The Dissertation Committee for Jin Zhang Certifies that this is the approved version of the following dissertation:

Coevolution between Nuclear and Plastid Genomes in Geraniaceae

Committee:

Robert K. Jansen, Supervisor

David L. Herrin

C. Randy Linder

Claus O. Wilke

Stanley J. Roux

Coevolution between Nuclear and Plastid Genomes in Geraniaceae

by

Jin Zhang, B.S.

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

August 2015

Acknowledgements

I would like to express my greatest appreciation to my wonderful supervisor, Dr. Robert K. Jansen, for giving me the opportunity to work in his lab, allowing me the freedom for exploration and making mistakes, being the ultimate supporter whenever I needed it, and motivating me every day to become a better scientist. I also want to thank Dr. Tracey Ruhlman for all the inspiration in my research, comprehensive guidance on my experimental work and my scientific writing skills, and I want to thank our collaborator Dr. Jeffery Mower for his guidance in bioinformatics techniques and the accommodations for my short visit to University of Nebraska Lincoln. Additionally, I would like to express my appreciation to my committee (Dr. David Herrin, Dr. Randy Linder, Dr. Claus Wilke and Dr. Stanley Roux) for their support, time and valuable suggestions.

I want to thank my long standing lab mates Dr. Chris Blazier, Maolun Weng, Erika Schwarz, Anna Yu and Dr. Seongjun Park for the inspiring discussions and being incredible resources of plant biology and evolutionary knowledge; Dr. Michael Gruenstaeudl for the help in my cloning experiments; Dhivya Arasappan for the help in scripting in my early PhD life; the Texas Advanced Computing Center for their valuable and patient assistance. I also want to thank all my friends during my PhD life, my best friend Dr. Shanshan Cao, who came to the USA with me in same year, for his incredible support and friendship; my five-year roommates, who are also my class mates, An Li, Zicheng Hu, Xianzhe Wang, Chao Liu and Dongyang Wang, for the inspiring discussion in research and the sharing of all the joyful moments in our lives.

Lastly I would like to thank my family, my father Qingyu Zhang, my mother Feng Jin and my girlfriend Xiwei Yan, for their lifelong love, support and understanding during all these years. From them I learned how to be a man who is honest, supportive and hard working.

Coevolution between Nuclear and Plastid Genomes in Geraniaceae

Jin Zhang, PhD

The University of Texas at Austin, 2015

Supervisor: Robert K. Jansen

Plastid genomes of angiosperms are highly conserved in both genome organization and nucleotide substitution rates. Geraniaceae have highly rearranged genomes and elevated nucleotide substitution rates, which provides an attractive system to study nuclear-plastid genome coevolution. My dissertation research has focused on two areas of nuclear-plastid genome coevolution in Geraniaceae. First, I have investigated the correlation of nucleotide substitution rates between nuclear and plastid genes that encode interacting subunits that form the multi-subunit complex of Plastid Encoded RNA Polymerase (PEP). Second, the hypothesis that the unusual changes of plastid genome organization and elevated nucleotide substitution rates of plastid encoded genes is the result of alterations in nuclear encoded DNA replication, recombination and repair (DNA RRR) genes is tested. The second chapter investigates the optimal methods for transcriptome sequencing/assembly. My findings supported the use of transcriptome assemblers optimized for Illumina sequencing platform (Trinity and SOAPdenovo-trans). The third chapter investigated coevolution of nucleotide substitution rates between plastid encoded RNAP (*rpoA*, *rpoB*, *rpoC1*, *rpoC2*) and nuclear encoded SIG (*sig1-6*) genes that are part of the multi-subunit complex PEP. Using the transcriptomes of 27 Geraniales species I extracted the PEP genes and performed a systematic correlation test. I detected strong correlations of *dN* (nonsynonymous substitutions) but not *dS* (synonymous

substitutions) between RNAP and SIG but no correlations were detected for the control genes, which provides a plausible explanation for the cause of plastome-genome incompatibility in Geraniaceae. The fourth chapter investigated the effect of DNA RRR system on the aberrant evolutionary phenomena in Geraniaceae plastid genomes. I extracted DNA RRR and nuclear control genes with different subcellular locations from 27 Geraniales transcriptomes and estimated genome complexity with various measures from plastid genomes of the same species. I detected significant correlations for *dN* but not *dS* for three DNA RRR genes, 10 nuclear encoded plastid targeted (NUCP) and three nuclear encoded mitochondrial targeted (NUMT) genes. The findings of a correlation between *dN* of DNA RRR genes and genome complexity support the hypothesis that changes of plastid genome complexity in Geraniaceae may be caused by dysfunction of DNA RRR systems.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Plastid genomes of angiosperms are highly conserved in organization (i.e., gene order and content) and nucleotide substitution rates. However, several unrelated lineages (Campanulaceae, Ericaceae, Geraniaceae and Fabaceae) have been identified with unusual genomic changes and highly elevated rates of nucleotide substitution. These aberrant plant families, especially Geraniaceae, provide attractive systems to study nuclear-plastid genome coevolution. The research in this dissertation focusses on two areas of nuclear-plastid genome coevolution in Geraniaceae. First, I have investigated the correlation of nucleotide substitution rates between nuclear and plastid genes that encode interacting subunits that form the multi-subunit complex of Plastid Encoded RNA Polymerase (PEP). Second, the hypothesis that the unusual changes of plastid genome organization and elevated nucleotide substitution rates of plastid encoded genes is the result of alterations in nuclear encoded DNA replication, recombination and repair (DNA RRR) genes is tested.

Chapter 2 investigates the optimal methods for *de novo* transcriptome sequencing and assembly. This study was a necessary prerequisite for gathering transcriptome data across the Geraniales for chapters 3 and 4. In chapter 2 transcriptome data of *Geranium maderense* and *Pelargonium* x *hortorum* were gathered using two different sequencing platforms (Illumina and 454) and assembled with five different *de novo* transcriptome assemblers (MIRA, Newbler, SOAPdenovo, SOAPdenovo-trans and Trinity). The results indicated that assemblers optimized for Illumina data (Trinity and SOAPdenovo-trans) produced the optimal transcriptome assembly. In addition, by comparing different amounts of Illumina data no major improvement in breadth of coverage was obtained by sequencing more than six billion nucleotides or sampling four different tissue types rather

than a single type of tissue. This study provided important data for optimizing methods for transcriptome sequencing in the remaining two chapters of the thesis. The results of the study have already been published in *BMC Plant Biology* (Zhang et al., 2013).

Chapter 3 investigates coevolution between nuclear and plastid genes encoding interacting subunits that form the multi-subunit protein complex PEP. Gene coevolution has been widely observed within individuals and between different organisms but rarely has this phenomenon been investigated within a phylogenetic framework. In plants, PEP is a protein complex composed of subunits encoded by both plastid encoded RNAP (*rpoA*, *rpoB*, *rpoC1*, *rpoC2*) and nuclear encoded SIG genes (*sig1-6*). I performed transcriptome and genome sequencing of 27 species of Geraniales, from which we extracted the RNAP and SIG genes. A systematic evaluation of nucleotide substitution rates of these genes was performed. I detected strong correlations of *dN* (nonsynonymous substitutions) but not *dS* (synonymous substitutions) within *rpoB*/*sig1* and *rpoC2*/*sig2* but not for other plastid/nuclear gene pairs, and identified a correlation of *dN*/*dS* ratio between *rpoB*/*C1*/*C2* and *sig1*/*5*/*6*, *rpoC1*/*C2* and *sig2*, and *rpoB*/*C2* and sig3 genes. Analyses of coevolved amino acid positions suggest that structurally-mediated coevolution is not the major driver of plastid-nuclear coevolution. I suggested that the strong correlation of evolutionary rates between SIG and RNAP genes is a plausible cause of plastome-genome incompatibility in Geraniaceae. The results of this chapter were published in *The Plant Cell* (Zhang et al., 2015).

Chapter 4 investigates the role of DNA replication, recombination and repair (DNA RRR) system on the aberrant evolutionary phenomena in Geraniaceae plastid genomes. Alteration of the DNA RRR system has been hypothesized as a potential cause for the unusual phenomena in Geraniaceae plastid genomes but has never been tested. I extracted 12 DNA RRR genes and 90 nuclear encoded control genes with different

subcellular locations (plastid, 30; mitochondrial, 30; other, 30) from the transcriptome of 27 Geraniales species. Plastid genome complexity was estimated in all species using four different measures (rearrangements, repeats, nucleotide insertions/deletions and substitution rates). A systematic evaluation of correlation between nucleotide substitution rates of DNA RRR, nuclear encoded control genes and plastid genome complexity was performed. I detected significant correlations for nonsynonymous (*dN*) but not synonymous (*dS*) substitution rates of three DNA RRR genes (*uvrB*, *why1* and *gyra*), 10 nuclear encoded plastid targeted (NUCP) and three nuclear encoded mitochondrial targeted (NUMT) genes but no other control genes. Comparisons between Geraniales and Brassicales suggested that all five gene groups (plastid encoded, NUCP, NUMT, NUOT and DNA RRR) have higher rates in Geraniales, but only plastid encoded and NUCP gene groups showed significant accelerations of *dN* in Geraniales, and only plastid encoded genes show significant acceleration of *dS* in Geraniales. Detection of a correlation between *dN* of DNA RRR genes and genome complexity support the hypothesis that the unusual changes in plastid genome complexity in Geraniaceae may be caused by the dysfunction of DNA RRR system. Furthermore, the significant acceleration of *dN* of NUCP genes is a possible explanation for the observed correlation between NUCP and genome complexity, and the acceleration of *dN* of NUCP genes could be caused by elevation of plastid genome complexity or other nuclear features.

Chapter 2: Comparative analyses of two Geraniaceae transcriptomes using next-generation sequencing

BACKGROUND

Four remarkable evolutionary phenomena are associated with organellar genomes of Geraniaceae. First, mitochondrial genomes show multiple, major shifts in rates of synonymous substitutions, especially in the genus *Pelargonium* (Parkinson et al., 2005; Bakker et al., 2006). Rate fluctuations of such magnitude have been documented in only two other plant lineages, *Plantago* (Cho et al., 2004) and *Silene* (Mower et al., 2007; Sloan et al., 2008, 2009). Second, mitochondrial genomes have experienced extensive loss of genes and sites of RNA editing. At least 12 putative gene losