evolutionary rate acceleration, indels, and loss of *ndh* genes in the LBC remain unclear. It is possible that the loss or functional transfer of *ndh* genes was promoted by the acceleration in the rate of evolution, if the rate in the plastome surpassed that in the nucleus, as has been recently suggested by Magee et al. (2010) for a hypervariable region in legume plastomes. A full evolutionary rate analysis of the genes from each functional class across eight additional *Erodium* species and monotypic sister species *California macrophylla* is underway to disentangle the relative roles of evolutionary rate accelerations, gene losses, indel frequency, and genomic rearrangements in *Erodium* plastome evolution. Finally, the recentness and restricted distribution of *ndh* gene loss in *Erodium* raises the possibility of finding additional recent losses among angiosperms, perhaps within lineages known to display plastome rearrangements.

Table 2.1. Comparison of *E. texanum* (clade I) and *E. carvifolium* (clade II) plastid genomes.

Genome Characteristic	***E. texanum***	***E. carvifolium***
Size (bp)	130,812	116,934
Number of different protein-coding genes	76	76
Number of different tRNA genes	26	28
Number of different rRNA genes	4	4
Number of genes with introns	14	14
GC content	39.50%	39%

Table 2.2. Number of indels present in *ndh* pseudogenes.

Full-length pseudogenes were found for these coding sequences. Of the 39 indel events, 29 (74%) are shared among the three LBC taxa. The great majority (84%) are deletions.

	E. gruinum	***E. chrysanthum***	***E. guicciardii***	**Shared indels**
***ndhA* exon 1**				
indels	2	2	2	2 of 2
***ndhB* exon 1**				
indels	4	5	5	4 of 5
***ndhB* exon 2**				
indels	3	4	3	3 of 4
ndhD				
indels	7	7	7	6 of 8
ndhF				
indels	17	13	14	13 of 17
ndhJ				
indels	1	3	1	1 of 3
				29/39= 74%

Table 2.3. Size of BLAST hits (in base pairs) returned for degraded *ndh* pseudogenes.

The gene length is from functional *E. texanum* sequences. The majority of BLAST hits are <50 bp, indicating substantial degradation. *Erodium gruinum* contains long, identifiable pseudogenes for two *ndh* genes degraded in *E. chrysanthum* and *E. guicciardii*. Short BLAST hits (20-30 bp long) are frequently spurious and are not necessarily pseudogenes.

	Gene length	***E. gruinum***	***E. chrysanthum***	***E. guicciardii***
ndhC	363	351	26	30
ndhE	306	81	19	14
ndhG	540	23	22	22
ndhH	1182	29	29	29
ndhI	516	47	21	55
ndhK	684	668	20	20

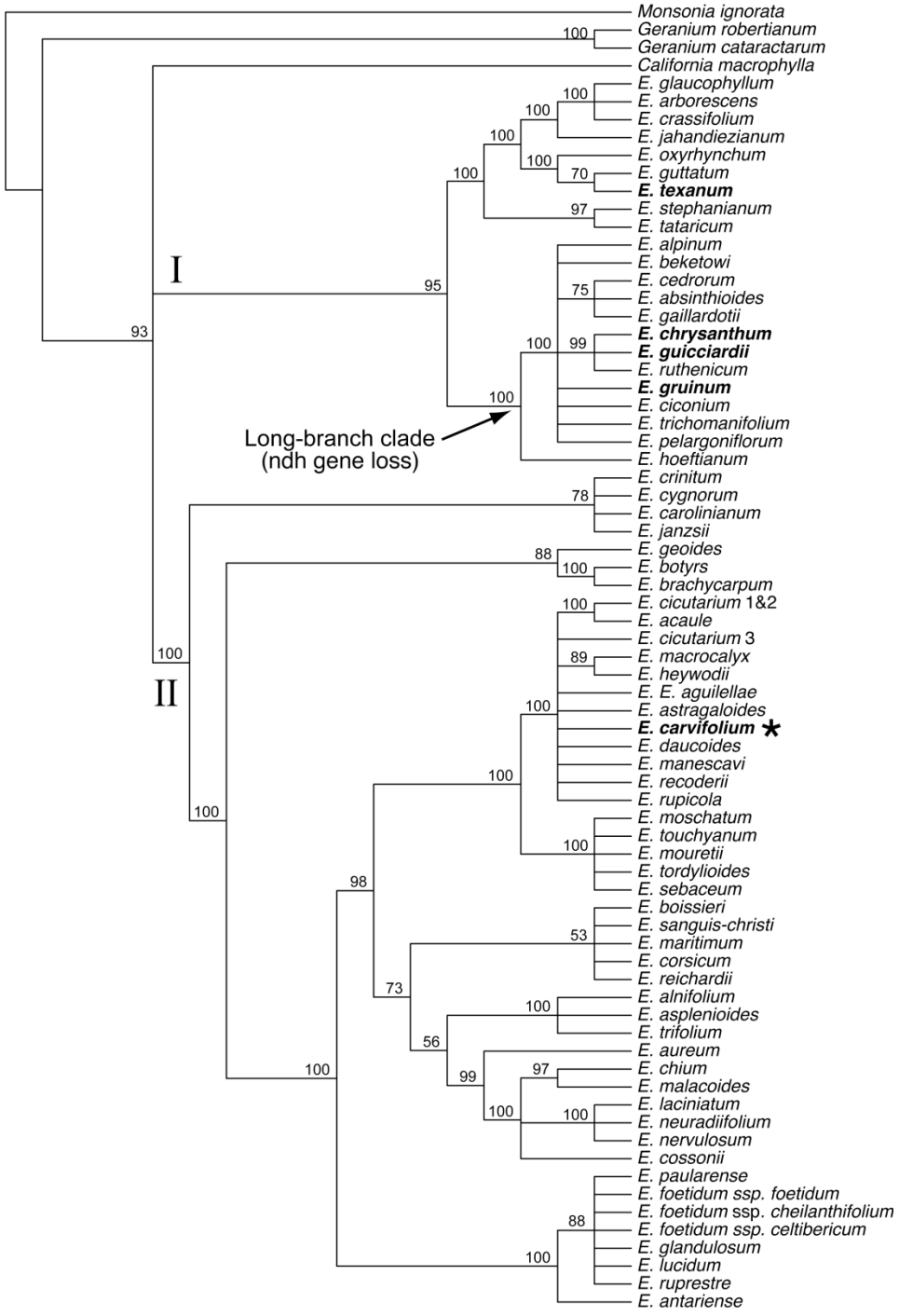

Figure 2.1. Phylogeny of *Erodium* adapted from Fiz et al. (2006). Clades I, II, and the long-branch clade (LBC) are indicated. Numbers at nodes are Bayesian posterior probabilities as percentages. Taxa in bold represent those with complete (*E. texanum*, *E. carvifolium*) or draft plastid genome sequences. The asterisk indicates the new genome sequence of *E. carvifolium* reported in this paper.

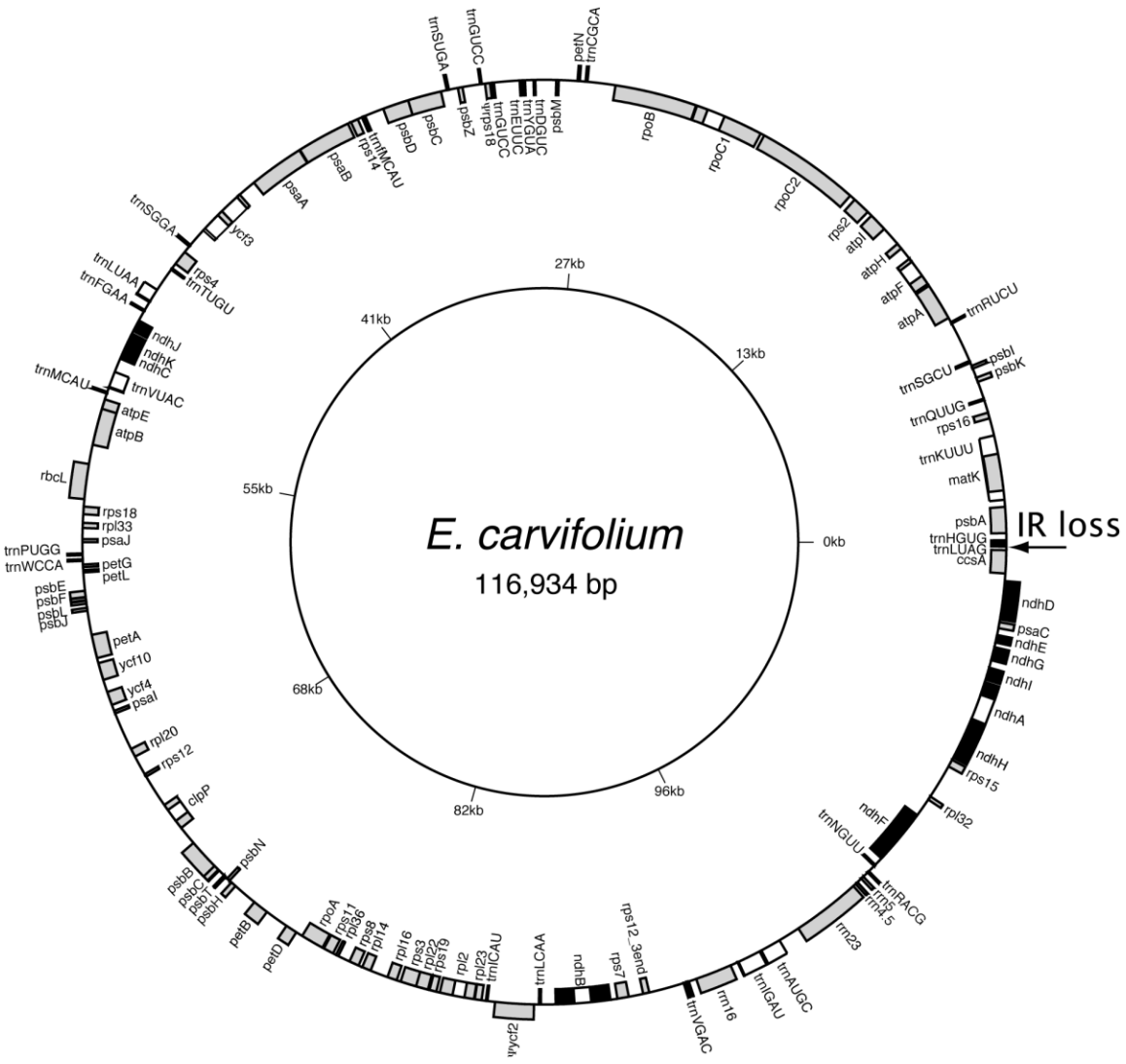

Figure 2.2. Circularized gene map of the *E. carvifolium* plastid genome. Genes on the outside of map are transcribed in the counterclockwise direction and genes on the inside of the map are transcribed in the clockwise direction. *ndh* genes are in black; all others in gray. The arrow indicates IR deletion location.

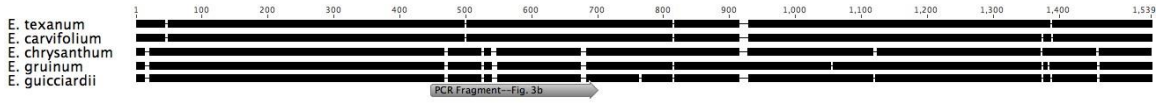

Figure 2.3a. Alignment of *ndhD* regions from *C. macrophylla*, *E. carvifolium*, and *E. texanum* with *ψndhD* from 3 LBC taxa. The *ψndhD* region amplified for all 13 LBC taxa (Figure 3b) is marked with an arrow.

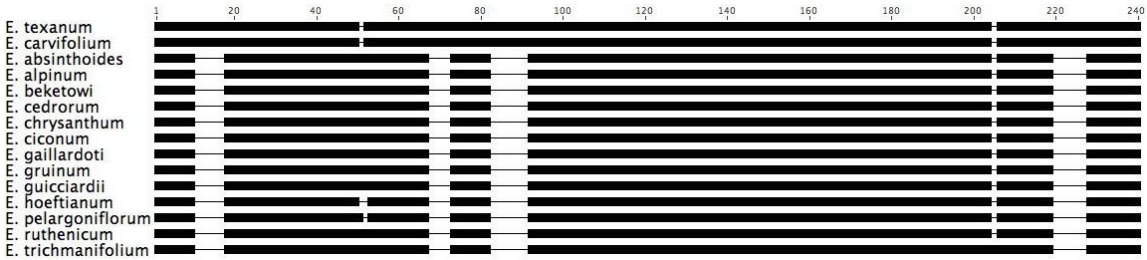

Figure 2.3b. Alignment of *ψndhD* fragments from all 13 LBC taxa against intact *ndhD* genes from *E. texanum* and *E. carvifolium*. The lengths of the shared deletions, in bp, from left to right are 7, 5, 9, and 8.

Chapter 3: Divergence of RNA polymerase α subunits in angiosperm plastid genomes is mediated by genomic rearrangement rather than positive selection.

INTRODUCTION

Before the advent of inexpensive DNA sequencing, the plastid genomes of flowering plants (angiosperms) were surveyed for gene content using Southern hybridization (Downie & Palmer 1992; Doyle et al. 1995; Gantt et al. 1991). These surveys revealed a remarkably conserved gene order and content across almost all angiosperms, yet also discovered a few isolated lineages with highly divergent, rearranged genomes also lacking genes and introns. The subsequent publication of several hundred complete plastid genomes has confirmed most of these early results. Plastid genomes typically contain 79 protein-coding genes, 30 tRNAs, and 4 ribosomal genes (Bock 2007). Plastid genes are often categorized as either photosynthesis-related or housekeeping genes, and it is housekeeping genes that are generally found to have been lost from plastid genomes and either functionally replaced or transferred to the nucleus (Millen et al. 2001; Jansen et al. 2007). Among the housekeeping genes encoded by the plastid genome are the four subunits of the eubacterial plastid-encoded RNA polymerase (PEP) that is responsible for most photosynthetic gene expression. A lineage was identified (*Pelargonium*, Geraniaceae) in which one of these PEP subunits, *rpoA*, failed to produce a signal in Southern hybridizations (Downie & Palmer 1992). The first plastid genome from this lineage to be sequenced, *Pelargonium x hortorum*, was found to contain multiple *rpoA*-like ORFs (Chumley et al. 2006), and it has been a long-standing mystery whether these represent pseudogenes or functional, divergent *rpoA* genes. Moreover, if these genes are functional, it is unclear whether they have diverged due to positive selection or by some unusual, locus-specific neutral process.

The three largest PEP subunits β, β', and β'' are co-transcribed as the adjacent *rpoB*, *rpoC1*, and *rpoC2* genes, respectively. The fourth and smallest subunit, *rpoA*, encodes the α subunit and is transcribed as the last gene in the conserved *rpl23* transcriptional unit consisting mostly of ribosomal protein genes. The three large subunits have only been found missing from a few parasitic and mycoheterotrophic plant plastid genomes (Delannoy et al. 2011; Wicke et al. 2013). In these cases it appears that the single-subunit nuclear-encoded RNA polymerase (NEP) has taken over all transcription of the residual plastid genome, which no longer encodes a functional photosynthetic apparatus. Deletions of individual PEP subunits from tobacco were all found to produce photosynthetically defective transformants, suggesting all four subunits are essential genes (Serino & Maliga, 1998). PEP and NEP express many of the same genes using distinctly different promoters; species lacking PEP have also lost PEP-specific promoters and nuclear-encoded σ factors (Krause et al. 2003; Wickett et al. 2011).

Determining the functionality of *rpoA* poses some unique difficulties. First, there is no published nuclear genome data for *Pelargonium*. It is possible that *rpoA* has been transferred to the nucleus and that the divergence of the gene reflects relaxed selection on the plastid copy in the wake of its functional replacement by a nuclear paralog. Although the transfer of *rpoA* has not been shown to have occurred in any angiosperm, it has been found in the moss *Physcomitrella patiens* (Sugiura et al. 2003), and it has been inferred that *rpoA* has been transferred to the nucleus twice in the bryophytes (Goffinet et al. 2005). Genes transferred to the nucleus from angiosperm plastid genomes have been found to persist intact in the plastid genome and degrade slowly, so it may be difficult to judge the functionality of an ORF if the gene has been transferred to the nucleus relatively recently (Jansen et al. 2011). Second, the *P. x hortorum* plastid genome is the largest and most complex angiosperm plastid genome yet discovered and houses three

distinct, divergent *rpoA*-like ORFs (Chumley et al. 2006). No other plastid genome is known to harbor multiple paralogs of any gene, and it is difficult to judge which, if any, of the ORFs is functional. Third, expression data is unlikely to give a conclusive answer regarding functionality. The *rpoA* gene is generally located at the end of a conserved transcriptional unit, so some read-through transcription is likely to occur even if the gene is nonfunctional (Shi et al. 2013). Fourth, the α subunit is the least conserved of the PEP subunits (Little & Hallick, 1988), so its degree of divergence may not be a useful criterion in determining functionality. Fifth, plastid transformation protocols have not been developed for any species of Geraniaceae, thus precluding direct experimentation by reverse genetics such as knockout or replacement of the *rpoA*-like ORFs. Lastly, commonly used methods, such as likelihood-based calculation of *dN*/*dS* ratios to detect selection, may be inappropriate for some of these ORFs, as some appear to be evolving in ways not anticipated by standard evolutionary models. For example, gene conversion is known to produce spurious signals of selection under likelihood-based models (Casola & Hahn, 2009). In addition, selection may be falsely detected due to alignment error (Schneider et al. 2009); some of the divergent *rpoA*-like ORFs share just 30% sequence identity with the same gene from an outgroup within the same angiosperm order. At this level of divergence, different alignment methods can produce different estimates of evolutionary rates, none of which is obviously superior to the others.

Due to the intractability of reverse genetics in most plastid genomes, we have adopted a sequence-based approach to resolve the long-standing mystery of whether *Pelargonium* plastid genomes still encode a functional PEP α subunit, and we apply these same methods to two additional lineages we found that also failed to produce a signal for *rpoA* in Southern hybridizations, *Passiflora* (Passifloraceae) and Annonaceae. We use several sequence-based analyses to demonstrate that these are functional genes, some of

which have been evolving in manner unlike that of other plastid genes due to illegitimate recombination and gene conversion.

MATERIALS AND METHODS

Southern blot hybridizations

Names of the 284 diverse species of angiosperms (representing 280 genera and 191 families) examined by Southern blot hybridization are provided in Supplementary Table 1. Immobilon-Ny+ membranes (Millipore) were prehybridized for 2-4 hours and then hybridized overnight, both at 60-62°C and in 5X SSC, 0.1% SDS, 50 mM Tris (pH 8.0), 10 mM EDTA, 2X Denhardt's solution, and 5% dextran sulfate. The membranes were then washed twice for 30-45 min at 60-62°C in 2X SSC and 0.5% SDS. Autoradiography was carried out at room temperature for 18-48 hours. Hybridized membranes were stripped by boiling 2-3 times for 5-10 min in 0.1X SSC and 0.1% SDS. To make sure there was no carryover signal, blots were overexposed for 3 days with intensifying screens at –80°C prior to a new hybridization cycle. Hybridization probes for all four plastid-encoded rpo genes and the plastid 16S rRNA gene (16S rDNA; used as a positive control) were generated from tobacco using the polymerase chain reaction (PCR) and gene-specific primers, followed by radiolabeling with 32P using random oligonucleotide primers.

Probe lengths and PCR primers

1) *rpoA*, probe length = 892 bp, primers SAGTGGAARTGTGTKGAATCA-(forward) and TCYTBVSTSTTABTCAAAAGKTCC-(reverse).
2) *rpoB*, probe length = 1,404 bp, TAGTYCTATYATCAGCTATGGG and TCCAAYCCMGTTCCAACAATGC.

3) *rpoC1*, probe length = 1,034 bp, GATACACTTCTTGATAATGG and GACCAACAGTGGTTCGAATG.
4) *rpoC2*, probe length = 1,428 bp CATTAAGAACTTTTCATACYGG and ATCCGYTGGACATAGATCCA.
5) 16S rDNA, probe length = 1,158 bp, GACACGGCCCAGACTCCTAC and ATCCAGCCGCACCTTCCAG.

DNA and RNA isolation

Illumina sequencing was performed on total genomic DNA extracted by a modified version of hexadecyltrimethylammonium bromide protocol from Doyle and Doyle (1987) with 2% Polyvinylpyrrolidone-40. Total genomic DNAs used for Southern blot hybridization were prepared by a slightly different modification of this same protocol, with further purification by CsCl/ethidium bromide centrifugation.

For RNA isolation, newly emergent leaves of *Pelargonium x hortorum* cv Ringo white were collected from live plants grown in the University of Texas (UT) greenhouse and frozen in liquid nitrogen. Total RNA was isolated by grinding in liquid nitrogen followed by 30 min incubation at 65°C in two volumes of extraction buffer (2% Cetyltrimethylammonium bromide, 3% Polyvinylpyrrolidone-40, 3% 2-Mercaptoethanol, 25 mM Ethylenediaminetetraacetic acid, 100 mM Tris(hydroxymethyl)aminomethane-HCl pH 8, 2 M NaCl, 2.5 mM spermidine trihydrochloride) with vortexing at 5 min intervals. Phase separation with chloroform:isolamyl alcohol (24:1) was performed twice and the aqueous phase was adjusted to 2M LiCl. Samples were precipitated overnight at 4°C and total RNA was pelleted by centrifugation at 17, 000 x g for 20 min at 4°C. RNA pellets were washed once with 70% ethanol and air dried at room temperature. Following

resuspension in RNase free water, RNAs were analyzed by denaturing gel electrophoresis and by spectrophotometry. DNase I was removed from the solution by extraction with phenol:chloroform:isoamyl alcohol (25:24:1) and the aqueous phase was adjusted to 0.3 M sodium acetate. RNA was precipitated with 0.6 volumes of cold isopropanol for 20 min at -80°C. Pellets were washed with 70% ethanol, air-dried and resuspended in water to 1 μg μL-1. Total RNA sample aliquots were frozen in liquid nitrogen.

DNA sequencing and assembly

Taxon sampling for the three data sets for DNA sequencing involved representatives of the Annonaceae, Geraniaceae, Passifloraceae, and associated outgroups (Table 1). Sanger and 454 sequencing were performed on products of Rolling Circle Amplification (RCA) of purified plastids as described in Jansen et al. (2005) and assembled using consed or Newbler and MIRA as described in Chumley et al. (2006) and Blazier et al. (2011), respectively (Gordon et al. 1998; Margulies et al. 2005; Chevreux et al. 1999). Total genomic DNAs were sequenced on the Illumina HiSeq 2000 at Bejing Genomics Institure (BGI). Sixty million 100 bp paired-end reads with a 750 bp insert size were generated for each sample. Illumina data was assembled using Velvet v. 1.2.07 with k-mer sizes ranging from 71 to 93 (Zerbino & Birney, 2008). Contigs representing nuclear and mitochondrial DNA had relatively low depth of coverage and were excluded using a 1000x coverage cutoff. Assembly and plastid contig extraction were performed on the Lonestar Linux Cluster from the Texas Advanced Computing Center (TACC) using custom Python scripts.

Sequence alignment and rates analyses

Plastid contigs were evaluated and genes were extracted in DOGMA (Wyman et al. 2004), and all sequence editing and alignment were conducted in Geneious (Drummond et al. 2011). Alignment of *rpo* genes was conducted using the L-INS-i algorithm in MAFFT as implemented in Geneious, as a single locally alignable block flanked by long terminal gaps was expected (Katoh et al. 2009). For other plastid genes, the MAFFT G-INS-i algorithm was used, as a global alignment without large terminal gaps was expected.

Constraint trees for the three data sets were created using a concatenated nucleotide alignment of seven plastid genes (*rpoA*, *rpoB*, *rpoC1*, *rpoC2*, *ndhF*, *matK*, and *rbcL*). Constraint trees were generated in Garli (Zwickl, 2008) under default settings as implemented in Geneious. Codon alignments were created using MAFFT in Geneious. For the *Pelargonium* data set, several additional alignment algorithms (CLUSTALW, MUSCLE, and the Geneious aligner) were used in order to control for alignment error, as these sequences were particularly difficult to align (Thompson et al. 1994; Edgar, 2004).

Plastid genes were analyzed with codon-based models to quantify the rates of synonymous (*dS*) and nonsynonymous (*dN*) substitution. Analyses were conducted in PAML 4.7 on the Lonestar Linux Cluster from TACC using custom Python scripts (Yang, 2007). Codon frequencies were calculated by the F3×4 model, and a free-ratio model was used to compute *dN*/*dS* values. Transition/transversion and *dN*/*dS* ratios were estimated with the initial values of 2 and 0.4, respectively, consistent with other studies examining evolutionary rate heterogeneity in angiosperm organellar genomes (Sloan et al. 2009; Weng et al. 2012).

Promoter analysis

The upstream regions containing the promoter sequence of *psbA* and *rbcL* were aligned in MAFFT as implemented in Geneious and conserved PEP promoter elements were annotated in accordance with Gruissem and Zurawski (1985). Upstream regions of *Cuscuta obtusaflora*, a parasitic plant lacking PEP, were also included in the alignments for comparison.

Conserved domain prediction

Conserved domains in *rpoA*-like ORFs were predicted by searching against the Conserved Domain Database at NCBI (CDD v.3.10) at an E-value threshold of 0.01 and low-complexity filters applied (Marchler-Bauer et al. 2010).

RT-PCR

Primers were designed in Primer3 as implemented in Geneious (Koressaar & Remm, 2007) to amplify transcripts of the two largest, oldest, and best conserved *rpoA*-like ORFs in *P. x hortorum*. The third ORF (ORF521) was not considered because it was still erroneously believed to be comprised of two fragments in different reading frames (Chumley et al. 2006) rather than a single ORF due to a sequencing error in the published plastid genome.

Approximately 1 ug of *P. x hortorum* DNase-free RNA was thawed on ice and used the template for reverse transcription polymerase chain reactions (RT-PCR). The RT reactions utilized ImProm-II™ Reverse Transcriptase (Promega, Madison WI) following the manufacturer's protocol. Reverse primers (*rpoA*-R: GTCCTTTTCGTTTTC; 597-R: GAATTCTCGATTTCCTCTTTTCCG) were used in RT reactions along with *rbcL* positive control (*rbcL*-R: ATTACGATAGGAACCCCCAACT). For each reaction a control reaction was also

performed where no enzyme was added. Reverse transcription products, 3 uL each, were used as templates for PCR reactions using the Phusion High-Fidelity DNA Polymerase (Thermo Scientific, Pittsburgh PA) according to the manufacturer's protocol and MgCl2-free buffer. Magnesium chloride concentration was adjusted to 2 mM. Primers for PCR reactions (*rpoA*-R; 578-F: CTCGCTAAAGTCCAT; 578_*petD*-R: CCACATAGTCCCAGTCT; petD-F: CCCGACTTGAATGATCCTG; 597-R; 597-F: AAATAACAAGAACTTCGC; rps11-F: GCTATTCGCACAGTAGTAAC; *rbcL*-R; *rbcL*-F: AAAGGGCATTACTTGAATGCTA).

Detection of gene conversion

Detection of gene conversion among *Pelargonium rpoA*-like ORFs was conducted both manually and using gene conversion algorithms implemented in ORGCONV (Hao, 2010). For manual detection of gene conversion events, the alignment was inspected for SNPs shared by two or three *rpoA* paralogs in a single species that were not shared across paralogs in multiple species.

RESULTS

Southern Hybridization

We used a probe comprising nearly the full-length *rpoA* gene from tobacco in a Southern blot hybridization survey of 284 diverse angiosperms, representing 280 genera, 191 families, and all major lineages except for *Amborella*, whose plastid genome has been sequenced (Goremykin et al. 2004). All examined taxa hybridized well to the *rpoA* probe except for three distantly related species – *Pelargonium x hortorum* (Geraniaceae), *Passiflora suberosa* (Passifloraceae), and *Annona muricata* (Annonaceae) – which showed either no detectable or a highly reduced signal (Supplemental Fig. 1). All three

of these species, along with the other 281 taxa, hybridized well to probes for the 16S rRNA gene and for the three other PEP subunits (*rpoB*, *rpoC1*, and *rpoC2*). These results suggest that *rpoA* is either missing from plastid genome or, if present, highly divergent in these three taxa.

RT-PCR

RT-PCR was performed for *Pelargonium x hortorum*, one of the three lineages that showed no hybridizaton signal for this gene. The RT-PCR results confirm the presence of transcripts for both *rpoA*-like ORFs that were assayed, the two longer ORFs, ORF578 and ORF597 (Chumley et al. 2006). Both products were amplified from the RT template twice, first using primers within the ORFs and then with the forward primers located in the upstream genes, *petD* and *rps11* (Figure 1). The results confirmed that there are at least discistronic transcripts for both of these ORFs but cannot distinguish between the ORF transcripts being present as monocistrons or as polycistrons including genes further upstream.

Promoter and Transcriptome Analysis

A high-coverage nuclear transcriptome assembly of *P. x hortorum* was surveyed using *rpoA* nucleotide and amino acid sequences from *A. thaliana* (Zhang et al. in press). No nuclear-encoded *rpoA* transcript was detected by BLAST search. Other nuclear-encoded components of the PEP holoenzyme, e.g. sigma factors, were found using the same BLAST parameters.

In *P. x hortorum* the promoter regions of *rbcL* and *psbA* closely resembled those of *Arabidopsis thaliana* and *Nicotiana tabacum*. The -35 and -10 elements as well as the transcription start sites were highly conserved, unlike in *Cuscuta obtusaflora*, a parasitic plant lacking PEP (Figures 2a and 2b).

Analysis of Signals of Selection

The ratio of non-synonymous nucleotide substitutions to synonymous substitutions (*dN*/*dS*) was calculated for the three different lineages of angiosperms.

Annonaceae

The Annonaceae data set consisted of nine species from the Magnoliales including three different genera in the Annonaceae: *Annona*, *Asimina*, and *Cananga* (Table 1). Maximum likelihood trees for *matK* and *rpoA* were generated for this data set (Fig. 3): in the *matK* tree, similar to trees generated from other individual plastid genes, the branch leading to *Piper* was long but branches within Annonaceae relatively short. In the *rpoA* tree, however, branch lengths within Annonaceae were extremely long, sufficiently long to produce an incorrect topology through long-branch attraction to *Piper*. In all, seven plastid genes (*rpoA*, *rpoB*, *rpoC1*, *rpoC2*, *ndhF*, *matK*, and *rbcL*) were analyzed in PAML to determine if the *dN*/*dS* ratio could indicate whether or not the divergent Annonaceae *rpoA* genes are under purifying selection. For this data set there were five branches of interest: the branch leading to Annonaceae, the branch leading to *Annona* and *Asimina*, and the three terminal branches leading to *Annona*, *Asimina*, and *Cananga* (Table 2). Only one of the branches had a *dN*/*dS* value >1, the terminal branch leading to *Asimina* for *matK* (*dN*/*dS* = 1.0069). All the *rpo* genes showed *dN*/*dS* values consistent with purifying selection in Annonaceae.

Passiflora

The *Passiflora* data set consisted of 12 taxa from the Malpighiales, including four *Passiflora* species (Table 1). Maximum likelihood trees for *matK* and *rpoA* were constructed for this data set: in the *matK* tree, similar to trees generated from other individual plastid genes, the branch leading to *Turnera* (Passifloraceae) was long but

branches within *Passiflora* are relatively short. In the *rpoA* tree, however, the terminal branch leading to *Passiflora biflora* was extremely long, sufficiently long to produce an incorrect topology through long-branch attraction to *Turnera*. In all, seven plastid genes (*rpoA*, *rpoB*, *rpoC1*, *rpoC2*, *ndhF*, *matK*, and *rbcL*) were analyzed in PAML to determine if the *dN*/*dS* ratio could indicate whether or not the divergent *Passiflora biflora rpoA* gene is under purifying selection. The branches of interest for this data set were the branch leading to the family, to the genus, to *P. ciliata* and *P. quadrangularis*, to *P. cirrhiflora* and *P. biflora*, and all the terminal branches (Table 3); however, the principal branch of interest was the terminal branch leading to *P. biflora*, since just this single species has a divergent *rpoA* (Fig. 4). The only gene for which a branch had a *dN*/*dS* value >1 was *rpoC1*, on the terminal branch leading to *P. quadrangularis*.

Pelargonium

The *Pelargonium* data set consisted of 26 species representing all major clades (Table 1). *Pelargonium rpoA* genes showed a complex pattern of divergence by clade that confounds the straightforward analysis of evolutionary rates conducted on the smaller Annonaceae and *Passiflora* data sets. A maximum likelihood tree of all *rpoA* genes/ORFs from the *Pelargonium* data set was generated (Fig. 5). Alignment of *Pelargonium rpoA* sequences is difficult across clades and with outgroups. Errors in alignment or in the models used to estimate *dN* and *dS* can cause estimates of *dN*/*dS* that are outside the range of what is biologically probable; for this reason four different alignment algorithms were utilized in *Pelargonium* rate comparisons. A *dN*/*dS* ratio of 50 was selected as an arbitrary cutoff over which a value was assumed to be an artifact. After conducting rate analyses on each *Pelargonium* clade separately (data not shown) and also on the combined data from clades A, B, and C1 (following the naming scheme of Bakker (2004)

we found that analysis of the combined data set produced fewer artifacts, with only a single *dN*/*dS* value >50 (Table 4).

rpoA genes in clades A and B were somewhat divergent between the two clades, sharing only 66-71% sequence identity, but showed high pairwise identity within each clade. Clade B of *Pelargonium* was the smallest data set. The five *rpoA* genes representing clade B shared 94% sequence identity; however, this percentage was lowered by indels associated with tandem repeats at the 3` end of the gene immediately preceding the predicted stop codon (Figure 6). When this repeat-rich region was excluded from the alignment the genes share over 98% sequence identity. In fact, four of the five genes were 100% identical when the 3` end was excluded, and the fifth sequence, *P. exstipulatum*, differed by only two nucleotides, both of which were nested in tandem repeats and cause non-synonymous substitutions. The repeat rich 3` ends of clade A and clade B *rpoA* genes were compared (see below).

Nine *rpoA* genes representing clade A shared 92% identical sites, or 95% identical sites if the 3` end was excluded. Similar to clade B, different numbers of tandem repeats towards the 3` end caused length differences in clade A *rpoA* (Fig. 7). Although the phenomenon underlying the length differences in clade B and clade A *rpoA* genes was the same—indels associated with tandem repeats--the actual repeats were not homologous sequences, nor were the ORFs extremely similar between the clades. In clade B there were two different tandem repeat units that underlie the length differences: a 6 bp motif of GCGAGG was present in all the ORFs, ranging from two repeat units in *P. australe* to eight in the same region of *P. grossularioides*. In *P. cotyledonis*, two copies of this 6 bp tandem repeat were nested inside a unique, larger 39 bp repeat, which expanded to four tandem copies, the last base pair of which was the first base pair of the predicted TAA stop codon (Figure 6). The 6 bp repeat was not found in any clade A *rpoA*

sequence, and instead, a 9 bp repeat unit, present as both tandem and dispersed repeats at the 3` end of the gene in all clade A species, appeared to have caused a deletion of 30 bp between two direct, dispersed 9 bp repeat units in *P. echinatum* and *P. fulgidum*. These two taxa were not sister species, and thus it appeared that this deletion occurred twice independently in clade A.

In contrast, the C1 and C2 clades were highly divergent both within and between clades, and the C2 clade contained species containing multiple (2 or 3) *rpoA*-like ORFs (Figure 5). For clade C2 species it was not clear which of the paralogous ORFs might be functional. We omitted *rpoA* genes from clade C2 from the analysis for two reasons: first, the sequences were even more difficult to align than those from clade C1 and thus even more likely to give spurious results. Second, the taxa from clade C2 with multiple *rpoA*-like ORFs showed evidence of gene conversion among the paralogs (Tables 4 and 5).

Clade C1 was represented by five species whose ORFs fell out into two groups of more closely related sequences. *Pelargonium dolomiticum* and *P. trifidum* were relatively similar at 95.5% pairwise sequence identity. *Pelargonium tetragonum*, and *P. worcesterae* had 99.5% identity and were identical in length at 912bp; *P. myrrhifolium* was more closely related to this second pair but nonetheless shared only 61.2% sequence identity with *P. tetragonum*. Between the groups, *P. dolomiticum* and *P. tetragonum* had only 64% pairwise identity.

The branches of interest for the *Pelargonium* rates analyses were different from those in the previous two data sets: the terminal branches were excluded as intra-clade divergence among species was extremely low due to much denser taxon sampling in this data set. Low sequence divergence between closely related taxa caused error values to be returned in the calculation of *dN*/*dS* when either or both of the parameters were calculated to be zero or close to zero. As a result, we have chosen as our branches of interest those

on which the greatest divergence in *rpoA* has occurred, namely the branches leading to the family, to *Pelargonium*, to clades A and B, to clade A, to clade B, and to clade C1.

Rates analyses of the non-*rpo* genes for *Pelargonium* found low *dN*/*dS* values consistent with purifying selection across all alignments for all branches of interest (Table 4). For the *rpo* genes a pattern emerged that was consistent across all different alignments used: *dN*/*dS* values for *rpoA* were uniformly low (<1), consistent with purifying selection on all branches of interest. However, *dN*/*dS* values for the other *rpo* genes were elevated along several branches of interest. For the branch leading to clades A and B, *rpoB*, *rpoC1*, and *rpoC2* all showed *dN*/*dS* values >1. For the branch leading to clade A, *rpoB* and *rpoC1* showed *dN*/*dS* values >1, and *rpoC2* showed values near or >1 depending on the alignment method used. For the branch leading to clade B, *rpoB* and *rpoC1* but not *rpoC2* showed *dN*/*dS* values >1 for all alignment methods used. On the branch leading to clade C1, *rpoC1* and *rpoC2* but not *rpoB* showed *dN*/*dS* values >1.

Conserved Domain Search

For each of the three data sets, the outgroup taxa—*Chloranthus*, *Populus*, and *Eucalyptus*—were queried against the Conserved Domain Database (CDD) to verify the recognition of functional domains. In all three cases the three functional domains were predicted to be present (Tables 7 and 8). Having verified the predictive capability of the CDD in these conserved plastid genes, all the other *rpoA* genes were queried against the database to determine whether the three interaction domains are predicted to be present.

In the case of the Annonaceae, all ORFs were predicted to encode the N-terminal domain of *rpoA*. Furthermore, all three interaction domains were predicted to be present in all nine *rpoA* genes, including those from *Annona*, *Asimina*, and *Cananga*, despite their substantial sequence divergence from outgroup *Chloranthus* (Table 8). *Annona*,

Asimina, and *Cananga* shared only 57%, 74.5%, and 55.9% pairwise sequence identity with *Chloranthus*, respectively.

In *Passiflora*, all ORFs were also predicted to encode the N-terminal domain of *rpoA*, and all three conserved domains were found in all 12 species (Table 8). *Passiflora* contained the *rpoA* genes with the most (*P. cirrhiflora*, 93.8% identity) and least (*P. biflora*, 53.6% identify) similarity to the outgroup *Populus*. In *Passiflora* the divergence was restricted to a single species surveyed, *P. biflora* (Figure 4). The other three *Passiflora* species contained *rpoA* genes more similar to *Populus* (~93%) than the genes from the other Malpighiales taxa, which ranged from 80% (*Turnera*) to 92% pairwise sequence identity (*Hevea*).

In *Pelargonium*, all ORFs were predicted to encode the N-terminal domain of *rpoA* as well as the homodimer interface. However, the conservation of functional domains showed a more complex pattern that differed by clade (Table 8). Clade B was the simplest; as all five representative *rpoA* sequences were predicted to contain all three functional domains despite retaining just 44%-49% sequence identity with outgroup *Eucalyptus*.

In *Pelargonium* clade A all nine representative *rpoA* genes were predicted to encode the N-terminal domain and homodimer interface, but the CDD search failed to locate the other two functional domains for two of the four species in clade A1 (*P. citronellum* and *P. cucullatum*) (Table 8). All five species from clade A2 were predicted to contain all three functional domains. Divergence from *Eucalyptus* in clade A is similar in magnitude to that in clade B, ranging from 44.9%-46.3% pairwise sequence identity. The two *rpoA* genes for which the CDD search did not predict the β and β` interface domains differed from those that were predicted to contain the domains by only three non-synonymous substitutions. When base 287 of *P. citronellum rpoA* was changed from

G to T, reversing the substitution of isoleucine for arginine, the CDD search was able to find all three functional domains. Separately, when base 425 of *P. citronellum rpoA* was changed from T to C, reversing the substitution of valine for alanine, the CDD search was once again able to predict all three functional domains. Reversing the third non-synonymous substitution had no effect on the CDD search prediction of functional domains. The fact that reversal of either one of the substitutions was sufficient to render all the functional domains recognizable by CDD search suggested that they are just beyond the threshold of recognition.

The *Pelargonium* C clade contained the most divergent and puzzling *rpoA*-like ORFs with respect to the prediction of conserved functional domains. From the taxa representing clade C1 all five were predicted to encode the N-terminus of *rpoA* and the homodimer interface, which spans the beginning and end of the *rpoA* N-terminus (Supplementary Figure 2), but only *P. tetragonum* and *P. worcesterae* were predicted to contain the other two functional domains, which were more dispersed through the N-terminus (Supplementary Figures 3 and 4). *Pelargonium tetragonum* and *P. worcesterae* had the highest pairwise sequence identity to the outgroup and at 912 bp were closest in length to *rpoA* in most angiosperms (versus 1014 bp in *Eucalyptus*), whereas the other three C1 taxa had shorter genes of 708-750 bp.

Using high-coverage Illumina sequence data we found two sequencing errors in the *rpoA*-like ORFs of the published *P. x hortorum* plastid genome annotation (Chumley et al. 2006). Both errors were single base pairs missing from ORFs, leading to a premature stop codon (ORF578) and to the division of one long ORF into two shorter ORFs (ORF521). The re-annotation of these ORFs was confirmed by comparison with those from the other three closely related taxa from *section Ciconium* included in the data set. After correction we found that these four genomes each contained three long *rpoA*-

like ORFs of similar length (1566 bp, 1737 bp, and 1794 bp in *P. x hortorum*; Table 9 and Fig. 9). These ORF names were used for the homologous ORFs in the other three *section Ciconium* species, even though some differed in length. Homology of these ORFs was inferred based on synteny. Although these ORFs were considerably longer than the *rpoA* genes in most angiosperms, the functional domains were nonetheless still predicted to reside in the N-terminus encoded by the first 800 bp of the ORFs.

Of the two species containing two *rpoA*-like ORFs, all ORFs were predicted to encode the N-terminal domain of *rpoA* and the homodimer interface, yet neither contained ORFs predicted to contain the other two functional domains (Table 8). *Pelargonium transvaalense* contained three ORFs predicted to encode the N-terminal domain of *rpoA* and the homodimer interface but not the other two functional domains. However, in the four *section Ciconium* taxa, at least one of the ORFs in each species was predicted to contain all three functional domains. In *section Ciconium*, one homolog, called ORF578 for its length in amino acids in *P. x hortorum* after re-annotation, was predicted to contain all domains in all four taxa (Table 8). Although the length of the other two ORFs varied between species, ORF578 was identical in length at 1737 bp in all four taxa and also displayed the highest percentage (99%) of identical sites across the four species.

Gene Conversion

The likelihood tree generated from clade C2 *rpoA*-like ORFs showed a pattern suggesting that gene conversion was an important phenomenon underlying the evolution of these unusual ORFs (Figure 5). First, ORFs from the two taxa containing only two ORFs grouped together by species and not by ORF, suggesting that these ORFs have not been evolving independently since their duplication in the ancestor of C2 taxa. For

example, the two ORFs in *P. endlicherianum* shared only 63-69% sequence identity with those from *P. spinosum*, whereas the ORFs in each species shared 85.7% and 72.0% identity with its paralog, respectively. The three ORFs in *P. transvaalense* grouped together as well, despite their apparent common ancestry with the ORFs in *section Ciconium*. For the four *section Ciconium* taxa, the ORFs grouped by ORF in the likelihood tree and not by species, despite showing evidence of gene conversion among ORFs, likely reflecting the relatively recent divergence of these taxa. The alignment of all three ORFs for all four species revealed regions of identity among ORFs in a single species, more similar to one another than to homologous ORFs from other species, suggesting that gene conversion played a role in the evolution of these ORFs as well.

ORGCONV (Hao 2010) found evidence of recombination among ORFs in all four species (Table 5). It predicted that gene conversion took place in the same region in all species, a region of approximately 600 bp from roughly the 120th to 720th base pair of the alignment of the three ORFs for each species. This was the region predicted by the CDD to encode the N-terminus of *rpoA*, which contains the functional domains. To investigate more closely the phenomenon of gene conversion among paralogs, a manual count of mutations potentially resulting from gene conversion was compiled. A simple parsimony criterion was used—substitutions common to multiple ORFs within a species but not among homologous ORFs across species were marked as results of putative gene conversion events (Table 6). Both ORGCONV and the manual, parsimony-based count indicated that gene conversion occurred among paralogs in all four *section Ciconium* species.

Table 3.1a. Taxon sampling for Annonaceae and Passiflora data sets.

Magnoliales	**Published**
Annona cherimola	Present study
Asimina incana	Present study
Calycanthus floridus	Gormekin et al., 2003
Cananga odorata	Present study
Chloranthus spicatus	Hansen et al., 2007
Drimys granadensis	Cai et al., 2006
Liriodendron tulipifera	Cai et al., 2006
Magnolia kwangsiensis	Kuang et al., 2011
Piper cenocladum	Cai et al., 2006
Malpighiales	
Hevea brasiliensis	Tangphatsornruang et al., 2011
Jatropha curcas	Asif et al., 2010
Linum usitatissimum	Present study
Manihot esculenta	Daniell et al., 2008
Oxalis latifolia	Wang et al., 2009
Passiflora biflora	Present study
Passiflora ciliata	Xi et al., 2012
Passiflora cirrhiflora	Present study
Passiflora quadrangularis	Present study
Populus trichocarpa	Direct submission, NCBI Genome Project
Ricinus communis	Rivarola et al., 2011
Turnera ulmifolia	Xi et al., 2012

Table 3.1b. Taxon sampling for *Pelargonium* data set.

Outgroups	
Eucalyptus globulus	D. Steane, 2005
Francoa sonchifolia	Weng et al., 2013
Melianthus villosus	Weng et al., 2013
Viviana marifolia	Weng et al., 2013
Hypseocharis bilobata	Weng et al., 2013
Clade A1	
Pelargonium citronellum	Present study
Pelargonium cucullatum	Present study
Pelargonium nanum	Present study
Pelargonium quercifolium	Present study
Clade A2	
Pelargonium alternans	Weng et al., 2013
Pelargonium echinatum	Present study
Pelargonium fulgidum	Present study
Pelargonium incrassatum	Present study
Pelargonium luridum	Present study
Clade B	
Pelargonium australe	Present study
Pelargonium cotyledonis	Guisniger et al. 2008
Pelargonium exstipulatum	Present study
Pelargonium grossularioides	Present study
Pelargonium reniforme	Present study
Clade C1	
Pelargonium dolomiticum	Present study
Pelargonium trifidum	Present study
Pelargonium myrrhifolium	Present study
Pelargonium tetragonum	Present study
Pelargonium worcesterae	Present study
Clade C2	
Pelargonium endlicherrianum	Present study
Pelargonium spinosum	Present study
Pelargonium transvaalense	Present study
Clade C2, sect. Ciconium	
Pelargonium alchemilloides	Present study
Pelargonium quinquelobatum	Present study
Pelargonium tongaense	Present study
Pelargonium xhortorum	Chumley et al., 2006

Table 3.2. *dN*/*dS* values from PAML for the branches of interest for the Annonaceae data set for all seven genes examined. The one value >1 is highlighted in bold.

Gene	**Annonaceae**	**Cananga**	**Asimina/Annona**	**Asimina**	**Annona**
matK	0.576	0.9757	0.2575	**1.0069**	0.2885
ndhF	0.3151	0.2535	0.2424	0.2423	0.2199
rbcL	0.0176	0.0757	0.0163	0.0776	0.2553
rpoA	0.7699	0.459	0.1851	0.6206	0.677
rpoB	0.1458	0.2798	0.2414	0.4706	0.4375
rpoC1	0.1748	0.287	0.139	0.1379	0.1156
rpoC2	0.4412	0.451	0.5286	0.2245	0.504

Table 3.3. *dN*/*dS* values from PAML for the branches of interest for the *Passiflora* data set for all seven genes examined. The one value >1 is highlighted in bold.

Gene	**Passifloraceae**	***Passiflora***	***P. ciliata*/*P. quad.***	***P. ciliata***	***P. quad.***	***P. cirrh.*/*P. biflora***	***P. cirrh.***	***P. biflora***
matK	0.5076	0.2837	0.2838	0.3646	0.4653	0.2081	0.3675	0.5375
ndhF	0.1312	0.2446	0.5146	0.1423	0.233	0.264	0.2846	0.3818
rbcL	0.1067	0.1967	0.0001	0.4484	0.9971	0.0001	0.3368	0.0641
rpoA	0.1352	0.2498	0.4067	0.1219	0.4935	0.0001	*	0.6263
rpoB	0.1184	0.2245	0.2742	0.1536	0.0895	0.4435	0.1778	0.8327
rpoC1	0.1272	0.2414	0.5605	0.0488	**1.2312**	*	0.1577	0.9736
rpoC2	0.3114	0.3637	0.3825	0.3777	0.505	0.6809	0.2248	0.9739

*Error when dS = 0

Table 3.4. dN/dS values from PAML for the branches of interest for the Pelargonium data set for all seven genes examined. Results from multiple alignment algorithms are given to show consistency of results. Values >1 are in bold. The one erroneous value (>50) is in bold italics.

Gene/Alignment	Geraniaceae	*Pelargonium*	Clade A/B	Clade A	Clade B	Clade C1
rpoA						
MAFFT	0.4417	0.2895	0.2717	0.4277	0.4405	0.1558
MUSCLE	***313.6861***	0.2241	0.3564	0.3917	0.5216	0.1157
CLUSTALW	0.5225	0.3317	0.3194	0.4756	0.4073	0.1095
Geneious	0.4797	0.306	0.2991	0.4364	0.4913	0.1149
rpoB						
MAFFT	0.3271	0.3397	**1.0804**	**5.9216**	**1.5002**	0.8164
MUSCLE	0.3153	0.3824	**1.1152**	**5.2133**	**1.3523**	0.813
CLUSTALW	0.3144	0.385	**1.0178**	**4.7961**	**1.4439**	0.8201
Geneious	0.3336	0.3451	**1.1462**	**2.5807**	**1.4648**	0.8637
rpoC1						
MAFFT	0.3249	0.5775	**1.8545**	**2.2963**	**4.1755**	**1.9503**
MUSCLE	0.3278	0.5671	**1.9809**	**2.53**	**4.0693**	**1.9034**
CLUSTAL	0.3301	0.5992	**1.9068**	**2.3202**	**2.7216**	**1.8449**
Geneious	0.3234	0.5772	**1.6394**	**1.6203**	**3.2839**	**1.8638**
rpoC2						
MAFFT	0.2796	0.3804	**2.1478**	0.9408	0.8104	**1.5229**
MUSCLE	0.2835	0.3717	**2.1931**	0.9949	0.795	**1.385**
CLUSTALW	0.2713	0.3887	**2.3449**	0.9751	0.8003	**1.4722**
Geneious	0.28	0.3679	**2.1464**	**1.1315**	0.7682	**1.271**
matK						
MAFFT	0.4176	0.3076	0.2874	0.4873	0.4892	0.1306
MUSCLE	0.4132	0.3116	0.2891	0.4883	0.4905	0.1325
CLUSTALW	0.4794	0.3489	0.2929	0.4944	0.5338	0.1331
Geneious	0.453	0.3212	0.2881	0.4864	0.5249	0.1313
ndhF						
MAFFT	0.1602	0.2563	0.3346	0.2362	0.4186	0.4322
MUSCLE	0.1598	0.2564	0.3438	0.2312	0.4254	0.4368
CLUSTAL	0.1706	0.2531	0.3377	0.2362	0.3328	0.4291
Geneious	0.1402	0.2453	0.339	0.2348	0.4136	0.4715
rbcL						
MAFFT	0.0304	0.001	0.0634	0.001	0.1309	0.2651
MUSCLE	0.0304	0.001	0.0634	0.001	0.1309	0.2651
CLUSTALW	0.0304	0.001	0.0634	0.001	0.1309	0.2651
Geneious	0.0304	0.001	0.0634	0.001	0.1309	0.2651

Table 3.5. Gene conversion events detected by ORGCONV. The donor and acceptor of each putative gene conversion event are given along with the coordinates of the converted region and the p-value of the conversion event.

Converted Sequence	Donor	Start	End	Pvalue(L/N)	Pvalue(L-N)
P_alchemilloides_ORF597	P_alchemilloides_ORF521	130	713	1.13E-07	5.14E-06
P_quinquelobatum_ORF597	P_quinquelobatum_ORF521	119	713	2.30E-10	1.09E-08
P_tongaense_ORF597	P_tongaense_ORF521	124	713	5.52E-10	2.74E-08
Pxhortorum_ORF578	Pxhortorum_ORF521	223	300	1.31E-03	5.20E-03
Pxhortorum_ORF597	Pxhortorum_ORF521	669	713	1.03E-03	1.77E-02

Table 3.6. Gene conversion events detected by manual count from an alignment of all 12 ORFs from the four *Pelargonium section Ciconium* species. Putatively converted bases (and one indel) are shown in red.

Alignment Coordinate (bp)	106	123	135	221	269	297	357	457	478	486	524	531	725	775	777	778	798	801	806	810	990	1012	1033	1043	1131	1166	1385	1685	1739	1803
P. alchemilloides_ORF521	G	A	G	A	+6bp	A	G	A	C	G	A	G	C	G	G	C	T	T	A	C	T	C	C	T	A	A	-	G	A	-
P. alchemilloides_ORF578	G	A	T	A		G	G	A	A	G	G	G	G	G	G	G	A	T	A	C	T	C	C	G	A	A	A	G	A	T
P. alchemilloides_ORF597	T	A	T	G	+6bp	A	G	A	A	G	G	G	T	G	G	G	A	T	A	C	T	C	C	G	A	C	A	A	G	G
P. quinquelobatum_ORF521	G	A	G	A	+6bp	A	G	C	C	G	T	A	C	T	C	G	A	G	G	T	G	G	A	T	A	A	A	T	A	-
P. quinquelobatum_ORF578	G	C	T	A		G	G	C	C	G	T	A	G	G	G	G	A	T	A	C	T	C	C	G	C	A	A	G	A	T
P. quinquelobatum_ORF597	T	A	G	G	+6bp	A	G	C	C	G	T	A	T	T	G	G	A	T	A	C	T	C	C	G	C	A	A	A	G	G
P. tongaense_ORF521	G	A	G	G		A	T	A	C	A	A	G	C	T	C	C	T	G	G	T	G	G	A	G	A	A	C	T	A	-
P. tongaense_ORF578	T	C	T	A		G	G	A	C	A	A	G	G	T	G	G	A	T	A	C	T	C	C	G	A	A	A	G	A	G
P. tongaense_ORF597	T	A	G	G	+6bp	A	T	A	C	G	G	G	G	T	G	G	A	T	A	C	T	C	C	G	A	C	A	A	G	G
P. xhortorum_ORF521	G	A	G	A		G	G	A	C	A	A	G	C	T	C	C	T	G	G	C	G	G	A	T	A	A	C	T	G	-
P. xhortorum_ORF578	G	C	T	A		G	G	A	C	G	G	G	G	T	G	G	A	T	A	C	T	C	C	G	A	A	A	G	A	T
P. xhortorum_ORF597	T	A	G	G	+6bp	A	G	A	C	G	G	G	T	T	G	G	A	T	A	C	T	C	C	G	A	C	A	A	G	G

Table 3.7. Summary of conserved domain database (CDD) search results for Annonaceae and *Passiflora* data sets. Predictions of *rpoA* N-terminus, homodimer interface, beta and beta prime interfaces are indicated (Y = Yes, N = No). The pairwise identity of each sequence with outgroup *Populus* or *Chloranthus* is given for nucleotide (nt) and amino acid (aa) alignments.

Annonaceae	**N-terminal**	**dimer**	**β**	**β'**	**nt identity**	**aa identity**	**ORF length**
Annona	Y	Y	Y	Y	57.0%	40.7%	1,035bp
Asimina	Y	Y	Y	Y	74.5%	63.8%	1,020bp
Calycanthus	Y	Y	Y	Y	89.0%	86.6%	1,020bp
Cananga	Y	Y	Y	Y	55.9%	38.5%	1,143bp
Chloranthus	Y	Y	Y	Y	100.0%	100.0%	1,002bp
Drimys	Y	Y	Y	Y	90.2%	85.4%	1,017bp
Liriodendron	Y	Y	Y	Y	91.3%	89.8%	1,014bp
Magnolia	Y	Y	Y	Y	91.8%	89.4%	1,014bp
Piper	Y	Y	Y	Y	85.7%	80.2%	1,017bp

Passifloraceae					**nt identity**	**aa identity**	**ORF length**
Hevea	Y	Y	Y	Y	92.0%	88.9%	1,023bp
Jatropha	Y	Y	Y	Y	91.8%	87.9%	1,017bp
Linum	Y	Y	Y	Y	86.9%	80.8%	1,011bp
Manihot	Y	Y	Y	Y	91.9%	87.8%	1,029bp
Oxalis	Y	Y	Y	Y	89.8%	85.3%	1,017bp
P. biflora	Y	Y	Y	Y	53.6%	37.4%	1,071bp
P. ciliata	Y	Y	Y	Y	93.4%	88.8%	1,005bp
P. cirrhiflora	Y	Y	Y	Y	93.8%	90.6%	1,017bp
P. quadrangularis	Y	Y	Y	Y	93.3%	89.1%	1,017bp
Populus	Y	Y	Y	Y	100.0%	100.0%	1,017bp
Ricinus	Y	Y	Y	Y	91.9%	88.0%	996bp
Turnera	Y	Y	Y	Y	80.7%	71.0%	891bp

Table 3.8. Summary of conserved domain database (CDD) search results for *Pelargonium* data set. Predictions of *rpoA* N-terminus, homodimer interface, beta and beta prime interfaces are indicated (Y = Yes, N = No). The pairwise identity of each sequence with outgroup *Eucalyptus* is given for nucleotide (nt) and amino acid (aa) alignments.

Outgroups	**N-terminal**	**dimer**	**β**	**β'**	**nt identity**	**aa identity**	**ORF length**
Eucalyptus	Y	Y	Y	Y	100.0%	100.0%	1,014bp
Francoa	Y	Y	Y	Y	92.2%	85.8%	1,020bp
Melianthus	Y	Y	Y	Y	91.3%	84.4%	1,020bp
Viviana	Y	Y	Y	Y	84.0%	73.2%	1,014bp
Hypseocharis	Y	Y	Y	Y	77.8%	65.4%	1,089bp
Clade A1							
P. citronellum	Y	Y	N	N	46.0%	31.6%	885bp
P. cucullatum	Y	Y	N	N	46.0%	31.6%	885bp
P. nanum	Y	Y	Y	Y	46.2%	31.9%	885bp
P. quercifolium	Y	Y	Y	Y	46.0%	31.6%	885bp
Clade A2							
P. alternans	Y	Y	Y	Y	46.3%	31.6%	885bp
P. echinatum	Y	Y	Y	Y	45.1%	32.0%	858bp
P. fulgidum	Y	Y	Y	Y	44.9%	31.4%	855bp
P. incrassatum	Y	Y	Y	Y	46.2%	32.2%	885bp
P. luridum	Y	Y	Y	Y	46.2%	31.6%	885bp
Clade B							
P. australe	Y	Y	Y	Y	44.0%	34.1%	828bp
P. cotyledonis	Y	Y	Y	Y	48.8%	33.8%	945bp
P. exstipulatum	Y	Y	Y	Y	46.0%	33.0%	879bp
P. grossularioides	Y	Y	Y	Y	45.1%	31.0%	864bp
P. reniforme	Y	Y	Y	Y	46.2%	32.7%	885bp
Clade C1							
P. dolomiticum	Y	Y	N	N	34.1%	24.4%	750bp
P. trifidum	Y	Y	N	N	32.6%	23.5%	708bp
P. myrrhifolium	Y	Y	N	N	35.1%	25.2%	714bp
P. tetragonum	Y	Y	Y	Y	46.0%	29.8%	912bp
P. worcesterae	Y	Y	Y	Y	46.2%	30.1%	912bp

Table 3.8. Summary of conserved domain database (CDD) search results for *Pelargonium* data set. Predictions of *rpoA* N-terminus, homodimer interface, beta and beta prime interfaces are indicated (Y = Yes, N = No). The pairwise identity of each sequence with outgroup *Eucalyptus* is given for nucleotide (nt) and amino acid (aa) alignments.

Outgroups	N-terminal	dimer	β	β'	nt identity	aa identity	ORF length
Clade C2							
P. endlicherrianum_578	Y	Y	N	N	32.3%	25.0%	1,701bp
P. endlicherrianum_597	Y	Y	N	N	31.6%	25.1%	1,737bp
P. spinosum_578	Y	Y	N	N	30.1%	19.9%	1,773bp
P. spinosum_597	Y	Y	N	N	35.0%	23.0%	1,479bp
P. transvaalense_521	Y	Y	N	N	30.0%	18.7%	1,863bp
P. transvaalense_578	Y	Y	N	N	31.4%	25.7%	1,755bp
P. transvaalense_597	Y	Y	N	N	31.4%	25.0%	1,779bp
Clade C2, sect. Ciconium							
P. alchemilloides_521	Y	Y	N	N	34.8%	33.5%	702bp
P. alchemilloides_578	Y	Y	Y	Y	31.8%	15.9%	1,737bp
P. alchemilloides_597	Y	Y	N	N	30.7%	14.8%	1,794bp
P. quinquelobatum_521	Y	Y	N	N	33.6%	20.6%	1,560bp
P. quinquelobatum_578	Y	Y	Y	Y	32.0%	15.9%	1,737bp
P. quinquelobatum_597	Y	Y	N	N	30.7%	15.3%	1,794bp
P. tongaense_521	Y	Y	Y	Y	33.0%	20.6%	1,554bp
P. tongaense_578	Y	Y	Y	Y	31.9%	15.7%	1,737bp
P. tongaense_597	Y	Y	N	N	30.3%	14.9%	1,815bp
P. xhortorum_521	Y	Y	Y	Y	33.4%	21.1%	1,566bp
P. xhortorum_578	Y	Y	Y	Y	31.9%	15.9%	1,737bp
P. xhortorum_597	Y	Y	N	N	30.6%	14.9%	1,794bp

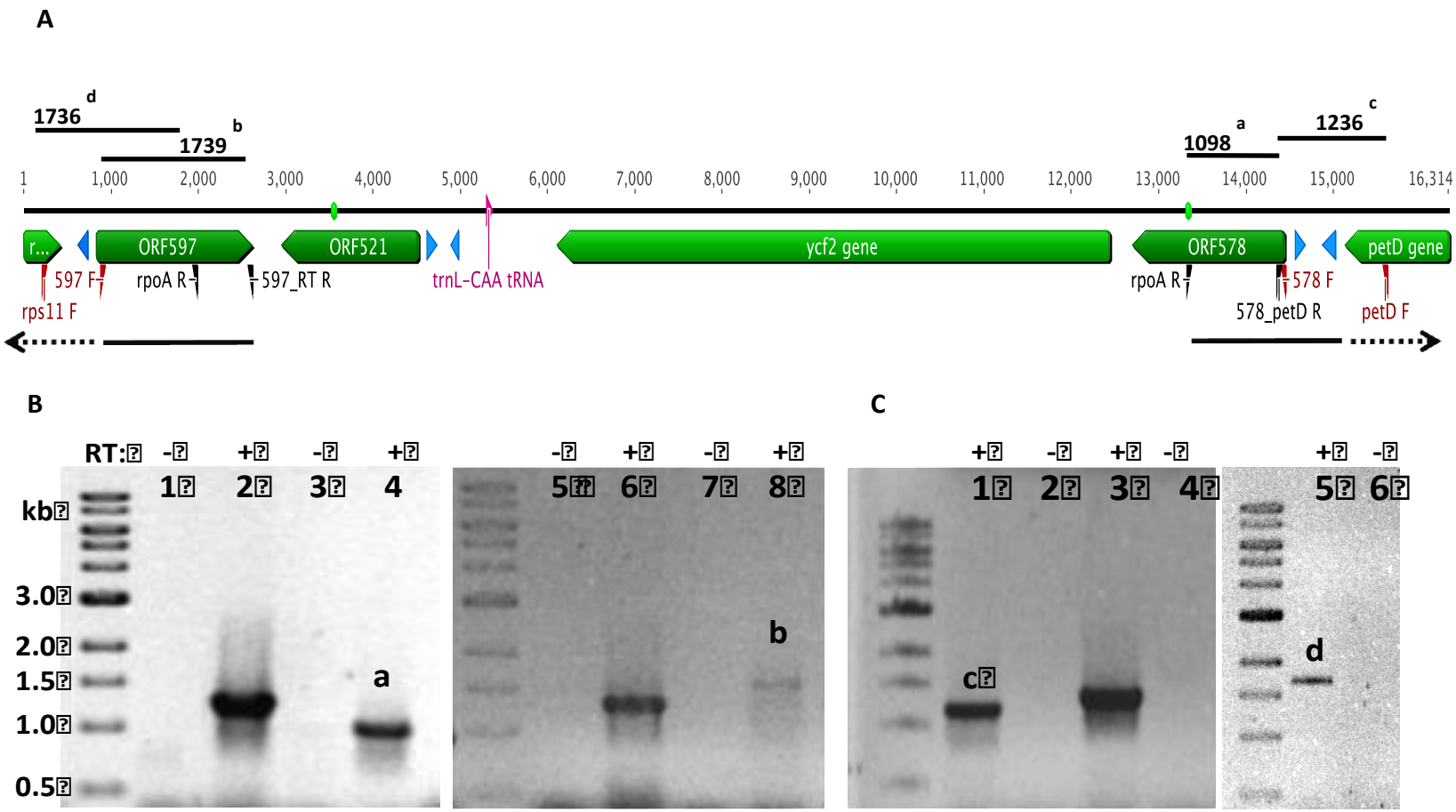

Figure 3.1a. Location of RT-PCR primers for amplification of *P. x hortorum* ORF578 and ORF597 transcripts.

Figure 3.1b. Agarose gel showing RT-PCR products for *P. x hortorum rpoA* ORFs. a,b) products representing monocistronic transcripts of ORF579 and ORF597, respectively. c,d) products representing dicistronic transcripts ORF579 and ORF597, respectively.

Figure 3.2a. Alignment of PEP promoter region for rbcL in three species with functional PEP (Nicotiana tabacum, Arabidopsis thaliana, P. x hortorum) and one lacking PEP (Cuscuta obtusiflora

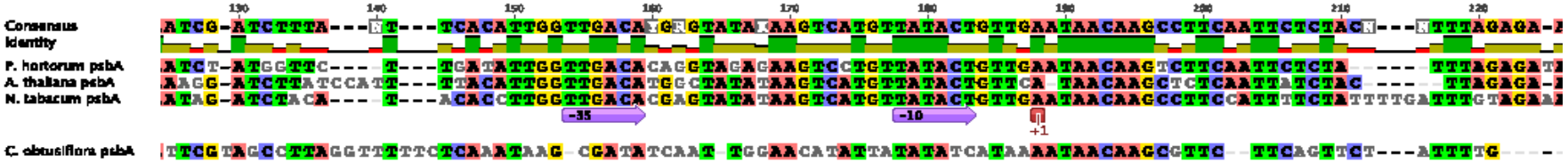

Figure 3.2b. Alignment of PEP promoter region for psbA in three species with functional PEP (Nicotiana tabacum, Arabidopsis thaliana, P. x hortorum) and one lacking PEP (Cuscuta obtusiflora).

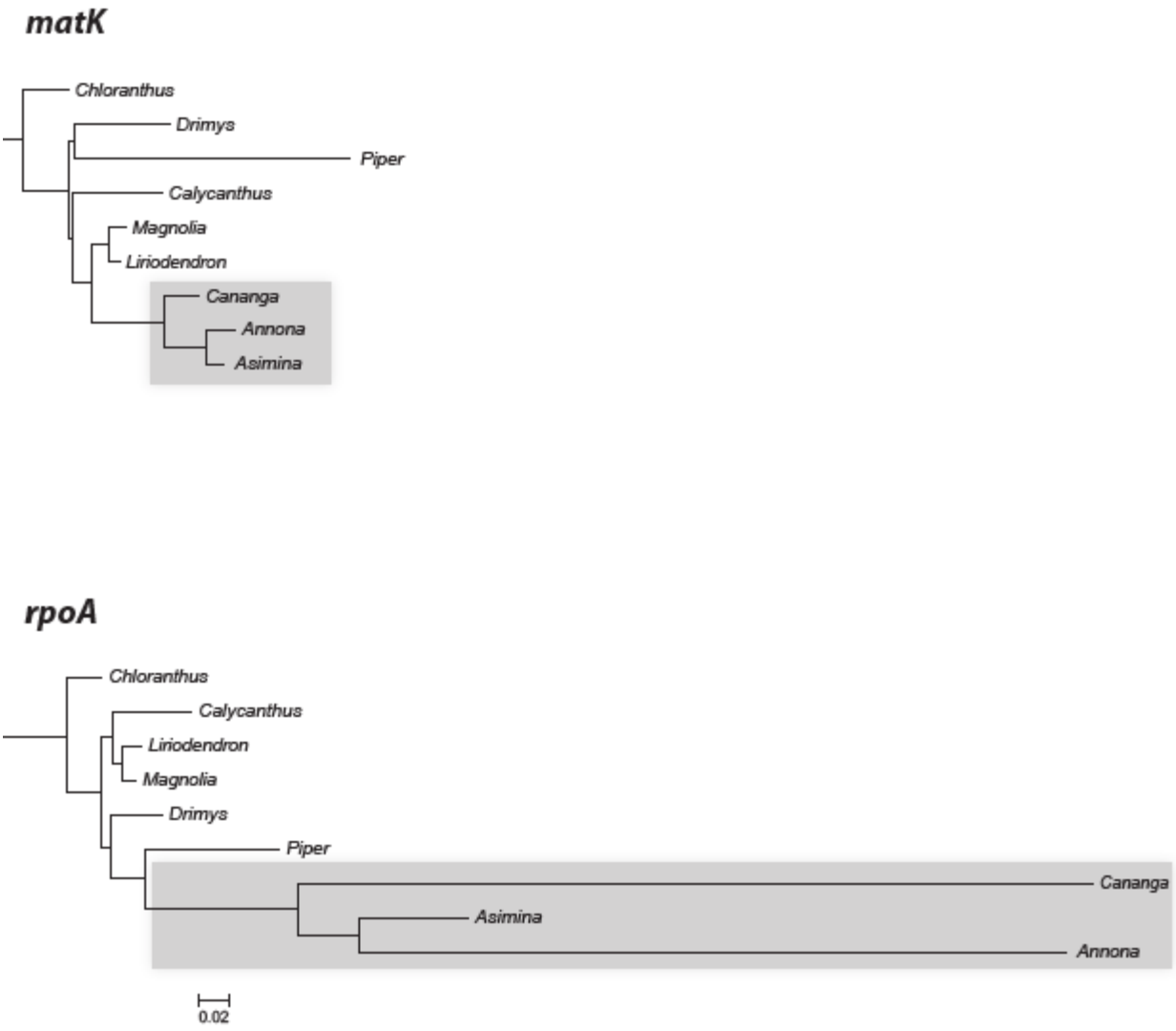

Figure 3.3.. Maximum likelihood trees generated from *matK* (top) and *rpoA* (bottom) for the nine Magnoliales taxa comprising the Annonaceae data set. Likelihood scores for the *matK* and *rpoA* trees were -5638.2661 lnL and -5506.0125 lnL, respectively. Annonaceae species are boxed in gray.

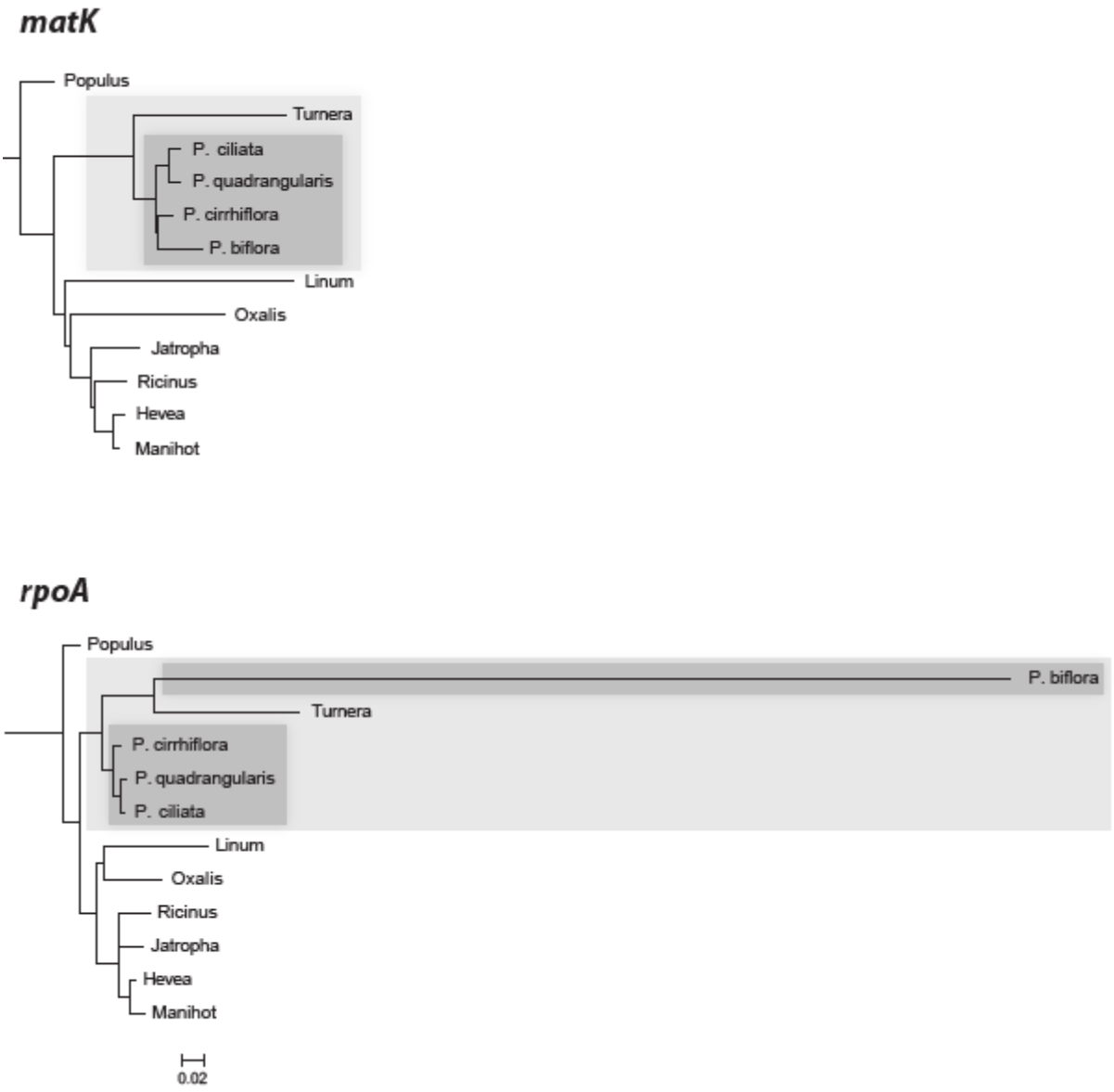

Figure 3.4. Maximum likelihood trees generated from matK (top) and *rpoA* (bottom) for the 12 Malpighiales taxa comprising the *Passiflora* data set. Likelihood scores for the *matK* and *rpoA* trees were lnL -7243.3426 and -4810.9045 lnL, respectively. *Passiflora* species are boxed in dark gray and Passifloraceae in light gray.

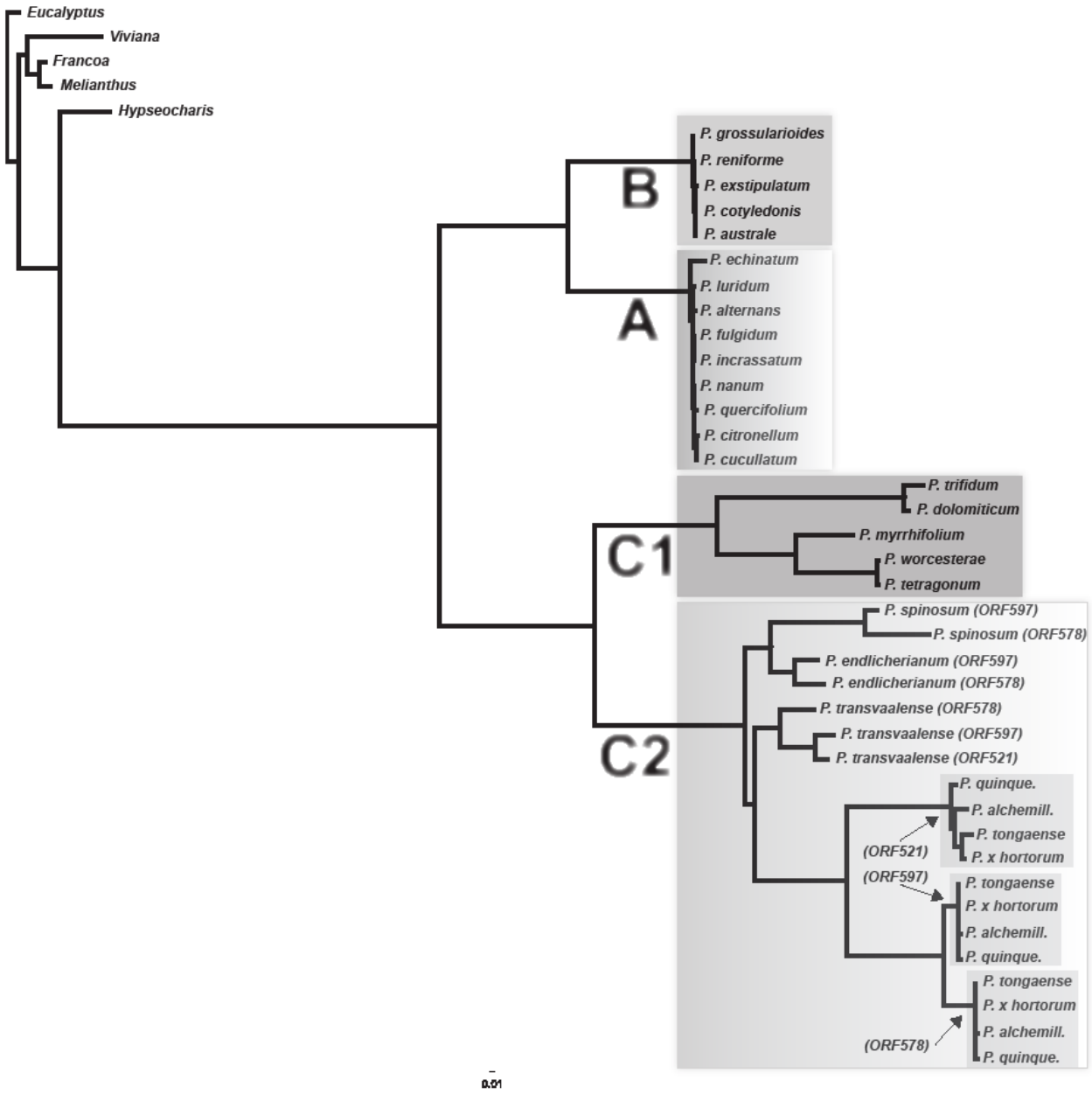

Figure 3.5. Maximum likelihood tree generated from all 43 *rpoA* ORFs from 26 *Pelargonium* species with likelihood score -19243.4093 lnL. Species in clade C2 contain either two (*P. spinosum* and *P. endlicherianum*) or three (*P. transvaalense* and four species from *section Ciconium*) *rpoA* paralogs.

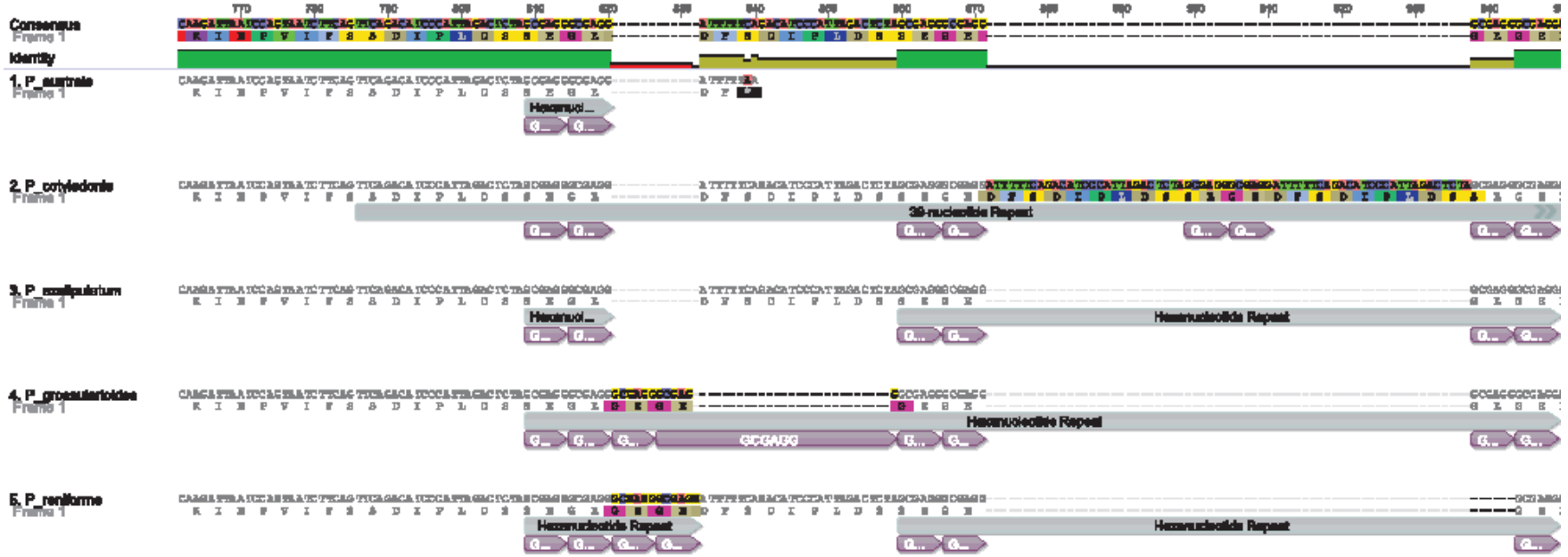

Figure 3.6. Figure 6. Diagram of the 3’ end of the MAFFT alignment of clade B *rpoA* genes. A 6 bp tandem repeat found in all species is shown as dark purple arrows, and a larger 39 bp tandem repeat only present in *P. cotyledonis* also containing the 6 bp repeat is shown in light purple.

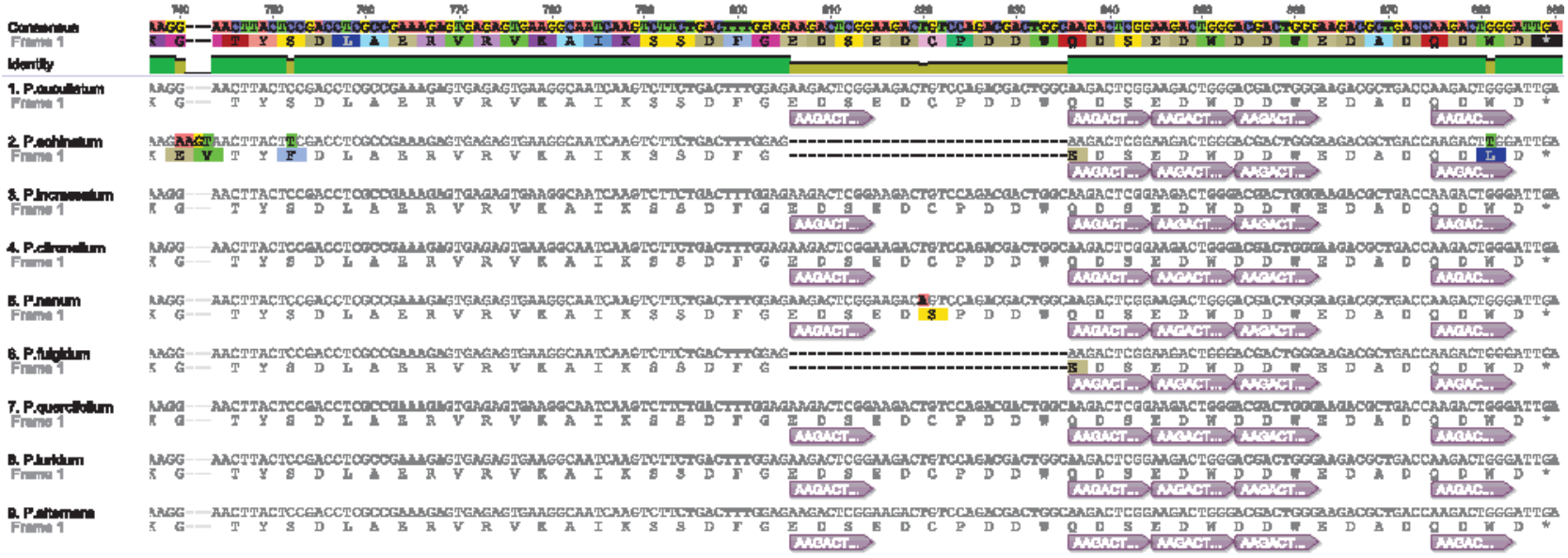

Figure 3.7. Diagram of the 3' end of the MAFFT alignment of clade A *rpoA* genes. Tandem repeats are shown as purple arrows, and deletions in *P. fulgidum* and *P. echinatum* are indicated by dashes in the coding sequences.

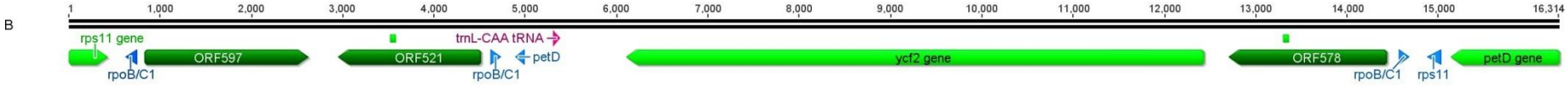

Figure 3.8. Diagram of the three corrected *rpoA*-like ORFs in the *P. xhortorum* plastid genome. The former ORF574 was assigned an upstream ATG start codon (instead of the previously annotated ATT alternative start codon) and renamed ORF597 to reflect the new length in amino acids. ORF221 and ORF332 were joined into a single ORF, ORF521, after a sequencing error was corrected—the insertion of a missing base pair is indicated as a small green square above the ORF. The former ORF365 had a sequencing error, a missing base pair, that caused a frameshift and premature stop codon—after correction the gene is similar in length to the other two ORFs and has been renamed ORF578. The insertion of a missing base pair is indicated as a small green square above the ORF.

Chapter 4: Plastid genome rearrangement and re-growth of the large inverted repeat in *Erodium* (Geraniaceae)

INTRODUCTION

The majority of angiosperms harbor plastid genomes with a nearly identical gene complement, gene order, and quadripartite structure (Bock 2007); however, early restriction fragment mapping studies identified two independent lineages in which one copy of the large, usually 25 kilobase (kb) inverted repeat (IR) had been lost (Downie and Palmer 1992): the genus *Erodium* (Geraniaceae) and a large clade with legumes (Fabaceae) termed the IR Lacking Clade (IRLC) (Wojciechowski et al. 2004). The IR is recombinogenic, and intermolecular recombination between IRs in the highly polyploid plastids gives rise to isomers that differ in the relative orientation of their single copy regions, termed the large and small single copy regions (LSC and SSC, respectively)(Palmer 1983). The function of the IR is unknown, but it is highly conserved across angiosperms and is present in many other land plant and algal lineages as well. With a single exception, *Monsonia speciosa* (Geraniaceae), the IR contains the entire ribosomal operon (Guisinger et al. 2011), and some algae with reduced IRs contain only the ribosomal operon (Yamada 1991). It has also been demonstrated that the primary origins of replication lie within the IR (Kunnimalaiyaan and Nielsen 1997). Several functions for the IR have been hypothesized, including replication (Heinhorst and Cannon 1993), stabilization of the plastid genome (Palmer and Thompson 1982; Hirao et al. 2008), and conservation of the translational machinery (Palmer and Thompson 1982), since genes in the IR have been shown to evolve threefold more slowly than those in the single-copy regions (Wolfe et al. 1987; Perry and Wolfe 2002). Although wholesale loss of the IR is limited to two angiosperm lineages, movement of the IR boundaries is more

common and is thought to occur through illegitimate recombination between opposite junctions of the IR and single copy regions (Goulding et al. 1996).

In Solanaceae, an unusual expansion of the IR into the LSC in *Nicotiana acuminata* was found to have occurred through illegitimate recombination and gene conversion between IR/LSC junctions ending in poly(A) tracts (Goulding et al. 1996). Presumably contraction of the IR can occur through a similar mechanism. Unfortunately, the losses of the IR in *Erodium* and some legumes did not leave behind clear footprints of illegitimate recombination—that is, clear homology on both sides of the tract that underwent gene conversion—or at least none remains, as neither loss is recent. Nonetheless it is easy to imagine the loss of the IR through illegitimate recombination in a manner similar to expansion of the IR in *N. acuminata*: illegitimate recombination between direct repeats in the LSC and SSC would cause the loss of the intervening IR. As no unique genes would be lost from the plastid genome through the deletion of one IR, there might be no immediate fitness cost associated with the deletion, and the reduced genome could have a replicative advantage over the copies containing the IR (Selosse et al. 2001).

A single lineage, the geranium family (Geraniaceae), contains not only one of the two angiosperm lineages lacking the IR (*Erodium*) but also the lineage with the largest known IR (*Pelargonium*) as well as the only lineage in which the IR has contracted such that it does not contain the entire ribosomal operon (*Monsonia*). Since the ancestral plastid genome organization for Geraniaceae has been inferred to include a relatively normal sized IR, these large fluctuations appear to have occurred independently in each genus (Weng et al. in press). Given the rarity of plastid genome rearrangement across angiosperms it is tempting to hypothesize that the rearrangement we see in the major

genera of Geraniaceae, through independent and different in outcome, share an underlying mechanism.

This study focuses on plastid genome evolution in *Erodium*, the second angiosperm lineage found to have lost the IR. *Erodium* plastid genomes have received considerably less attention, likely due to the great economic importance of legume crops. The observation that legumes lacking the IR have undergone more frequent genomic rearrangement than those retaining the IR led to an early hypothesis that the IR functioned to stabilize the plastid genome (Palmer and Thompson 1982). And indeed the publication of the first *Erodium* plastid genome, *Erodium texanum*, seemed to support this hypothesis since it is highly rearranged (Guisinger et al. 2011). However, the subsequent publication of plastid genomes of monotypic sister species *California macrophylla* and *Erodium carvifolium* (Blazier et al. 2011; Weng et al. in press), a representative of the other major clade within the genus, did not support the hypothesis of IR genome stabilization. First, the *C. macrophylla* plastid genome shows few unique rearrangements compared to the ancestral angiosperm gene order. Second, *E. carvifolium* closely resembles *C. macrophylla* except that one copy of the IR (IRa, the copy adjacent to *trnH* and *psbA*) has been deleted leaving less than 300 bp between the two former LSC and SSC regions. Together these two plastid genomes demonstrate that the extensive rearrangement in all major genera within Geraniaceae has occurred separately and that IR loss is not necessary or sufficient to destabilize the conserved gene order. In fact, next to *C. macrophylla*, the *Erodium* plastid genomes from clade II (that containing *E. carvifolium*) are among the least rearranged plastid genomes in the family. And conversely, the great expansion of the IR in *Pelargonium*, as demonstrated by the genome of *P. x hortorum*, accompanied the greatest rearrangement.

In addition to plastid genomes, the quadripartite structure consisting of two single copy regions separated by inverted repeats has also been found in herpes simplex virus 1 (HSV-1) (Horiuchi and Watanabe 2011; Lehman and Boehmer 1999) and in the 2-micron circle plasmids of *Saccharomyces cerevisiae* (Broach and Volkert 1991). Though the patterns of IR expansion, loss, and genomic rearrangement in Geraniaceae plastid genomes do not support the genome stability hypothesis for maintenance of the IR, convergence on this genome architecture in a virus, a plasmid, and in plant organellar genomes suggests that it may have some functional significance. The re-appearance of a new IR and quadripartite structure in the plastid genome of *E. gruinum* provides additional evidence that this configuration may have a conserved function. Demonstrating a precedent for loss and regain of the IR also impacts models of evolution for other highly rearranged plastid genomes.

We categorize *Erodium* plastid genomes into four types and describe distinct evolutionary trajectories in the two major clades in *Erodium*. For both clades we describe gene and intron losses and relate genomic rearrangements to repeat content. We analyze rates of evolution on the long branch leading to *E. gruinum* and compare patterns of nucleotide substitution between the two major clades.

MATERIALS AND METHODS

Taxon Sampling

The data set consists of 19 species, 16 whole plastid genomes and three draft genomes from which proteins-coding genes were extracted for evolutionary rate analysis (Table 1); the protein-coding genes of *E. chrysanthum* were previously published (Guisinger et al. 2008). Of the 16 whole plastid genomes, three were previously published (Blazier et al. 2011; Guisinger et al. 2011; Weng et al. in press). The thirteen

new taxa were chosen based on a molecular phylogeny of the genus to ensure that all clades are represented (Fiz et al. 2006) (Figure 1). Plants were obtained from commercial sources (Geraniaceae.com and B & T World Seeds) or grown from seed provided by J.J. Aldasoro.

DNA Isolation and Sequencing

For pyrosequenced genomes, plastids were isolated and plastid DNA was amplified as previously described (Jansen et al. 2005; Blazier et al. 2011). Sequencing was conducted on the 454 FLX platform at the W. M. Keck Center for Comparative and Functional Genomics at University of Illinois.

For Illumina-sequenced genomes, total genomic DNA was isolated from fresh leaf tissue using a modified version of hexadecyltrimethyl-ammonium bromide procedure in Doyle and Doyle (1987) (Weng et al. in press). Total genomic DNA was sequenced using Illumina HiSeq 2000 at Beijing Genomics Institute Corporation or the Genome Sequence and Analysis Facility (GSAF) at the University of Texas at Austin. For each species, about 60 million 100 bp paired-end reads were generated from a sequencing library with ~750 bp insert.

Genome Assembly and Annotation

Pyrosequenced genomes were assembled de novo in the native 454 assembler (Newbler) under default settings as well as in MIRA v.3.4 using the "accurate" setting (Chevreux et al. 1999).

Illumina data was assembled de novo with Velvet v. 1.2.07 (Zerbino and Birney 2008) using a range of kmer sizes from 71 to 93, with and without scaffolding enabled. Plastid contigs were identified by BLAST search against a database of Geraniaceae plastid protein-coding genes using custom Python scripts. Further, nuclear and

mitochondrial contigs containing plastid DNA insertions were excluded using 1000x coverage cutoff. Assembly and filtering were performed on the Lonestar Linux Cluster from the Texas Advanced Computing Center (TACC).

For both data types, contigs were assembled and edited in Geneious 7.0.4 by Biomatters and annotated in DOGMA (Wyman et al. 2004). Circular genome maps were generated using OGDraw (Lohse et al. 2007). Whole genome alignments were created using MAUVE as implemented in Geneious under default settings (Darling et al. 2004).

Verification of IR Boundaries

Presence of the IR was verified bioinformatically and empirically. For the bioinformatic verification, 10% of the paired-end Illumina reads were extracted at random from the 75 million read data set and mapped to the *E. gruinum* genome with BLAST. Using custom Perl scripts, only read pairs in which both reads mapped at the proper distance and in the proper orientation were counted, and read depth was calculated and plotted in Excel at 100 bp intervals. IR regions are expected to have twice the coverage as the single copy regions (Supplementary Figure 1). For the empirical verification, four sets of primers were designed to span the putative LSC/IR and SSC/IR junctions. Primer sequences are given in Supplementary Table 1. Sanger sequencing of PCR products was performed on an ABI 3730 platform at The University of Texas at Austin.

Sequence Analysis

For the 19 gene data set, genes were extracted from DOGMA and edited in Geneious. Genes were aligned in MAFFT (Katoh et al. 2009) as implemented in Geneious, and a concatenated alignment of all 19 genes (26,985 bp) was used to generate

a constraint tree in Garli (Zwickl 2008) under default settings as implemented in Geneious. For rates analyses, codon alignments were generated using MAFFT and the translation align function in Geneious.

Using the constraint tree, plastid genes were analyzed with codon-based models to quantify the rates of synonymous (*dS*) and nonsynonymous (*dN*) substitution. Analyses were conducted in PAML 4.7 (Yang 2007) using custom Python scripts on the Lonestar Linux Cluster at TACC. The F3×4 model was used to calculate codon frequencies, and the free-ratio model was used to compute *dN*/*dS* values. Transition/transversion and *dN*/*dS* ratios were estimated with the initial values of 2 and 0.4, respectively, consistent with other studies examining evolutionary rate heterogeneity in angiosperm organellar genomes (Sloan et al. 2009; Weng et al. 2012).

Rates Analyses

Likelihood ratio tests (LRTs) were conducted in HyPhy v2.1.1 Beta for Mac (Pond and Muse 2005) to detect whether the branch leading to the long-branch clade was significantly different from the other branches. The LRTs were conducted between two models, the null model with globally constrained *dS* shared by all branches and the alternative model with the branch leading to the long-branch clade free from this constraint. The same setting was applied to the LRTs for *dN*. Because the constraint tree used for estimating rates has 34 branches, the p-value was multiplied by a Bonferroni correction factor of 34 to account for multiple comparisons.

To test whether clades I and II have significantly different evolutionary rates, the internal branches from *dN* and *dS* trees were extracted and compared using the Wilcoxon rank sum test in R v2.15 (R Development Core Team).

Repeat Analyses

Repetitive DNA was identified by BLASTing each genome against itself using blastn under default parameters and an e-value of 1e-10. For the two genomes with an IR (*C. macrophylla* and *E. gruinum*), one copy of the IR was removed.

Prediction of tRNA genes

Genes encoding tRNAs were predicted in DOGMA under default settings. For two tRNA genes found to be missing from some species, *trnG-gcc* and *trnV-gac*, tRNAscan-SE (Schattner et al. 2005) was used under "cove-only" and "Organellar" parameters in order to verify the absence of these genes.

RESULTS

Genome Organization Overview

The data set contains 16 completed plastid genomes representing all major clades in *Erodium* (Figure 1, Table 1). *Erodium* plastid genomes fall into four types (Figure 2, Table 2): type 1 is represented by *E. texanum* (Guisinger et al., 2011) and differs from the other representative, *E. guttatum*, by a single inversion (Figure 3). Rearrangement in this type has been so severe—with an estimated 14 inversions in *E. texanum* (Guisinger et al., 2011)—that it is not possible to determine whether the inversion distinguishing the two species has taken place in *E. texanum* or *E. guttatum*. Type 2 is represented by *E. carvifolium* (Blazier et al., 2011) and nine additional genomes from Clade II; these genomes are mostly collinear but a few differ by one or two rearrangements. Type 3 is

represented by two new genomes, *E. crassifolium* and *E. jahandiezianum*, which are collinear (Figures 4 and 5). These genomes are much less rearranged than types 1 or 4, differing from the relatively unrearranged Type 2 genomes by an estimated five inversions (Figure 6). Type 4 is represented by a single species, *E. gruinum* (Figure 7). We have estimated that 10 inversions distinguish the type 4 genome from unrearranged type 2 genomes (Figure 8), but this is likely an underestimate since large portions of the *E. gruinum* genome no longer encode functional genes following the loss of 13 genes found in other Clade I species. Several additional regions of the *E. gruinum* plastid genome have likely undergone rearrangement, but their low sequence identity to other *Erodium* plastid genomes hinders the reconstruction of rearrangement events. Most strikingly, *E. gruinum* was found to contain a large 25 kb inverted repeat that must be a recent development since the IR is inferred to have been lost on the branch separating *Erodium* from *C. macrophylla*.

As representatives of type 1 and 2 genomes have been described in detail previously (Blazier et al., 2011; Guisinger et al., 2011), we will focus on the two new types, types 3 and 4, represented by *E. jahandiezianum* and *E. gruinum*, respectively. We will also provide a detailed comparison of the 10 plastid genomes representing type 2. Table 2 shows characteristics of the four types. Table 3 shows characteristics of all complete genomes from Clades I and II. Plastid genome types 1-3 provide evidence for a single loss of the IR in *Erodium*, while type 4 has undergone too many rearrangements in the vicinity of the former IR boundary to be informative.

Loss of the IR

The paucity of rearrangements separating the IR-containing *C. macrophylla* plastid genome from Clade II (Type 2) plastid genome *E. carvifolium* makes it clear that

the copy of the IR adjacent to *psbA-trnH* was deleted in this lineage and that this deletion was not accompanied by further changes in gene order. However, many gene and intron losses and genomic rearrangements show homoplasy in Geraniaceae. For this reason we have compared additional *Erodium* plastid genomes in order to establish that only one loss of the IR has occurred. Figure 9 is an alignment of the region that used to contain the copy of the IR (IRa) that is lost in all *Erodium* and has been reduced to just a few hundred base pairs in all Clade II species. Some Clade I species (types 1 and 3) contain additional sequence in this region including a pseudogene of *ndhA* (exon 1) and a full copy of *trnI-cau*, genes located on opposite ends of the IR in the inferred ancestral genome arrangement for all Geraniaceae genera (Weng et al. in press). Comparison of the region containing the former IR-SSC and IR-LSC boundaries (*ccsA-trnL-uag* and *trnH-gug-psbA*, respectively) in types 1, 2, and 3 supports a single loss of the IR in *Erodium*, with greater reduction of the resulting intergenic region in type 2.

Plastid genome type 3: *Erodium jahandiezianum*

The plastid genome of *E. jahandiezianum* (Figure 4) is relatively small (121,692 bp) and is just 300 bp larger than the other completed type 3 genome, *E. crassifolium*. These two genomes are identical in gene order, gene and intron content (Figure 5, Table 3), and GC content (39.1%). *Erodium jahandiezianum* contains a slightly lower proportion of repetitive DNA (2.61% versus 3.64%), but both genomes contain less repetitive DNA than the other genome types from clade I (Table 3). Based on a MAUVE alignment of *E. jahandiezianum* and the simplest representative of Clade II (*E. foetidum ssp. foetidum*), we estimate that five inversions are required to derive the gene order found in type 3 plastid genomes (Figure 6). As in *E. texanum*, the *rps2-atpA* transcriptional unit has been split in two. It is difficult to judge whether this

rearrangement is shared between types 1 and 3. *Ycf3* flanks *atpI* in both types 1 and 3, but multiple subsequent rearrangements in type 1 confound reconstruction of this gene order change. This transcriptional unit has been found to have been broken independently in *Geranium palmatum*, also between *atpH* and *atpI* (Guisinger et al. 2011).

Like the other genome types from Clade I, both type 3 genomes appear to lack a functional copy of the *trnG-gcc* gene. *Erodium jahandiezianum* contains a pseudogene of *trnG-gcc* in the conserved location for that gene that retains only 58.1% sequence identity with the functional gene, which is lower than the similarity between *ψtrnG-gcc* and another conserved tRNA, *trnD-guc* (65.3% identity). A second, slightly less conserved pseudogene (57.7% identity) is found next to *rpl32*. Both pseudogenes are located at rearrangement endpoints. In *E. crassifolium*, *ψtrnG-gcc* is slightly more conserved with 74.6% sequence identity to the *trnG-gcc* gene from *E. carvifolium*.

The variability of the spacer region containing *ψtrnG-gcc* in types 1, 3 and, 4 suggests that it may be a mutational hotspot in Clade I. The length of the intergenic region between *ψtrnG-gcc* and *psbZ* is 1 kb in *E. jahandiezianum*, over 3kb in *E. texanum*, and just 217 bp in *E. gruinum*. And although all Clade II *Erodium* taxa contain an apparently functional *trnG-gcc* gene, we show below that there is evidence of the region containing *trnG-gcc* acting as a hotspot for illegitimate recombination in type 2 plastid genomes as well.

Plastid genome type 4: *Erodium gruinum*

The plastid genome of *E. gruinum* is highly unusual (Figure 7). First, it contains a large 25,508 bp inverted repeat that includes the ribosomal operon and other genes commonly found in the IR in most angiosperms. As previously indicated, there is strong evidence that the IR was lost once on the branch leading to *Erodium*. As such, the IR in

E. gruinum must have a relatively recent origin. In addition to this bizarre development in genome architecture, *E. gruinum* has lost all 11 plastid-encoded *ndh* genes, a ribosomal protein gene (*rpl23*), and three introns (*clpP*, *rpoC1*, and the cis intron of *3' rps12*). A tRNA gene, *trnV-gac*, also appears to be a pseudogene in *E. gruinum*, as the closest match to this gene from other *Erodium* species retains only 59.7% sequence identity. All these gene and intron losses are shared by other members of this clade, which has been designated the long-branch clade (LBC) due to its long branch in phylogenetic reconstructions

The IR of *E. gruinum* contains some of the genes found in a typical angiosperm IR, but even these genes have undergone several rearrangements. Figure 10 shows an alignment of the IRs of *E. gruinum*, *C. macrophylla*, and unrearranged Geraniales outgroup *Francoa sonchifolia* (25,508 bp, 22,304 bp, and 26,509 bp, respectively). Starting at the LSC boundary, the *atpB/E-trnM-cau* transcriptional unit is interrupted by the IR boundary such that the first 340 bp of *atpB* are situated in the LSC. After *atpE* there is a tandem duplication of *trnM-cau*. The copy adjacent to *atpE* is a pseudogene with 64.4% identity to *E. texanum* and *E. carvifolium*, in which the gene is identical. The second copy of *trnM-cau* appears to be functional, with 95.9% identity to *E. texanum* and *E. carvifolium*. The first collinear block of genes among the three species contains *ycf2*, *trnL-caa*, and *ndhB*. Although this region is collinear, it is highly variable among the three species, as *ycf2* is a greatly reduced pseudogene (2 kb) in *C. macrophylla* absent in *E. gruinum* but intact (7 kb) in *Francoa*. For this first block, *E. gruinum* contains *trnL-caa* and pseudogenes representing the two exons of *ndhB*. After this collinear block, *E. gruinum* contains two genes normally found in the SSC, *ccsA* and *trnL-uag*. Next, the IR of *E. gruinum* shares a collinear block with *California* and *Francoa* that contains the entire ribosomal operon and several tRNA genes, but this block is in an inverted

orientation relative to the other IRs. After the *rrn16* gene, the conserved gene order would be *trnV-gac*, *rps12* and *rps7*, but instead *E. gruinum* contains two non-identical tandemly repeated copies of *rpl32* (described in more detail below) followed by *rps7*, *rps12*, which are inverted relative to the conserved IR gene order, and a pseudogene of *trnV-gac* adjacent to a pseudogene of *ndhA* exon 1. Immediately following the *ndhA* pseudogene is *rps15*, another gene normally found in the IR. After *rps15* there is 2.2 kb of intergenic spacer until the SSC boundary. In summary, the IR of *E. gruinum* contains many genes commonly found in the conserved IR of other angiosperms, albeit in an unusual order that is interrupted by genes normally found in the SSC. The only large block of IR genes that is conserved is the ribosomal operon spanning from *trnN-guu* to *rrn16*, and this block has been inverted relative to other angiosperm IRs.

The IR of *E. gruinum* contains two non-identical tandemly repeated copies of *rpl32*. It is not clear if these genes are functional, as they are very similar to each other at the conserved 5' end (86.9% identical) but less similar over the entire coding sequence (64.1% identical). The conserved 5' ends of the two genes are only 64.5% and 66.4% identical to *rpl32* from *E. texanum*. When the two full *rpl32* genes *from E. gruinum* are aligned with the *rpl32* gene from *E. texanum*, the sequence identity drops to 48.4% and 40.4% identity, respectively. These two paralogs are also present in the other three LBC draft genomes (data not shown). We have annotated them both as functional genes in *E. gruinum* because there is no evidence that either is nonfunctional.

Finally, GC content is unusually high in *E. gruinum* at 43%. Since the *ndh* genes are the most GC-poor plastid genes, we wanted to investigate whether the loss of *ndh* genes might be responsible for the elevation in GC content in *E. gruinum*. Thus we examined the GC content of the 19 protein-coding genes used in the evolutionary rates analysis and found that the elevation in GC content was also evident in the 19 genes from

E. gruinum (44.1% GC) as well as the other three LBC species included for the 19 protein coding genes (Table 3). Thus the loss of *ndh* genes does not appear to be the cause of the elevated GC content in *E. gruinum* or other LBC taxa. An elevation in GC content evident across representatives of all functional categories of protein-coding genes suggests some genome-wide mutational bias favoring GC was in effect on the long branch leading to this clade.

Survey of type 2 plastid genomes

Summary of Clade II/type 2 plastid genomes:

Type 2 plastid genomes are more conserved in gene order and gene content than genomes from Clade I. Differences in gene order among type 2 plastid genomes are described in detail below. All type 2 plastid genomes contain the same gene complement and the same number of introns, except for the remaining *clpP* intron (exon 2) which has been lost at least twice in Clade II. *Erodium cossonii* and *E. reichardii* both lack this intron, and it is unclear whether this represents one or two independent losses. *Erodium moschatum* also lacks the intron, but this loss is clearly independent of the other two (Figure 2). The *clpP* intron is also lost independently in the LBC, represented by the Type 4 plastid genome of *E. gruinum*. Finally, it is unclear whether all type 2 genomes encode *trnK-uuu*, since no first exon could be found in *E. moschatum*; however, this gene is divergent in all four genome types, and there is no clear pattern to the loss/retention of this gene.

Type 2 plastid genomes range in size from 115,794 bp in *E. foetidum ssp. foetidum* to 123,865 bp in *E. trifolium* (Table 4). This 8 kb range in genome sizes can be explained by two factors, the percentage of repetitive DNA and the size of the remaining pseudogenes for the large ycf genes, *ycf1* and *ycf2* (Table 5). The largest genome, *E.*

trifolium, has a relatively high repetitive DNA content (6.08%) as well as a relatively large *ycf2* pseudogene (5,962 bp), whereas the smallest genome, *E. foetidum ssp. foetidum* has a low repeat content (1.48%) and a small *ycf2* pseudogene (204 bp). Two other genomes of intermediate size happen to have exactly the same length at 116,810bp: *E. manescavi* has a large *ycf2* pseudogene and low repeat content (3,472 bp and 0.9%, respectively), whereas *E. rupestre* has a small *ycf2* pseudogene but a higher repeat content (1,027 bp and 2.55%, respectively). Although the functional *ycf1* is larger than *ycf2* (7,659 bp and 6,333 bp in *Pelargonium x hortorum*, respectively), in *Erodium* the *ycf1* pseudogene has eroded far more rapidly than that of ycf2. The largest fragment of *ycf1* in a type 2 plastid genome is less than 1 kb (917 bp in *E. carvifolium*), and it is completely absent from *E. manescavi*. By contrast, the *ycf2* pseudogene in *E. trifolium* and *E. cossonii* are nearly the full length of the intact gene (5,962bp and 5,892bp, respectively). There is no obvious cause for the more rapid degradation of *ycf1* in *Erodium*, as both genes are inferred to have lost functionality on the branch leading to *Monsonia*, *Geranium*, and *Erodium* (Guisinger et al., 2011).

Gene Order

The nine additional plastid genomes from Clade II are similar overall to that of *E. carvifolium* (Blazier et al., 2011) with a few notable exceptions. First, there are four genomes that are not collinear with *E. carvifolium*. Two species, *E. reichardii* and *E. cossonii*, share a similar 24 kb inversion that appears to have occurred independently. The endpoints of this inversion are unusual in *E. carvifolium*: the first endpoint has two tandemly repeated copies of *trnQ-uug*, and the inversion in *E. reichardii* and *E. cossonii* is located between these two tRNAs, leaving one copy of *trnQ-uug* on each side of the inversion. The second endpoint has two copies of *trnG-gcc* in an inverted orientation

surrounding a pseudogene of *rps18* in *E. carvifolium*, but this duplication and pseudogene are absent from the species for which this region is a rearrangement endpoint. In fact, the only species that resembles *E. carvifolium* in this region is *E. manescavii*, which has the *rps18* pseudogene but only one copy of *trnG-gcc*. This region between *trnE-uuc* and *psbZ* appears to be a hotspot for illegitimate recombination in Clade II, not just in Clade I, from which *trnG-gcc* has been lost.

The third species that differs in gene order from *E. carvifolium* is *E. trifolium*. The genomes are collinear except that two genes have moved from their conserved location (*rpl20* and the 5' exon of *rps12*). This gene order can be explained by five inversions. Since gene duplications have occurred in many *Erodium* plastid genomes, it is also possible that these genes were duplicated in their new locations and the original copies lost.

The fourth species that differs in gene order from *E. carvifolium* is *E. cygnorum*. This species is the earliest diverging representative of Clade II and also contains the only Clade II plastid genome in which a conserved transcriptional unit has been broken by genomic rearrangement. The *rpl23* transcriptional unit has been broken between *rpl16* and *rpl14*; this transcriptional unit is also broken in type I genomes (*E. texanum* and *E. guttatum*), but in that type, the breakpoint is on the opposite side of *rpl16*, between *rpl16* and *rps3*. Guisinger (2011) found that this transcriptional unit was broken in *Geranium palmatum* and *Monsonia speciosa* as well. In *G. palmatum* the transcriptional unit is broken into three pieces, and in *M. speciosa* it is broken once in the same location as in *Erodium* Type I genomes. We can infer that each of these disruptions of the *rpl23* transcriptional unit has occurred independently. In *E. cygnorum* the rearrangement endpoints contain partial copies of *rpl23*, but it is not possible to

determine whether these repeats played a role in the inversions or resulted from some other process.

Evolutionary Rates

Aside from the many gene and intron losses and re-appearance of the IR in *E. gruinum*, the branch leading to these species lacking *ndh* genes is distinctly long in phylogenetic reconstructions. We conducted an analysis of evolutionary rates in *Erodium* to confirm that the branch leading to the LBC is in fact significantly longer than other branches. We further sought to answer the following questions: First, is the acceleration limited to *dN*, *dS*, or present in both? Second, is the rate acceleration limited to a specific functional class of genes? Third, are the internal branches in the highly rearranged Clade I significantly longer than those in the less rearranged Clade II? The data set contained 19 genes, with two representatives of each functional class, all four *rpo* genes, and *matK*, *clpP*, and *rbcL*. The branch leading to the LBC was significantly longer for *dS* for all genes and for *dN* for 14 of the 19 genes (Table 6). For the five genes for which the branch length difference was not significant, two of them had been significant before correction for multiple comparisons. The other three genes, *petB*, *psbC*, and *rbcL*, have a very low rate of nonsynonymous substitution such that many branches in the analysis had *dN* values of zero or close to zero, which likely confounded the likelihood ratio test. The concatenated dataset with 19 genes showed that the branch leading to the long-branch clade was significant for both *dS* and *dN*. The long branch leading to the LBC thus appears to be accelerated with respect to *dS*, and to *dN* for all but the slowest evolving genes.

We also compared *dN* and *dS* between the internal branches of Clade I and Clade II to investigate whether the highly rearranged clade (I) showed significantly higher rates

of evolution than the clade showing few rearrangements (II). The Wilcoxon rank sum test showed that the difference in *dS* between Clade I and Clade II was significant in 8 out of 19 genes tested (Figure 6a) whereas the difference in *dN* was significant in 6 out of 19 genes tested (Figure 6b). Among these genes, *psaA*, *rbcL*, *rpoB*, and *rpoC1* showed significant difference between these two clades in both *dS* and *dN*.

DISCUSSION

Gene and intron loss in *Erodium*

The pattern of gene and intron loss in *Erodium* parallels that of genomic rearrangement, with Clade I showing much greater variation than Clade II. *Erodium* genomes lack many canonical tRNA genes, protein-coding genes, and introns. The tRNA *trnT-ggu* is inferred to have been lost on the branch leading to Geraniaceae, and its former location is the endpoint of an inversion shared by all members of the family (Guisinger et al. 2011; Weng et al. in press). The *trnK-uuu* gene, whose intron contains the maturase gene *matK*, shows a complex pattern of divergence in *Erodium*. The two exons of *trnK* appear to be intact in most Clade II taxa except for *E. trifolium*, in which the first exon contains a 10 bp deletion, in *E. manescavi*, in which the second exon is either divergent or lacks the 4 bp at its 5' end, and *E. moschatum*, in which the first exon could not be located. The status of *trnK* is even less clear in Clade I. In the two type 1 genomes, *E. texanum* and *E. guttatum*, the 3' end of the first exon is divergent but almost identical to the same region in the type 3 genomes, *E. jahandiezianum* and *E. crassifolium*. The second exon is divergent and contains a 6 bp insertion in type 1 genomes but is highly conserved in type 3 genomes. The two exons in *E. gruinum* are similar to the type 3 genomes. Because *trnK-uuu* is the only tRNA for lysine in the plastid genome, we are reluctant to annotate it as a pseudogene, whereas the other

missing tRNAs all have a "backup" tRNA that could conceivably take over the function of the lost gene through "wobble" base pairing (Sugiura et al. 1998). *trnV-uac* could potentially take over the function of *trnV-gac*, which is lost from *E. gruinum* as well as from *Monsonia speciosa* and *Geranium palmatum* (Guisinger et al. 2011). Similarly, *trnG-ucc* could potentially compensate for the loss of *trnG-gcc* in Clade I. This gene loss has been shown to have a paradoxical effect on codon usage: the use of the CCG codon for glycine specifically recognized by *trnG-gcc* is used ~60% more frequently in Geraniaceae than in other rosids, despite its loss in several Geraniaceae lineages (Guisinger et al. 2011).

Loss and re-gain of the IR

The presence of additional sequence between *trnL-uag* and *trnH-gug* in type 1 and type 3 plastid genomes changes our model for IR loss in *Erodium* somewhat. Previously we hypothesized that the loss of the IR was a "clean" deletion (Blazier et al., 2011), but it now appears that the deletion of the IR left behind fragments (*ndhA* and *trnI-cau*) that are still present in Clade I and that the miniscule intergenic spacer (233 bp in *E. carvifolium*) seen in Clade II (type 2) species is a derived character that has resulted from the erosion of remnants of the deleted IR. Type 1 plastid genomes contain the largest traces of the IR, and type 3 shows erosion of this region and thus represents an intermediate state between this region of types 1 and 2 plastid genomes: the *ndhA* pseudogene is 112 bp long in *E. texanum* and *E. guttatum* but has been reduced to just 33 bp in *E. jahandiezianum* and *E. crassifolium*. The type 4 plastid genome, *E. gruinum*, has undergone subsequent rearrangements in this region and provides no information about loss of the ancestral IR. Although our model for loss of the IR has changed, the additional

genomes nonetheless support a single loss of the IR in *Erodium* and a single re-acquisition of an IR in *E. gruinum*.

Evidence for regrowth of the IR in *E. gruinum* has implications for modeling rearrangements in other plastid genomes. One of the models suggested to derive the highly rearranged plastid gene order of *Trachelium caeruleum* (Campanulaceae) involved loss and re-growth of the IR along with ten inversions (Cosner et al. 1997; Haberle et al. 2008). The other two models involve a mixture of inversions, transpositions, and expansions and contractions of the IR. In light of new evidence that re-growth of the IR is possible, the model invoking only inversions along with loss and regain of the IR appears more likely than before. The genome of *Geranium palmatum* (Guisinger et al. 2011), like that of *T. caeruleum*, has a SSC with many genes usually found in the LSC. Since inversions have never been shown to cross IR boundaries, it is difficult to explain how 33 kb of LSC genes found their way into the SSC in *G. palmatum*. A model that invoked transient loss of the IR would greatly simplify the reconstruction of gene order changes; moreover, genes normally found at the boundaries of the SSC and LSC are adjacent in *G. palmatum* (*ccsA-trnL-uag* and *trnH-gug-psbA*) just as in *Erodium*, suggesting that an IR may have been lost from this region as well.

Rates of nucleotide substitution in *Erodium* versus *Pelargonium*

Although rates of nucleotide substitution have been shown to be elevated in Geraniaceae compared to other rosids (Guisinger et al., 2008), the most striking long branches in phylogenetic reconstructions are internal branches in *Pelargonium* (leading to the major clades and to clade C2 in particular) (Weng et al. 2012) and the long branch leading to the LBC in *Erodium*. Our dense taxon sampling in *Erodium* allows us to draw distinctions between rate accelerations in the two genera. First, Weng et al. (2012) found

that in *Pelargonium* just two branches show acceleration in *dS*, the branch leading to Geraniaceae and the branch leading to the genus. The *dS* acceleration was not locus specific and was highly correlated among the four genes sampled (*rpoC1*, *matK*, *ndhF*, and *rbcL*). However, *dN* showed a more complex locus-specific pattern in which the same two branches were accelerated for *matK* and *ndhF*, but no significantly accelerated branches were detected for *rbcL*. For *rpoC1* not only were the branches to the family and genus accelerated but also the branches leading to the major clades and many internal branches in clade C2. These results are consistent with the results for *rpo* genes in Chapter 3, which found elevated *dN*/*dS* for the branches leading to the major clades for the three large *rpo* subunits (*rpoB*, *rpoC2*, and *rpoC2*) but not for *matK*, *ndhF*, or *rbcL*. In *Erodium* we find that the branch leading to the LBC is accelerated for *dS* for all genes and for *dN* for all but the slowest evolving genes in the 19 gene data set.

It has been shown that GC content is elevated in Geraniaceae plastid genomes (Guisinger et al., 2011), but *E. gruinum* is elevated even further. In fact, *E. gruinum* has the highest GC content of any angiosperm plastid genome on GenBank as of 11/2013. This acceleration in both *dN* and *dS*, together with the increase in GC content seen in the four LCB species represented in the 19 gene data set, suggest that an aberration in DNA repair is responsible for this long branch.

CONCLUSIONS

A dense sampling of *Erodium* plastid genomes demonstrates that loss of the IR does not necessarily lead to further destabilization of plastid gene order. In fact, we find two distinct trajectories of plastid genome evolution in *Erodium*. Clade I genomes contain more repetitive DNA and display a greater number of genomic rearrangements. Clade II genomes have a relatively low repeat content, though it is elevated in the two

genomes that display multiple unique gene order changes (*E. trifolium*, 6.08%; *E. cygnorum*, 8.43%) compared to the average repeat content of genomes containing just one or no unique gene order changes (2.1%). Thus both clades of *Erodium* appear to follow the trend that an elevation in repeat content coincides with an increase in genomic rearrangements (Jansen et al. 2007; Haberle et al. 2008; Guisinger et al. 2011; Weng et al. 2012). Unrearranged plastid genomes such as that of Geraniales outgroup *Francoa sonchifolia* have a very low repeat content (0.45%), half that of *E. manescavi*, the *Erodium* with the lowest repeat content (0.98%).

The most interesting repetitive DNA in *Erodium* is the newly formed 25 kb IR found in *E. gruinum*. The re-appearance of a large IR in a lineage from which the ancestral IR has been lost has never been demonstrated before. *Erodium gruinum* has also lost 12 of the 75 protein-coding genes present in other Clade I taxa, as well as three of the 14 introns. In addition, it shows an elevated GC content (43%), making it the most GC-rich angiosperm plastid genome yet discovered. Finally, *E. gruinum* and other LBC species are separated from the rest of Clade I by a branch that is significantly longer for *dS* for all 19 genes examined and for *dN* for 14 of the 19 genes. *Erodium gruinum* and the LBC as a whole are highly unusual, and it is unclear whether their abnormal plastid genomes represent the most severe outcome of forces shaping plastid genome evolution in other Clade I species or whether the forces underlying the rate acceleration, gene and intron loss, increase in GC content, and the re-appearance of the IR are entirely different.

Of the three hypotheses raised to explain the persistence of the IR, one can be clearly eliminated: the IR does not exist to protect the ribosomal operon since we find two angiosperm lineages lacking an IR and another in which the IR does not contain the entire ribosomal operon (*Monsonia*). The second hypothesis, that the IRs function in plastid DNA replication, seems promising since the IR does contain the primary origins

of replication, and replication is also initiated in the inverted repeat regions in HSV-1 and in the yeast 2-micron circular plasmid system (Broach 1991); however, the chloroplast is highly polyploid, so it is unclear whether two copies of the IR would ever need to undergo intramolecular recombination to initiate DNA replication. Moreover, plastid DNA replication does not appear to be impaired in *Erodium* or in legumes lacking the IR, so the IR does not appear to be critical for plastid DNA replication. It is tempting also to eliminate the hypothesis that the IR persists in order to stabilize the plastid genome, since the most highly rearranged plastid genomes have an IR (e.g. *Pelargonium*, *Geranium*, *Trachelium* (Campanulaceae), *Menodora* (Oleaceae). However, consistent with other studies, we do notice a correlation between genomic rearrangement and repeat content in *Erodium* and in the other Geraniaceae genera (Jansen et al. 2007; Guisinger et al. 2011; Weng et al. in press). Thus, while there may be insufficient evidence that the large inverted repeat stabilizes the plastid genome, there is nonetheless ample evidence that repeats, broadly speaking, act to destabilize it. Perhaps then the IR does act to stabilize the plastid genome, but only in genomes with a low repeat content. In rearranged plastid genomes we see evidence of illegitimate recombination, which is thought to underlie gene order changes and movement of the IR boundaries and which we have recently hypothesized to underlie the divergence of less constrained plastid genes such as *rpoA* (Chapter 3). When the repeat content of a plastid genome is very low, illegitimate recombination should be minimized. Under these conditions, recombination could be largely confined to the IRs and genome stability would be maintained. However, once the repeat content of a plastid genome increases, illegitimate recombination within single copy regions or between regions flanking the IRs become more likely, causing inversions and expansion or contraction of the IR, respectively. Thus the presence of the IR may be less important to plastid genome stability than the absence of other large repeats.

Finally, now that loss and regain of the IR has been demonstrated, it may be useful to revisit models to explain gene order changes in highly rearranged plastid genomes. The gene orders found in *Trachelium caeruleum* and *Geranium palmatum*, in which many LSC genes have migrated to the SSC, may be more parsimoniously explained by loss and regain of the IR than through a complex mixture of inversions, gene relocations, and expansions and contractions of the IR. The gene order in *G. palmatum* in particular is suggestive of the loss of one IR, between the same genes where the IR was lost in *Erodium*. As the recombination and DNA repair machinery targeted to the plastids becomes better understood, we anticipate that it will become possible to associate characteristics of rearranged plastid genomes with specific defects in the DNA repair machinery.

Table 4.1. Taxon sampling for genome sequencing and evolutionary rates analysis.

Species	Clade	Accession number	Publication
Whole genomes			
C. macrophylla	NA	TBD	Weng et al., 2013
E. carvifolium	Clade I	NC_015083	Blazier et al., 2011
E. cossonii	Clade II	JN989541	this study
E. crassifolium	Clade I	TBD	this study
E. cygnorum	Clade II	JN815232	this study
E. foetidum ssp. cheilantifolium	Clade II	JQ014209	this study
E. foetidum ssp. foetidum	Clade II	KF771022	this study
E. gruinum	Clade I--LBC	KF804069	this study
E. guttatum	Clade I	JN688871	this study
E. jahandiezianum	Clade I	JN997456	this study
E. manescavi	Clade II	KF751825	this study
E. moschatum	Clade II	JQ063026	this study
E. reichardii	Clade II	KF771021	this study
E. rupestre	Clade II	KF751824	this study
E. texanum	Clade I	NC_014569	Guisinger et al., 2011
E. trifolium	Clade II	KF441758	this study
Genes only			
E. absinthoides	Clade I--LBC		this study
E. chrysanthum	Clade I--LBC		Guisinger et al., 2008
E. guicciardii	Clade I--LBC		this study

Table 4.2. General characteristics of the four types of *Erodium* plastid genomes and outgroup *California macrophylla*.

Genome characteristics		**Type 1**	**Type 2**	**Type 3**	**Type 4**
Species	*C. macrophylla*	*E. texanum*	*E. carvifolium*	*E. jahandiezianum*	*E. gruinum*
Size/with one IR	149,202 bp/126,898 bp	130,812 bp	116,935 bp	121,692 bp	142,208 bp/116,700 bp
Number of protein-coding genes	75	75	75	75	63
Number of tRNA genes	29	28	29	28	27
Number introns		14	17	14	11
%GC content	38.7%	36.8%	39.0%	39.1%	43.0%
%Repetitive DNA	1.50%	23.10%	2.85%	2.62%	4.56%
LSC	88,738 bp	-	-	-	56,293 bp
IR	22,304 bp	-	-	-	25,508 bp
SSC	15,856 bp	-	-	-	34,899 bp

Table 4.3 lists general characteristics of *Erodium* plastid genomes by clade. The number of estimated gene order changes per species is also given. *E. cossonii* and *E. reichardii* have an identical inversion (marked with an asterix), but it is unclear whether this is due to a single event or to homoplasy.

Clade I	**Type 1**		**Type 3**		**Type 3 (LBC)**	**LBC taxa included in 19 gene data set**		
Genome characteristics	*E. texanum*	*E. guttatum*	*E. jahandiezianum*	*E. crassifolium*	*E. gruinum*	*E. absinthoides*	*E. chrysanthum*	*E. guicciardii*
Size (bp)	130,812 bp	128,510 bp	121,692 bp	121,393 bp	142,208 bp	-	-	-
Number of protein-coding genes	75	75	75	75	63	63	63	63
Number of tRNA genes	28	28	28	28	28	28	28	28
Number introns	14	14	14	14	11	11	11	11
%GC content	39.5%	39.4%	39.1%	39.1%	43.0%	-	-	-
%GC content in 19 gene data set	41.9%	41.8%	41.7%	41.6%	44.1%	44.0%	44.1%	44.0%
%Repetitive DNA	23.09%	17.61%	2.61%	3.24%	4.56%	-	-	-
Estimated gene order changes	14	13	5	5	10			

Clade II (Type 2)										
Genome characteristics	*E. carvifolium*	*E. cossonii*	*E. cygnorum*	*E. foetidum ssp. cheil.*	*E. foetidum ssp. foet.*	*E. manescavi*	*E. moschatum*	*E. reichardii*	*E. rupestre*	*E. trifolium*
Size (bp)	116,935 bp	121,465 bp	121,905 bp	116,340 bp	115,794 bp	116,810 bp	119,078 bp	117,753 bp	116,810 bp	123,865 bp
Number of protein-coding genes	75	75	75	75	75	75	75	75	75	75
Number of tRNA genes	28	28	28	28	28	28	28	28	28	28
Number introns	17	16	17	17	17	17	16	16	17	17
%GC content	39.0%	39.2%	39.2%	38.9%	38.9%	39.1%	39.1%	39.2%	38.9%	39.3%
%GC content in 19 gene data set	41.9%	41.8%	41.8%	41.8%	41.8%	41.9%	41.8%	41.8%	41.8%	41.9%
%Repetitive DNA	2.85%	3.54%	8.43%	1.53%	1.48%	0.98%	2.20%	1.68%	2.55%	6.08%
Estimated gene order changes	-	1*	2	-	-	-	-	1*	-	2

Table 4.4. The size of *ycf1* and *ycf2* pseudogenes for each species in Clade II as well as the percentage of repetitive DNA. These two factors account for the range of genome sizes within the clade.

Clade II species	***ψycf1***	***ψycf2***	**% Repetitive DNA**	**Genome Size**
E. trifolium	217 bp	5,962 bp	6.08%	123,865 bp
E. cygnorum	769 bp	2,739 bp	8.43%	121,905bp
E. cossonii	823 bp	5,892 bp	3.54%	121,465 bp
E. moschatum	867 bp	3.574 bp	2.20%	119,078 bp
E. reichardii	229 bp	3,771 bp	1.68%	117,753 bp
E. carvifolium	917 bp	1,568 bp	2.85%	116,935 bp
E. rupestre	605 bp	1,027 bp	2.55%	116,810 bp
E. manescavi	0	3,472 bp	0.98%	116,810 bp
E. foetidum ssp. cheilantifolium	836 bp	1,016 bp	1.53%	116,340 bp
E. foetidum ssp. foetidum	245 bp	204 bp	1.48%	115,794 bp

Table 4.5. The results of LRTs for *dS* and *dN* for the 19 gene data set and for a concatenated alignment of the 19 genes. Significant values are given in bold. All genes and the concatenated alignment are significant for *dS*, and 14 of the 19 genes, plus the concatenated alignment, are significant for *dN*.

Gene	Null mode for dS[1]	Alternative model for dS[2]	*p*-value[3]	Null mode for dN[1]	Alternative model for dN[2]	*p*-value[3]
atpA	-3627.58	-3557.56	**9.05E-31**	-3542.98	-3531.42	**5.20E-05**
atpB	-3307.86	-3264.62	**4.80E-19**	-3247.7	-3229.83	**8.81E-08**
clpP	-3312.36	-3303.08	**5.60E-04**	-3458.15	-3402.6	**2.01E-24**
matK	-4555.24	-4490.75	**2.33E-28**	-4635.78	-4547.3	**7.60E-39**
petA	-2284.82	-2257.43	**4.58E-12**	-2253.48	-2248.54	0.0568
petB	-1361.93	-1355.05	**0.0071**	-1323.39	-1323.05	13.9259
psaA	-4652	-4579.41	**6.67E-32**	-4532.4	-4518.84	**6.50E-06**
psaC	-466.241	-455.08	**7.84E-05**	-425.208	-425.208	34
psbA	-2060.86	-2040.63	**6.82E-09**	-1981.01	-1975.01	**0.0181**
psbC	-3035.13	-2978.44	**6.06E-25**	-2930.49	-2927.31	0.3969
rbcL	-3326.93	-3270.41	**7.19E-25**	-3247.74	-3247.51	16.919216
rpl2	-2798.02	-2751.28	**1.40E-20**	-2834.79	-2795.17	**1.87E-17**
rpl14	-897.761	-891.732	**0.0175**	-885.963	-881.696	0.1185
rpoA	-2681.99	-2644.65	**1.88E-16**	-2676.84	-2640.93	**8.02E-16**
rpoB	-9782.58	-9661.4	**4.09E-53**	-9926.3	-9719.44	**1.93E-90**
rpoC1	-5953.33	-5866.07	**2.59E-38**	-6096.91	-5983.51	**1.01E-49**
rpoC2	-12431.2	-12312.6	**5.45E-52**	-12723	-12424.6	**2.83E-130**
rps2	-2307.67	-2283.08	**7.94E-11**	-2452.77	-2356.62	**3.40E-42**
rps7	-1587.56	-1486.6	**2.71E-44**	-1676.57	-1521.81	**9.45E-68**
19 genes	-72374.8	-71415.1	**0**	-72767.3	-71711.6	**0**

1. The null model constrained one evolutionary rate shared by all branches.

2. The alternative model frees the branch leading to the long-branch clade from the constraint.

3. p-values were Bonferoni corrected. Significant p-values with 5% cut-off are in bold.

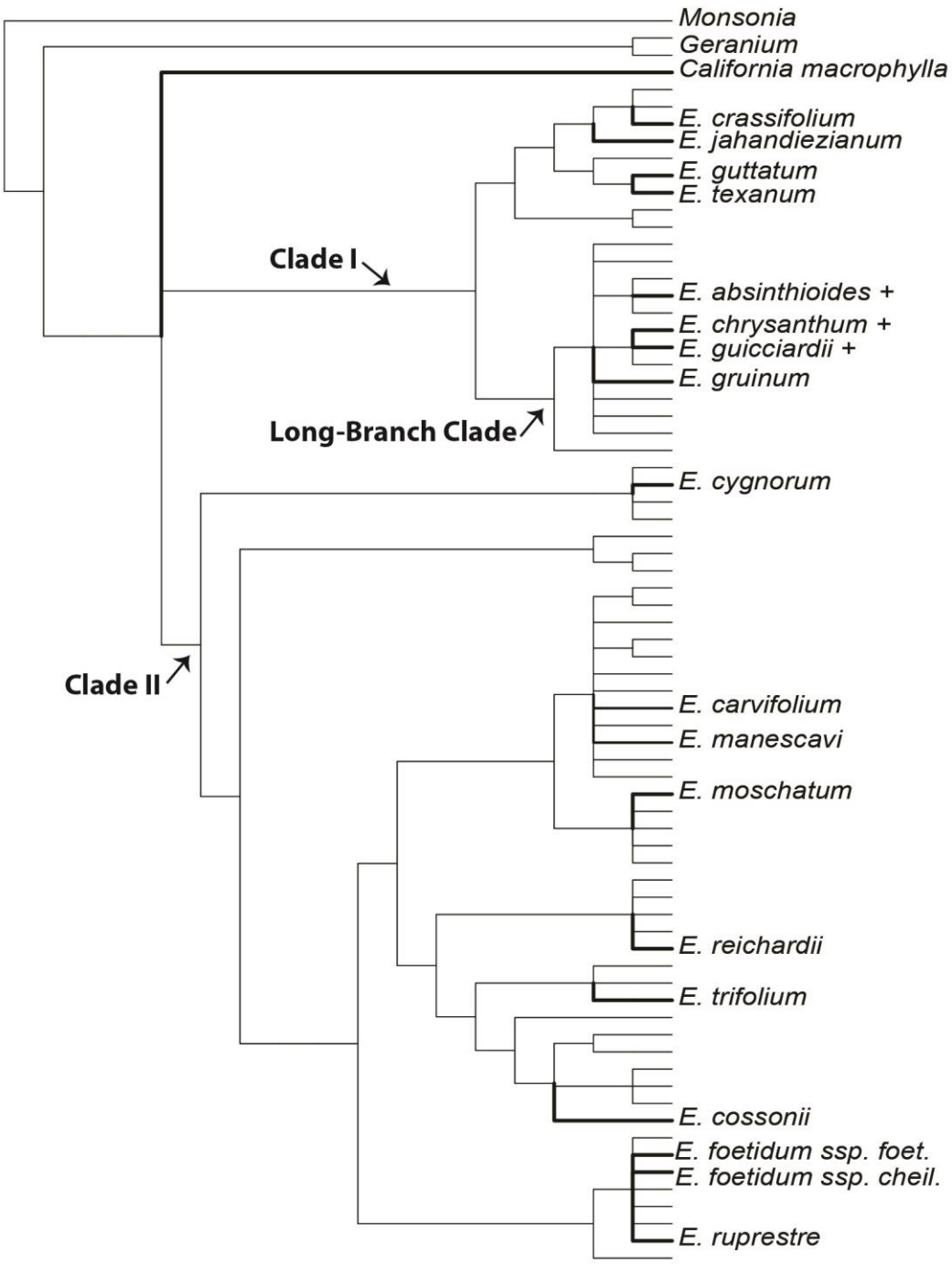

Figure 4.1. Figure 1. A ML tree for *Erodium* based on the *trnL-F* spacer for all 72 species, adapted from Fiz et al. (2006) and Blazier et al. (2011). The 19 species labelled are included in the 19 gene data set for evolutionary rates analysis. Genes from the three species marked with a plus sign are included in the rates analysis but the genomes have not been completed.

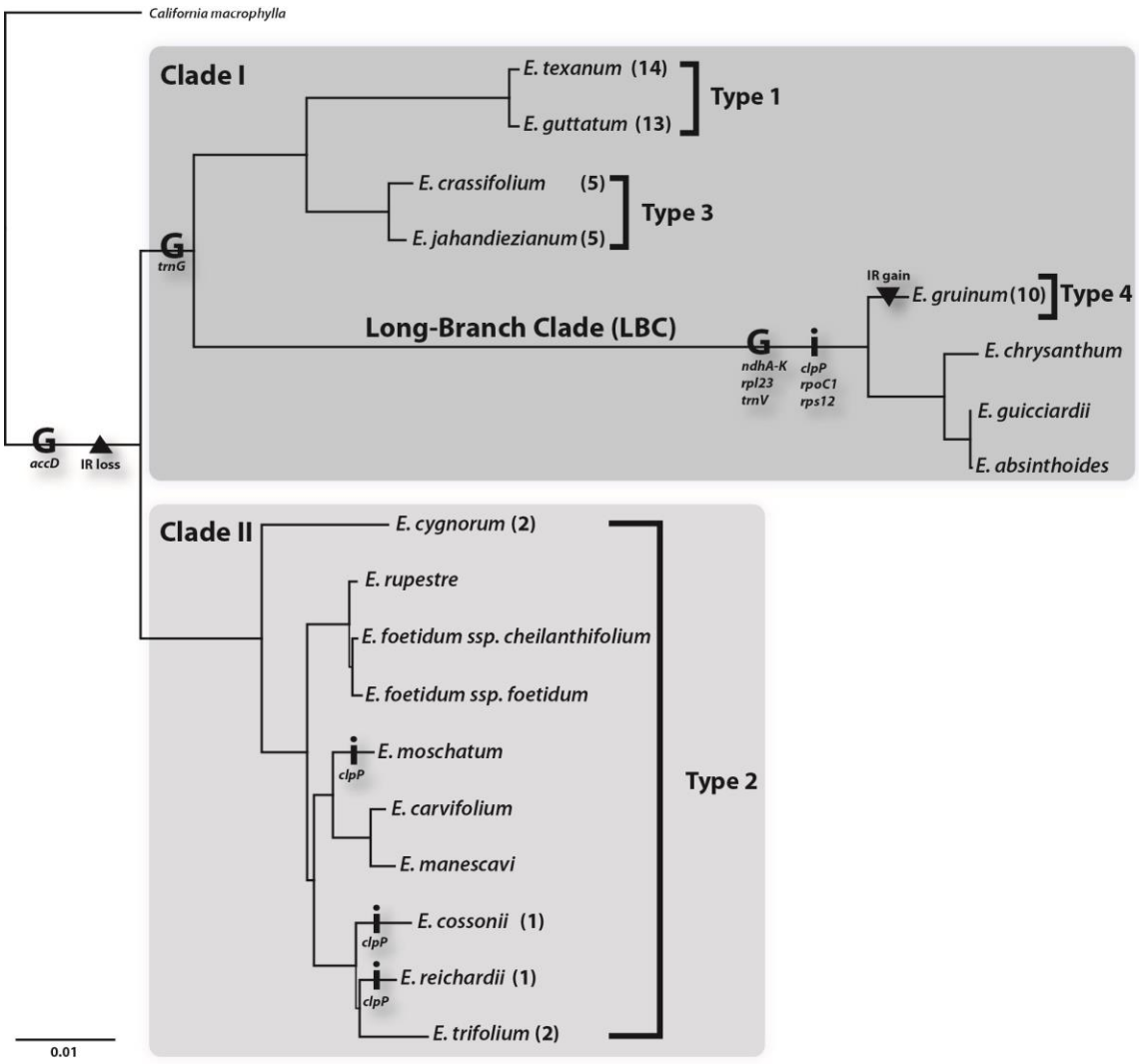

Figure 4.2. ML tree of the 19 species in the rates data set generated from a concatenated alignment of all 19 genes (26,985 bp). The likelihood score of the tree is -70914.4940 lnL. Plastid genome rearrangements have been mapped on to the tree. Gene and intron losses are listed below the letters "G" and "i", respectively. IR loss and gain are indicated by a triangle and an inverted triangle, respectively. The estimated number of unique gene order changes is given for each species in parentheses after the species name. The plastid genome type (Types 1-4) are also indicated to the left of the species names for each clade.

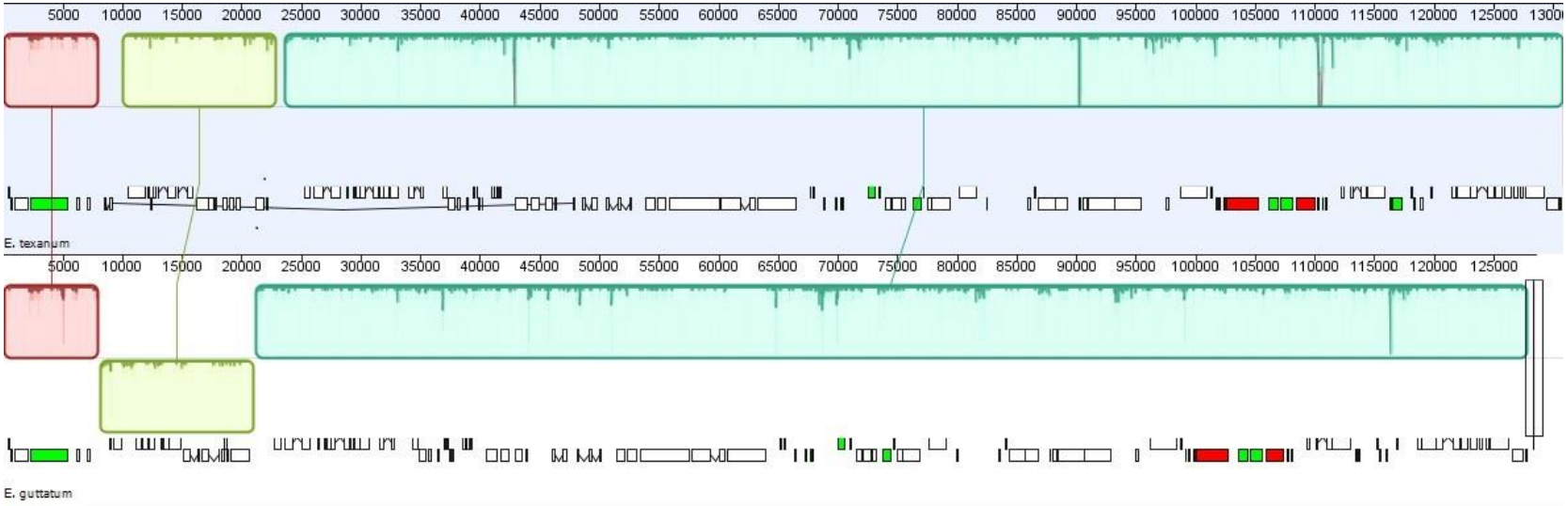

Figure 4.3. MAUVE alignment of type 1 genomes *E. texanum* and *E. guttatum* showing that one inversion (yellow block) distinguishes their gene orders.

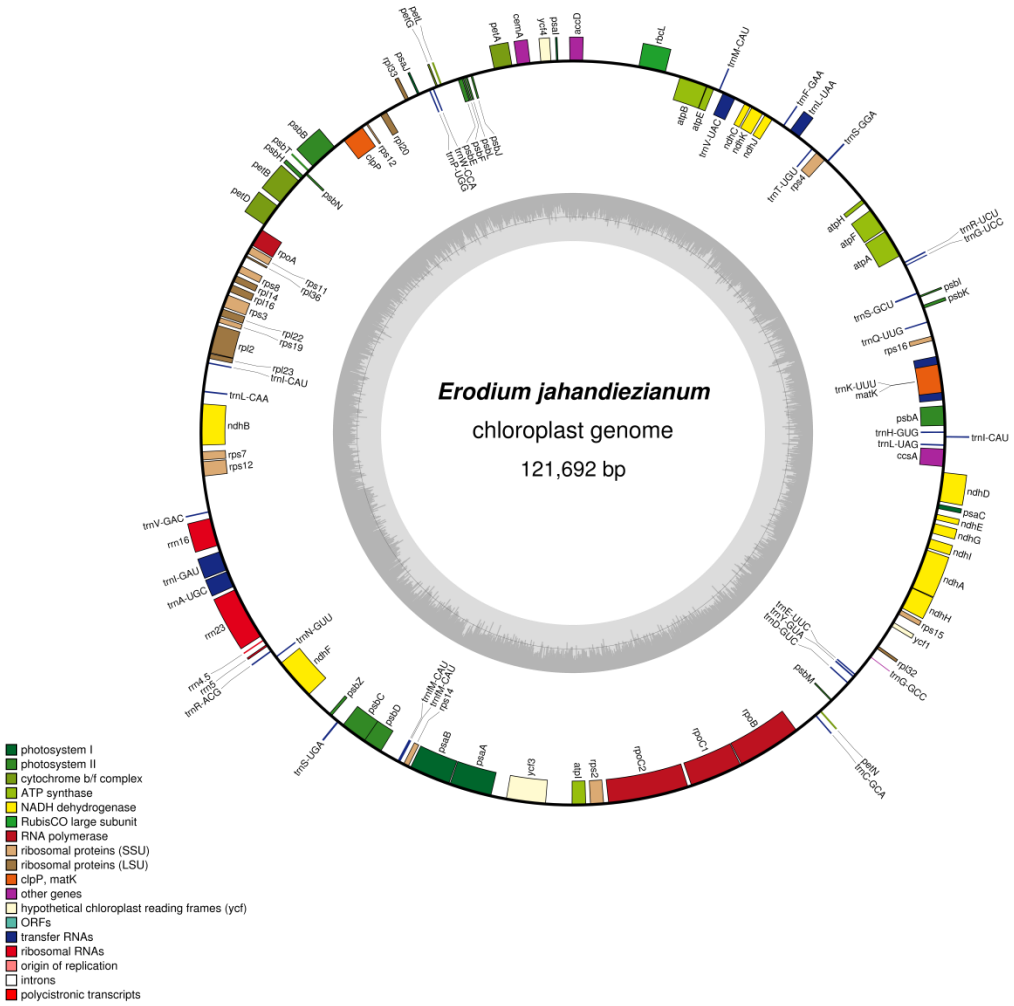

Figure 4.4. The genome map of *E. jahandiezianum* representing Type 3 *Erodium* plastid genomes, including that of *E. crassifolium*, which is identical in gene order and nearly identical in size.

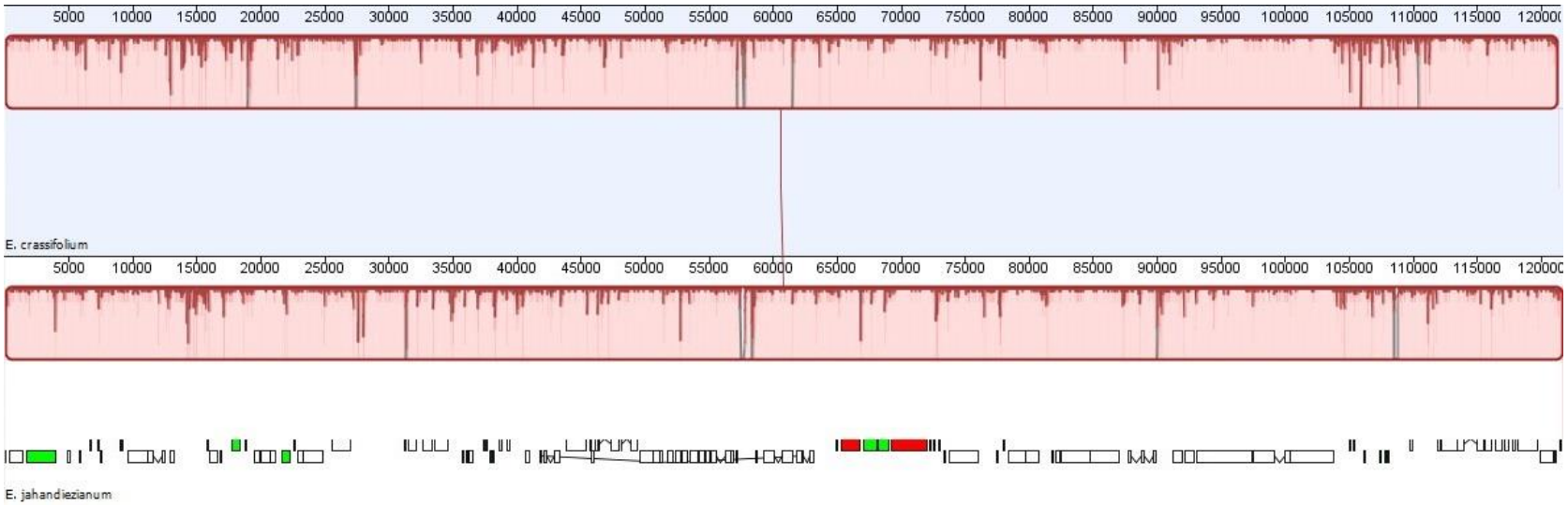

Figure 4.5. MAUVE alignment of type 3 genomes *E. jahandiezianum* and *E. crassifolium* showing that the two genomes are collinear.

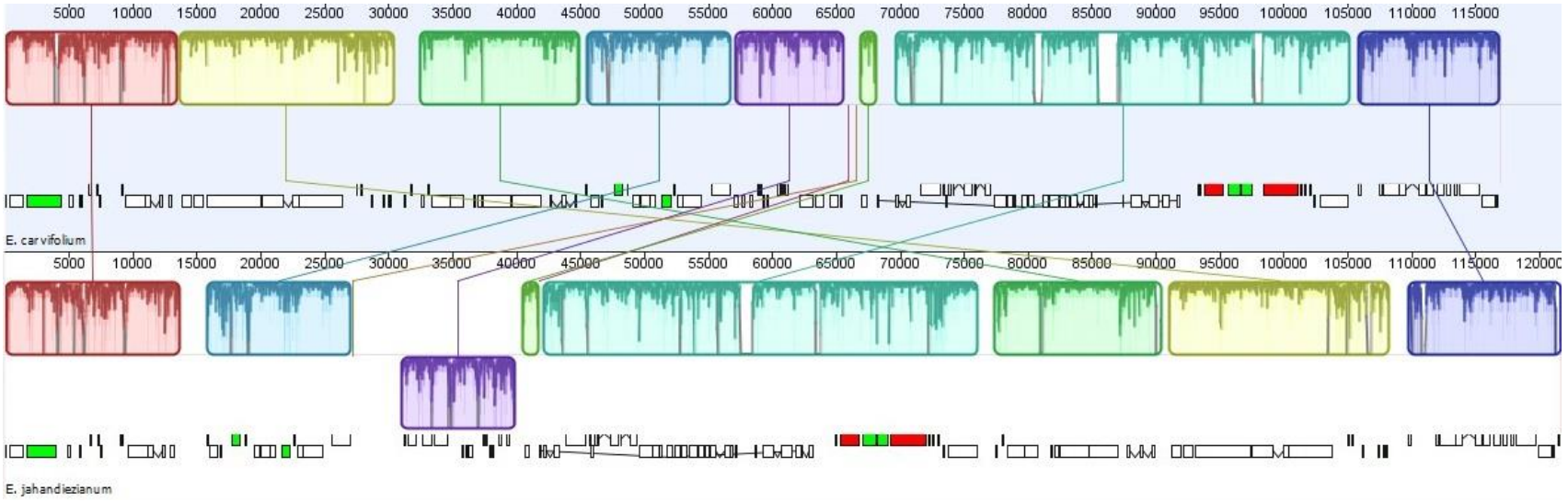

Figure 4.6. MAUVE alignment of type 3 genome *E. jahandiezianum* and type 2 genome *E. carvifolium* showing that five inversions are necessary to derive the type 3 gene order from the inferred ancestral order for the genus.

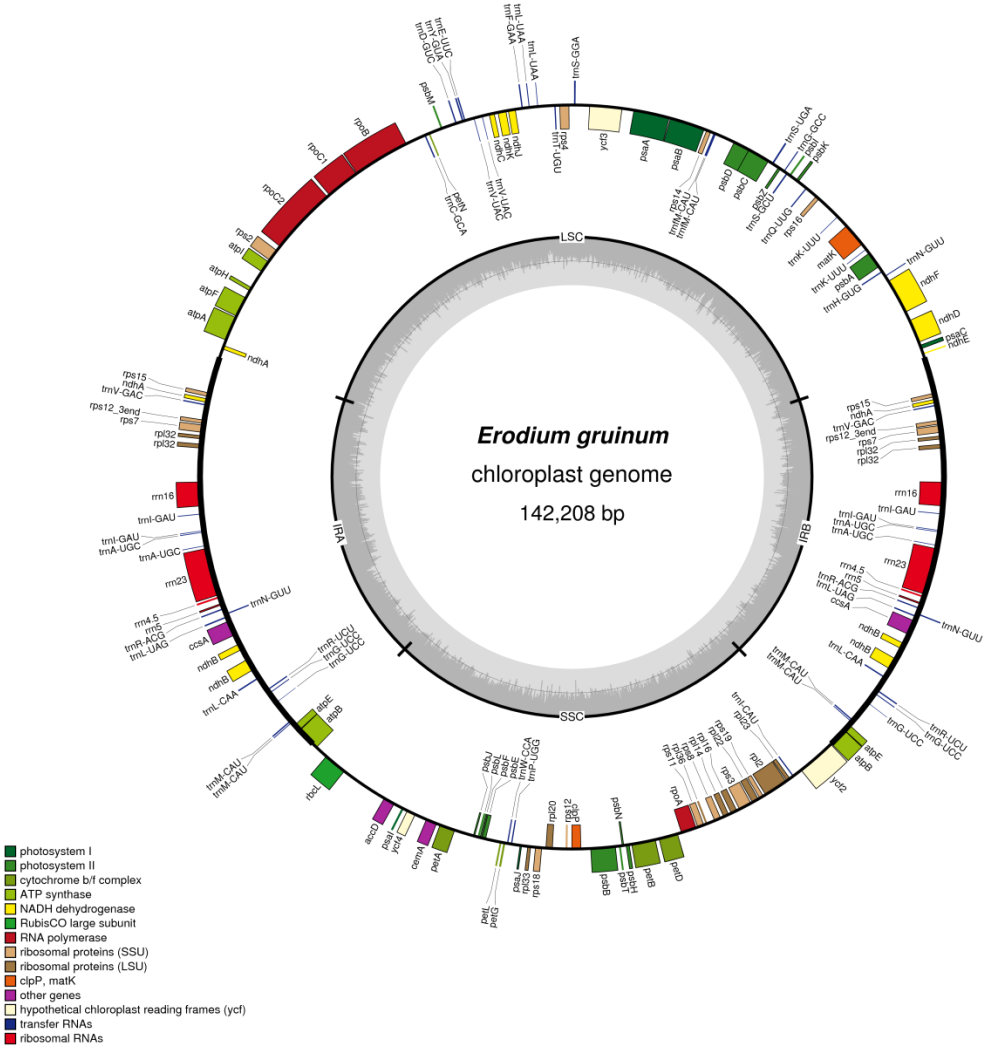

Figure 4.7. Genome map of *E. gruinum*, the Type 4 *Erodium* plastid genome. *E. gruinum* has a novel large, 25kb inverted repeat of recent origin.

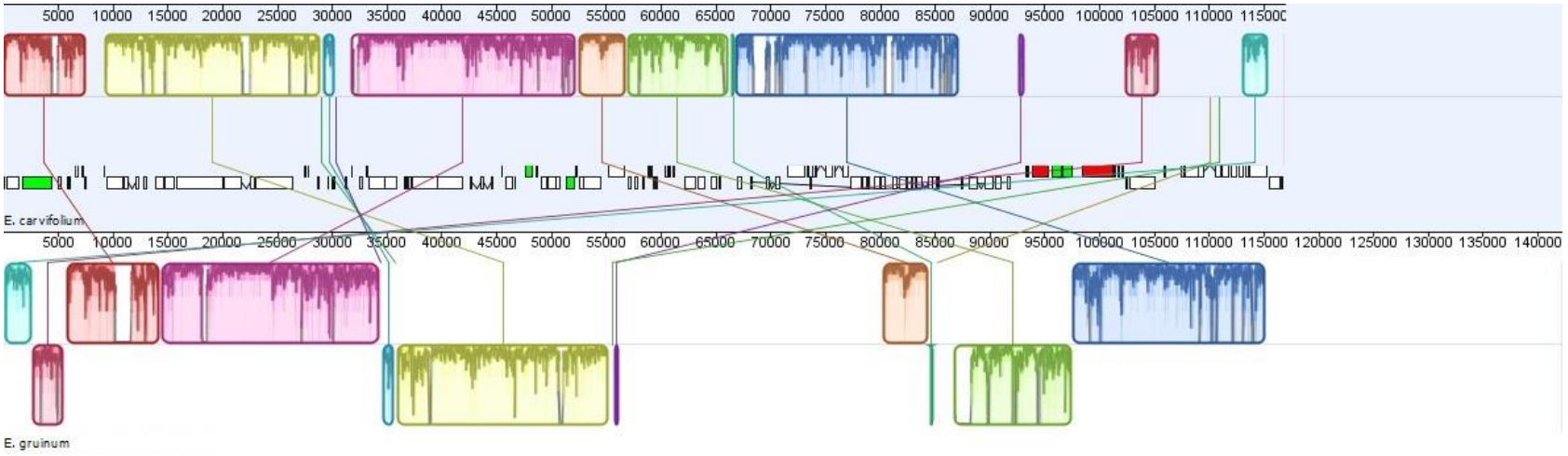

Figure 4.8. MAUVE alignment of type 4 genome *E. gruinum* and type 2 genome *E. carvifolium* showing that 10 inversions are necessary to derive the type 4 gene order from the inferred ancestral order for the genus.

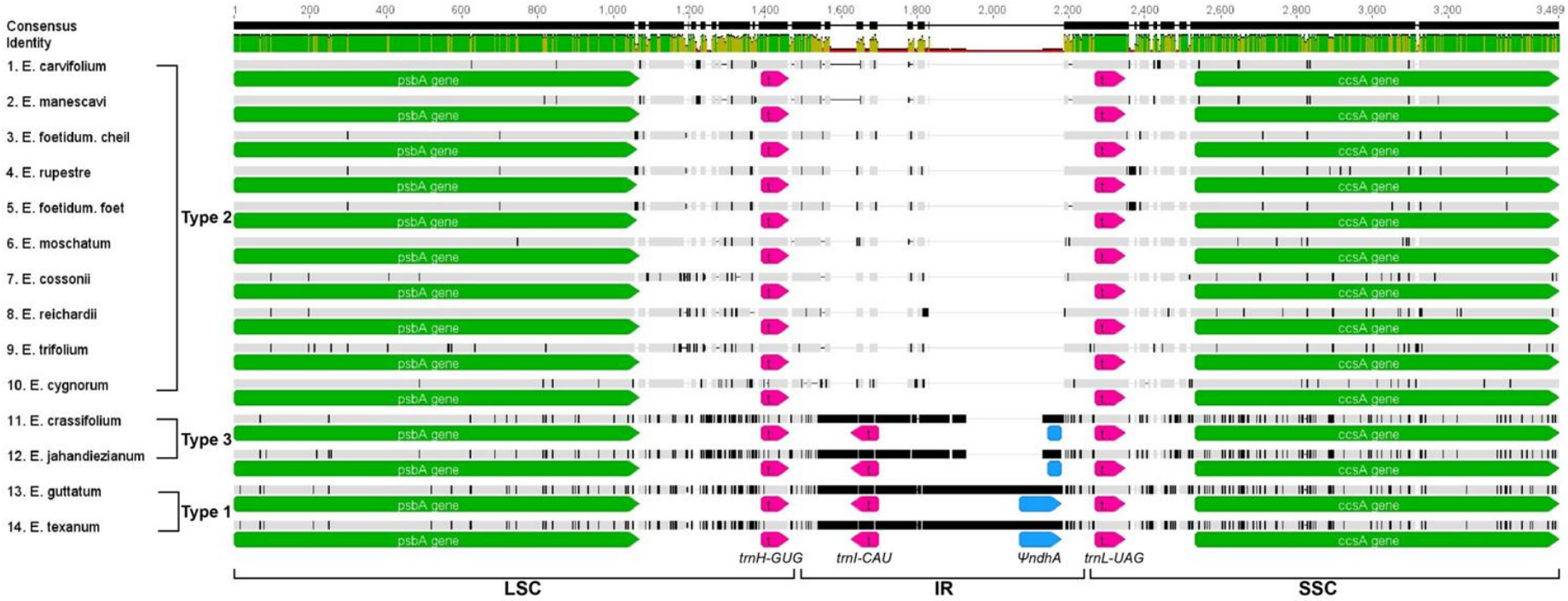

Figure 4.9. An annotated nucleotide alignment of the region formerly flanking the copy of the IR lost on the branch leading to *Erodium*. Type 1 and Type 3 plastid genomes retain a pseudogene of *ndhA* in this region as well as a second copy of *trnI-CAU*. In Type 2 genomes this region has been reduced to >300bp

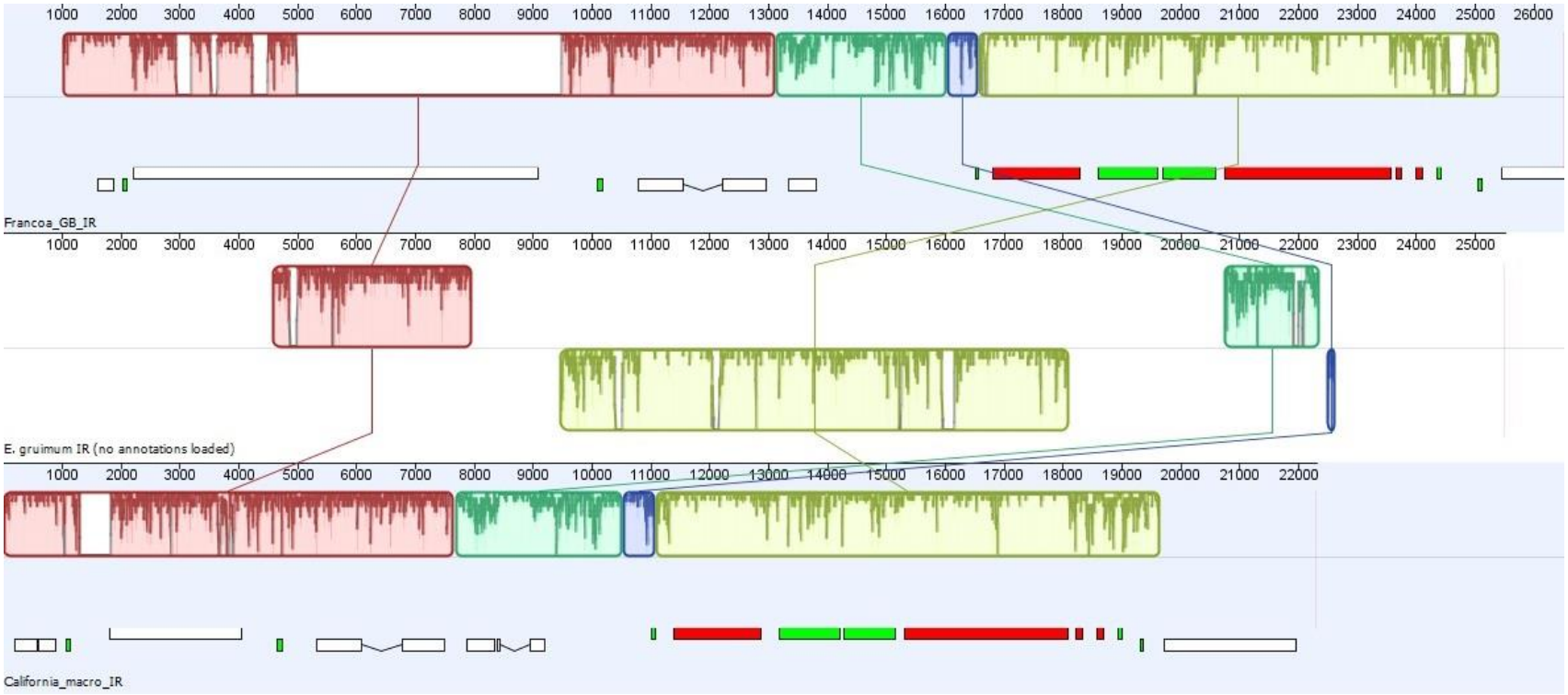

Figure 4.10. MAUVE alignment of the IR regions of *E. gruinum*, *California macrophylla*, and unrearranged Geraniales outgroup *Francoa sonchifolia*.

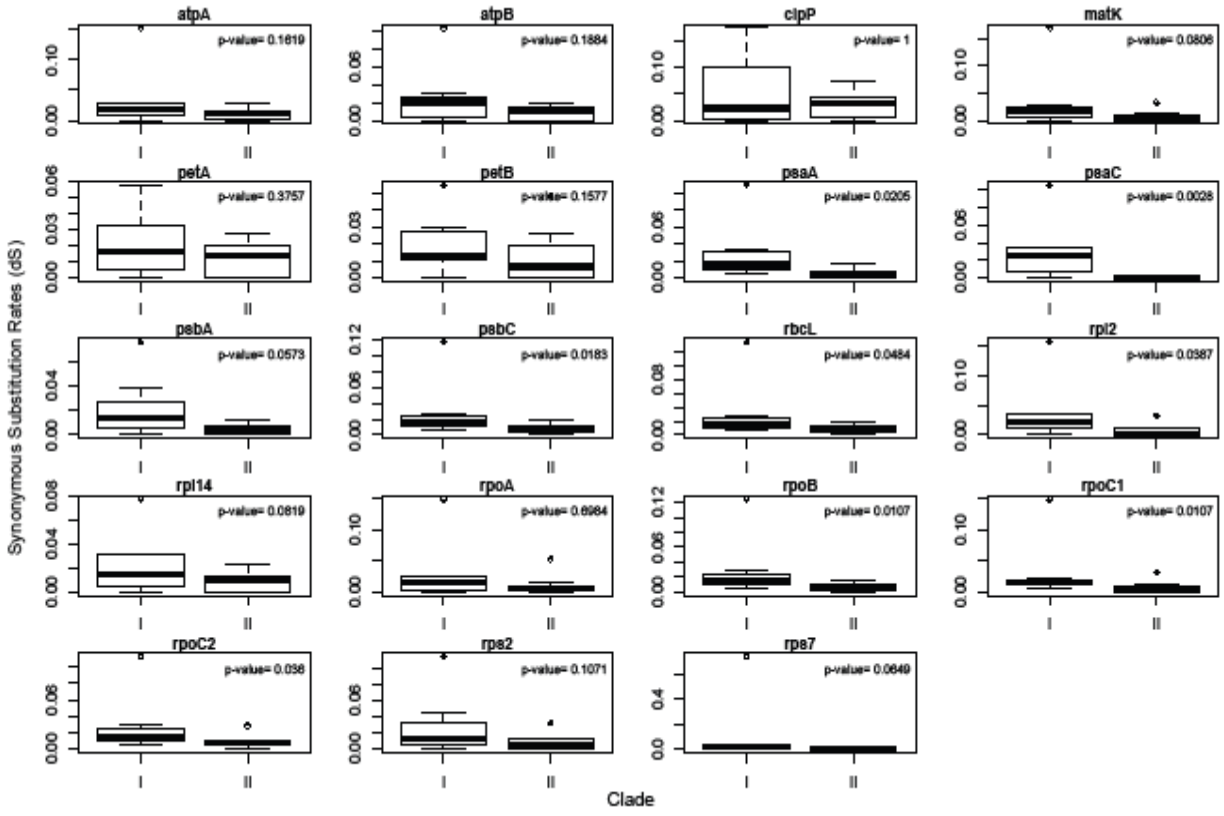

Figure 4.11. Boxplots of the Wilcoxon test comparing *dS* on internal branches of Clade I and Clade II for the 19 gene data set. *dS* was significantly different between the clades for 8 of the 19 genes.

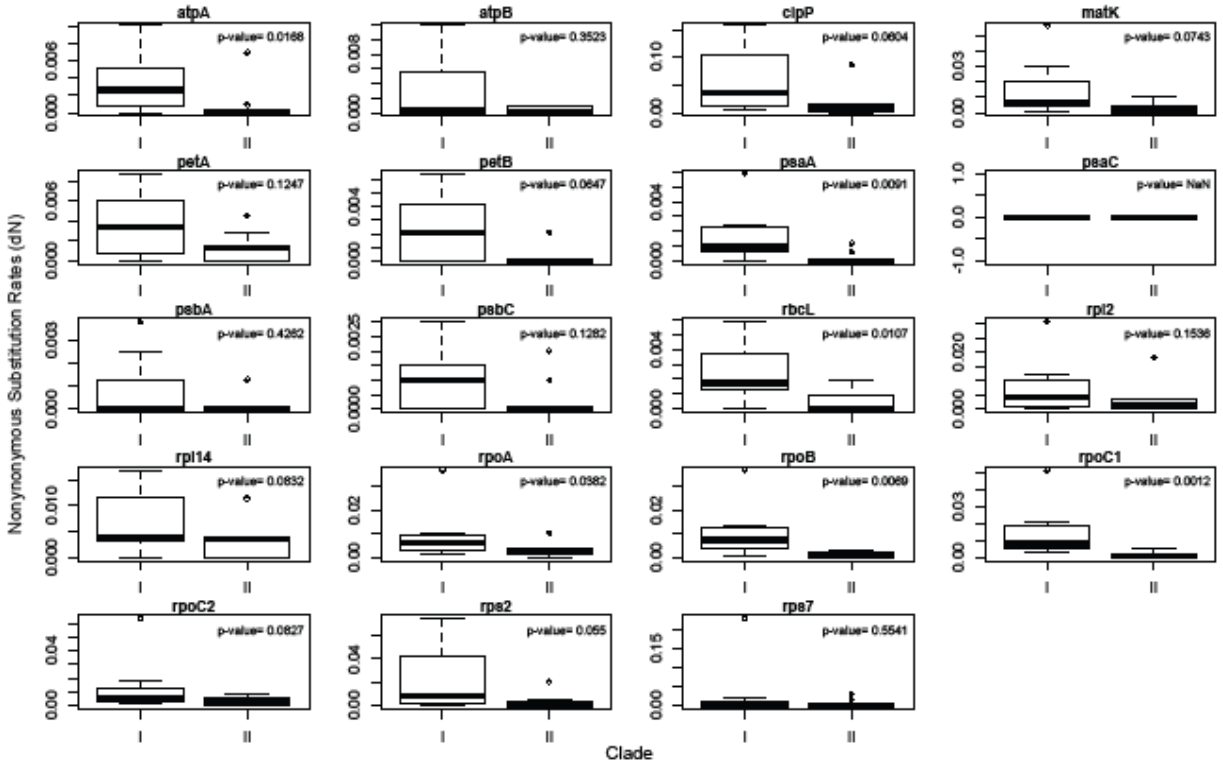

Figure 4.12. Boxplots of the Wilcoxon test comparing *dN* on internal branches of Clade I and Clade II for the 19 gene data set. *dN* was significantly different between the clades for 6 of the 19 genes.

References

Chapter 2

Atwood, JT (1986) The size of the Orchidaceae and the systematic distribution of epiphytic orchids. Selbyana 9:171-186.

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

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

Bowe, LM, Coat, G, dePamphilis, CW (2000) Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. Proc Natl Acad Sci USA 97:4092–4097.

Braukmann, TWA, Kuzmina, M, Stefanovic, S (2009) Loss of all plastid *ndh* genes in Gnetales and conifers: extent and evolutionary significance for the seed plant phylogeny. Curr Gen 55: 323-337.

Casano, LM, Martin, M, and Sabater, B (2001) Hydrogen peroxide mediates the induction of chloroplast Ndh complex under photooxidative stress in barley. Plant Physiol. 125:1450-1458.

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

Chaw, SM, Parkinson, CL, Cheng, Y, Vincent, TM, Palmer, JD (2000) Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. Proc Natl Acad Sci USA 97:4086–4091.

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

Chevreux B MIRA: An Automated Genome and EST Assembler. Available at: http://chevreux.org/thesis/index.html [Accessed August 22, 2009].

Chumley, TM, Palmer, JD, Mower, JP, Fourcade, HM, Calie, PJ, Boore, JL, and Jansen, RK (2006) The complete chloroplast genome sequence of *Pelargonium* x *hortorum*: Organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Mol Biol Evol 23:2175-2190.

Conant, GC and Wolfe KH (2007) GenomeVx: Simple web-based creation of editable circular chromosome maps. Bioinformatics 24: 861-862.

Dean, FB, Nelson, JR, Giesler, TL, and Lasken, RS (2001) Rapid amplification of plasmid and phage DNA using phi29 DNA polymerase and multiply-primed rolling circle amplification. Gen Res 11:1095-1099.

Drummond, AJ, Ashton, B, Buxton, S, Cheung, M, Heled, J, Kearse, M, Moir, R, Stones-Havas, S, Thierer, T, and Wilson, A (2009) Geneious v4.8, Available from http://www.geneious.com.

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

Endo, T, Shikanai, T, Takabayashi, A, Asada, K, and Sato, F (1999) The role of chloroplastid NAD(P)H dehydrogenase in photoprotection. FEBS Lett 457:5-8.

Fiz, O., Vargas, P., Alarcón M.L., and Aldasoro, J.J. 2006. Phylogenetic relationships and evolution in *Erodium* (Geraniaceae) based on *trnL-trnF* sequences. Syst Bot 31:739-763.

Fiz, O, Vargas, P, Alarcón ML, Aedo, C, Garcia, JL, and Aldasoro, JJ (2008) Phylogeny and historical biogeography of Geraniaceae in relation to climate changes and pollination ecology. Syst Bot 33:326-342.

Guisinger, MM, Kuehl, JV, Boore, JL, and Jansen, RK (2008) Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. Proc Natl Acad Sci USA 105:18424-18429.

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

Guittonneau, GG Taxonomy, ecology, and phylogeny of genus *Erodium* l'Her. in the Mediterranean region. In: Vorster, P ed(s). Proc Intnl Geraniaceae Symp, Univ Stellenbosch, Sept. 1990 pp. 69-91.

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

Hiratsuka, J, Shimada, H, Whittier, R, Ishibashi, T, Sakamoto, M, Mori, M, Kondo, C, Honji, Y, Sun, C, Meng, B, Li, Y, Kanno, A, Nishizawa, Y, Hirai, A, Shinozaki, K, and Sugiura, M (1989) The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Mol Gen Genet 217:185-94.

Horváth, EM, Peter, SO, Joët, T, Rumeau, D, Cournac, L, Horváth, GV, Kavanaugh, TA, Schäfer, C, Peltier, G, and Medgyesy, P (2000) Targeted inactivation of the

plastid *ndhB* gene in tobacco results in an enhanced sensitivity of photosynthesis to moderate stomatal closure. Plant Physiol 123:1337-1350.

Huang, CY, Ayliffe, MA, and Timmis, JN (2003) Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature 422:72-76.

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

Konishi, T, Shinohara, K, Yamada, K, Sasaki, Y (1996) Acetyl-CoA carboxylase in higher plants: Most plants other than Gramineae have both the prokaryotic and eukaryotic forms of this enzyme. Plant Cell Physiol 37: 117-122.

Jansen, RK, Cai, Z, Raubeson, LA, Daniell, H, dePamphilis, CW, Leebens-Mack, J, Müller, KF, Guisinger-Bellian, M, Haberle, RC, Hansen, AK, Chumley, TW, Lee, S, Peery, R, McNeal, JR, Kuehl, JV, Boore, JL (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA 104:19369-19374.

Maier, RM, Zeltz, P, Kössel, H, Bonnard, G, Gualberto, JM, Grienenberger, JM (1996) RNA editing in plant mitochondria and chloroplasts. Plant Mol Biol 32:343–65.

Martin, M, Casano, LM, Zapata, JM, Guéra, A, Del Campo, EM, Schmitz-Linneweber, C, Maier, RM, and Sabater, B (2004) Role of thylakoid Ndh complex and peroxidase in the protection against photo-oxidative stress: fluorescence and enzyme activities in wild-type and *ndhF*-deficient tobacco. Physiol Plant 122:442-452.

Martin, M and Sabater, B (2010) Plastid *ndh* genes in plant evolution. Plant Physiol and Biochem 48: 636-645.

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

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

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

Neyland, R and Urbatsch, LE (1996) Phylogeny of Subfamily Epidendroideae (Orchidaceae) inferred from *ndhF* chloroplast gene sequences. Amer J Bot 83:1195-1206.

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

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

Palmer, JD, Osorio, B, and Thomson, WF (1988) Evolutionary significance of inversions in legume chloroplast DNA. Curr Genet 14:65-74.

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

Ramirez, SR, Gravendeel, B, Singer, RB, Marshall, CR, and Pierce, NE (2007) Dating the origin of the Orchidaceae from a fossil orchid with its pollinator. Nature 448:1042-1045.

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

Rumeau, D, Peltier, G, and Cornac, L (2007) Chlororespiration and cyclic electron flow around PSI during photosynthesis and plant stress response. Plant Cell Environ 30:1041-1051.

Rumeau, D, Bécuwe-Linka, N, Beyly, A, Louwagie, M, Garin, J, and Peltier, G (2005) New subunits NDH-M, -N, and –O, encoded by nuclear genes, are essential for plastid Ndh complex functioning in higher plants. Plant Cell 17:219-232.

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

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

Wang, X, Tank, DC, and Sang, T (2000) Phylogeny and divergence times in Pinaceae: evidence from three genomes. Mol Biol Evol 17:773-781.

Werner, T, Braukmann, A, Kuzmina, M, and Stefanovic, S (2009) Loss of all plastid *ndh* genes in Gnetales and conifers: extent and evolutionary significance for the seed plant phylogeny. Curr Genet 55:323-337.

Wu, F, Chan, M, Liao, D, Hsu, C, Lee, Y, Daniell, H, Duvall, MR, and Lin, C (2010) Complete chloroplast genome of *Oncidium* Gower Ramsey and evaluation of

molecular markers for identification and breeding of Oncidiinae. BMC Plant Bio 10:68.

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

Zhong, B, Yonezawa, T, Zhong, Y, and Hasegawa, M (2010) The position of Gnetales among seed plants: Overcoming pitfalls of chloroplast phylogenomics. Mol Biol Evol doi:10.1093/molbev/msq170.

Chapter 3

AK Hansen, & JC Blazier. (n.d.). Comparative Chloroplast Genomics in Passiflora.

Bakker, F. T., Culham, A., Hettiarachi, P., Touloumenidou, T., & Gibby, M. (2004). Phylogeny of Pelargonium (Geraniaceae) based on DNA sequences from three genomes. *Taxon*, 17–28.

Blazier, J. C., Guisinger, M. M., & Jansen, R. K. (2011). Recent loss of plastid-encoded ndh genes within Erodium (Geraniaceae). *Plant Molecular Biology*, *76*(3-5), 263–272. doi:10.1007/s11103-011-9753-5

Casola, C., & Hahn, M. W. (2009). Gene Conversion Among Paralogs Results in Moderate False Detection of Positive Selection Using Likelihood Methods. *Journal of Molecular Evolution*, *68*(6), 679–687. doi:10.1007/s00239-009-9241-6

Chevreux, B., Wetter, T., & Suhai, S. (1999). Genome sequence assembly using trace signals and additional sequence information. In *Computer science and biology: proceedings of the German conference on bioinformatics (GCB)* (Vol. 99, pp. 45–56). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.23.7465&rep=rep1&type=pdf

Chumley, T. W., Palmer, J. D., Mower, J. P., Fourcade, H. M., Calie, P. J., Boore, J. L., & Jansen, R. K. (2006). The Complete Chloroplast Genome Sequence of Pelargonium × hortorum: Organization and Evolution of the Largest and Most Highly Rearranged Chloroplast Genome of Land Plants. *Molecular Biology and Evolution*, *23*(11), 2175–2190. doi:10.1093/molbev/msl089

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. *Molecular Biology and Evolution*, *28*(7), 2077–2086. doi:10.1093/molbev/msr028

Downie, S. R., & Palmer, J. D. (1992). Use of Chloroplast DNA Rearrangements in Reconstructing Plant Phylogeny. In P. S. Soltis, D. E. Soltis, & J. J. Doyle (Eds.), *Molecular Systematics of Plants* (pp. 14–35). Springer US. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4615-3276-7_2

Doyle, J. J., Doyle, J. L., & Palmer, J. D. (1995). Multiple Independent Losses of Two Genes and One Intron from Legume Chloroplast Genomes. *Systematic Botany*, *20*(3), 272–294. doi:10.2307/2419496

Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, Field M, Heled J, Kearse M, Markowitz S, Moir R, Stones-Havas S, Sturrock S, Thierer T, Wilson A. (n.d.). (Version 6.1.5). Biomatters Ltd.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. doi:10.1093/nar/gkh340

Gantt, J. S., Baldauf, S. L., Calie, P. J., Weeden, N. F., & Palmer, J. D. (1991). Transfer of rpl22 to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. *The EMBO Journal*, *10*(10), 3073–3078.

Goffinet, B., Wickett, N. J., Shaw, A. J., & Cox, C. J. (2005). Phylogenetic significance of the rpoA loss in the chloroplast genome of mosses. *Taxon*, *54*(2), 353–360.

Gordon, D., Abajian, C., & Green, P. (1998). Consed: A Graphical Tool for Sequence Finishing. *Genome Research*, *8*(3), 195–202. doi:10.1101/gr.8.3.195

Gruissem, W., & Zurawski, G. (1985). Analysis of promoter regions for the spinach chloroplast rbcL, atpB and psbA genes. *The EMBO Journal*, *4*(13A), 3375.

Hao, W. (2010). OrgConv: detection of gene conversion using consensus sequences and its application in plant mitochondrial and chloroplast homologs. *BMC Bioinformatics*, *11*(1), 114. doi:10.1186/1471-2105-11-114

Jansen, R. K., Saski, C., Lee, S.-B., Hansen, A. K., & Daniell, H. (2010). Complete Plastid Genome Sequences of Three Rosids (Castanea, Prunus, Theobroma): Evidence for At Least Two Independent Transfers of rpl22 to the Nucleus. *Molecular Biology and Evolution*, *28*(1), 835–847. doi:10.1093/molbev/msq261

Jansen, Robert K., Cai, Z., Raubeson, L. A., Daniell, H., dePamphilis, C. W., Leebens-Mack, J., … Boore, J. L. (2007). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences*, *104*(49), 19369–19374. doi:10.1073/pnas.0709121104

Jansen, Robert K., Raubeson, L. A., Boore, J. L., dePamphilis, C. W., Chumley, T. W., Haberle, R. C., … Cui, L. (2005). Methods for Obtaining and Analyzing Whole Chloroplast Genome Sequences. In and E. H. R. Elizabeth A. Zimmer (Ed.), *Methods in Enzymology* (Vol. Volume 395, pp. 348–384). Academic Press. Retrieved from http://www.sciencedirect.com/science/article/pii/S0076687905950209

Katoh, K., Asimenos, G., & Toh, H. (2009). Multiple Alignment of DNA Sequences with MAFFT. In D. Posada (Ed.), *Bioinformatics for DNA Sequence Analysis* (Vol.

537, pp. 39–64). Totowa, NJ: Humana Press. Retrieved from http://www.springerlink.com/index/10.1007/978-1-59745-251-9_3

Koressaar, T., & Remm, M. (2007). Enhancements and modifications of primer design program Primer3. *Bioinformatics*, *23*(10), 1289–1291. doi:10.1093/bioinformatics/btm091

Krause, K., Berg, S., & Krupinska, K. (2003). Plastid transcription in the holoparasitic plant genus Cuscuta: parallel loss of the rrn16 PEP-promoter and of the rpoA and rpoB genes coding for the plastid-encoded RNA polymerase. *Planta*, *216*(5), 815–823.

Little, M. C., & Hallick, R. B. (1988). Chloroplast rpoA, rpoB, and rpoC genes specify at least three components of a chloroplast DNA-dependent RNA polymerase active in tRNA and mRNA transcription. *Journal of Biological Chemistry*, *263*(28), 14302–14307.

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., … Bryant, S. H. (2010). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research*, *39*(Database), D225–D229. doi:10.1093/nar/gkq1189

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380. doi:10.1038/nature03959

Millen, R. S., Olmstead, R. G., Adams, K. L., Palmer, J. D., Lao, N. T., Heggie, L., … Wolfe, K. H. (2001). Many Parallel Losses of infA from Chloroplast DNA during Angiosperm Evolution with Multiple Independent Transfers to the Nucleus. *The Plant Cell Online*, *13*(3), 645–658. doi:10.1105/tpc.13.3.645

Morton, B. R., & Clegg, M. T. (1993). A chloroplast DNA mutational hotspot and gene conversion in a noncoding region near rbcL in the grass family (Poaceae). *Current Genetics*, *24*(4), 357–365.

Schneider, A., Souvorov, A., Sabath, N., Landan, G., Gonnet, G. H., & Graur, D. (2009). Estimates of Positive Darwinian Selection Are Inflated by Errors in Sequencing, Annotation, and Alignment. *Genome Biology and Evolution*, *1*(0), 114–118. doi:10.1093/gbe/evp012

Serino, G., & Maliga, P. (1998). RNA Polymerase Subunits Encoded by the Plastid rpoGenes Are Not Shared with the Nucleus-Encoded Plastid Enzyme. *Plant Physiology*, *117*(4), 1165–1170. doi:10.1104/pp.117.4.1165

Shi, C., Liu, Y., Huang, H., Xia, E.-H., Zhang, H.-B., & Gao, L.-Z. (2013). Contradiction between Plastid Gene Transcription and Function Due to Complex Posttranscriptional Splicing: An Exemplary Study of ycf15 Function and Evolution in Angiosperms. *PLoS ONE*, *8*(3), e59620. doi:10.1371/journal.pone.0059620

Sloan, D. B., Oxelman, B., Rautenberg, A., & Taylor, D. R. (2009). Phylogenetic analysis of mitochondrial substitution rate variation in the angiosperm tribe Sileneae. *BMC Evolutionary Biology*, *9*(1), 260. doi:10.1186/1471-2148-9-260

Sugiura, C., Kobayashi, Y., Aoki, S., Sugita, C., & Sugita, M. (2003). Complete chloroplast DNA sequence of the moss Physcomitrella patens: evidence for the loss and relocation of rpoA from the chloroplast to the nucleus. *Nucleic Acids Research*, *31*(18), 5324–5331. doi:10.1093/nar/gkg726

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*(22), 4673–4680. doi:10.1093/nar/22.22.4673

Weng, M.-L., Ruhlman, T. A., Gibby, M., & Jansen, R. K. (2012). Phylogeny, rate variation, and genome size evolution of Pelargonium (Geraniaceae). *Molecular Phylogenetics and Evolution*, *64*(3), 654–670. doi:10.1016/j.ympev.2012.05.026

Wickett, N. J., Honaas, L. A., Wafula, E. K., Das, M., Huang, K., Wu, B., … dePamphilis, C. W. (2011). Transcriptomes of the Parasitic Plant Family Orobanchaceae Reveal Surprising Conservation of Chlorophyll Synthesis. *Current Biology*, *21*(24), 2098–2104. doi:10.1016/j.cub.2011.11.011

Wyman, S. K., Jansen, R. K., & Boore, J. L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, *20*(17), 3252–3255. doi:10.1093/bioinformatics/bth352

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. doi:10.1093/molbev/msm088

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. doi:10.1101/gr.074492.107

Zwickl, D. J. (2008). Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Retrieved from http://repositories.lib.utexas.edu/handle/2152/2666

Chapter 4

Blazier, J.C., Guisinger, M.M., and Jansen, R.K. (2011). Recent loss of plastid-encoded ndh genes within Erodium (Geraniaceae). Plant Mol. Biol. *76*, 263–272.

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

Chevreux, B., Wetter, T., and Suhai, S. (1999). Genome sequence assembly using trace signals and additional sequence information. In Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB), pp. 45–56.

Cosner, M.E., Jansen, R.K., Palmer, J.D., and Downie, S.R. (1997). The highly rearranged chloroplast genome of Trachelium caeruleum (Campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. Curr. Genet. *31*, 419–429.

Downie, S.R., and Palmer, J.D. (1992). Use of Chloroplast DNA Rearrangements in Reconstructing Plant Phylogeny. In Molecular Systematics of Plants, P.S. Soltis, D.E. Soltis, and J.J. Doyle, eds. (Springer US), pp. 14–35.

Fiz, O., Vargas, P., Alarcon, M.L., and Aldasoro, J.J. (2006). Phylogenetic relationships and evolution in Erodium (Geraniaceae) based on trnL-trnF sequences. Syst. Bot. *31*, 739–763.

Goulding, S.E., Wolfe, K.H., Olmstead, R.G., and Morden, C.W. (1996). Ebb and flow of the chloroplast inverted repeat. Mol. Gen. Genet. MGG *252*, 195–206.

Guisinger, M.M., Kuehl, J.V., Boore, J.L., and Jansen, R.K. (2008). Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. Proc. Natl. Acad. Sci. *105*, 18424–18429.

Guisinger, M.M., Kuehl, J.V., Boore, J.L., and Jansen, R.K. (2011). Extreme Reconfiguration of Plastid Genomes in the Angiosperm Family Geraniaceae: Rearrangements, Repeats, and Codon Usage. Mol. Biol. Evol. *28*, 583–600.

Haberle, R.C., Fourcade, H.M., Boore, J.L., and Jansen, R.K. (2008). Extensive Rearrangements in the Chloroplast Genome of Trachelium caeruleum Are Associated with Repeats and tRNA Genes. J. Mol. Evol. *66*, 350–361.

Heinhorst, S., and Cannon, G.C. (1993). DNA replication in chloroplasts. J. Cell Sci. *104*, 1–9.

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

Horiuchi, T., and Watanabe, T. Double Rolling Circle Replication (DRCR): Involvement in Gene Amplification and Genome Replication including HSV.

James R. Broach, and Fredric C. Volkert (1991). Circular DNA Plasmids of Yeasts. In Volume I: The Molecular and Cellular Biology of the Yeast Saccharomyces: Genome Dynamics, Protein Synthesis, and Energetics, (Cold Spring Harbor Laboratory Press), pp. 297–331.

Jansen, R.K., Raubeson, L.A., Boore, J.L., dePamphilis, C.W., Chumley, T.W., Haberle, R.C., Wyman, S.K., Alverson, A.J., Peery, R., Herman, S.J., et al. (2005).

Methods for Obtaining and Analyzing Whole Chloroplast Genome Sequences. In Methods in Enzymology, and E.H.R. Elizabeth A. Zimmer, ed. (Academic Press), pp. 348–384.

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., dePamphilis, C.W., Leebens-Mack, J., Müller, K.F., Guisinger-Bellian, M., Haberle, R.C., Hansen, A.K., et al. (2007). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc. Natl. Acad. Sci. *104*, 19369–19374.

Katoh, K., Asimenos, G., and Toh, H. (2009). Multiple Alignment of DNA Sequences with MAFFT. In Bioinformatics for DNA Sequence Analysis, D. Posada, ed. (Totowa, NJ: Humana Press), pp. 39–64.

Kunnimalaiyaan, M., and Nielsen, B.L. (1997). Fine mapping of replication origins (oriA and oriB) in Nicotiana tabacum chloroplast DNA. Nucleic Acids Res. *25*, 3681–3686.

Lehman, I.R., and Boehmer, P.E. (1999). Replication of Herpes Simplex Virus DNA. J. Biol. Chem. *274*, 28059–28062.

Lohse, M., Drechsel, O., and Bock, R. (2007). OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr. Genet. *52*, 267–274.

Palmer, J.D. (1983). Chloroplast DNA exists in two orientations. Nature *301*, 92–93.

Palmer, J.D., and Thompson, W.F. (1982). Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. Cell *29*, 537–550.

Perry, A.S., and Wolfe, K.H. (2002). Nucleotide Substitution Rates in Legume Chloroplast DNA Depend on the Presence of the Inverted Repeat. J. Mol. Evol. *55*, 501–508.

Pond, S.L.K., and Muse, S.V. (2005). HyPhy: hypothesis testing using phylogenies. In Statistical Methods in Molecular Evolution, (Springer), pp. 125–181.

Schattner, P., Brooks, A.N., and Lowe, T.M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. *33*, W686–W689.

Selosse, M.-A., Albert, B., and Godelle, B. (2001). Reducing the genome size of organelles favours gene transfer to the nucleus. Trends Ecol. Evol. *16*, 135–141.

Sloan, D.B., Oxelman, B., Rautenberg, A., and Taylor, D.R. (2009). Phylogenetic analysis of mitochondrial substitution rate variation in the angiosperm tribe Sileneae. BMC Evol. Biol. *9*, 260.

Sugiura, M., Hirose, T., and Sugita, M. (1998). Evolution and Mechanism of Translation in Chloroplasts. Annu. Rev. Genet. *32*, 437–459.

Weng, M.-L., Ruhlman, T.A., Gibby, M., and Jansen, R.K. (2012). Phylogeny, rate variation, and genome size evolution of Pelargonium (Geraniaceae). Mol. Phylogenet. Evol. *64*, 654–670.

Wojciechowski, M.F., Lavin, M., and Sanderson, M.J. (2004). A phylogeny of legumes (Leguminosae) based on analysis of the plastid matK gene resolves many well-supported subclades within the family. Am. J. Bot. *91*, 1846–1862.

Wolfe, K.H., Li, W.H., and Sharp, P.M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. *84*, 9054–9058.

Wyman, S.K., Jansen, R.K., and Boore, J.L. (2004). Automatic annotation of organellar genomes with DOGMA. Bioinformatics *20*, 3252–3255.

Yamada, T. (1991). Repetitive sequence-mediated rearragements in Chlorella ellipsoidea chloroplast DNA: completion of nucleotide sequence of the large inverted repeat. Curr. Genet. *19*, 139–147.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. *24*, 1586–1591.

Zerbino, D.R., and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829.

Zwickl, D.J. (2008). Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion.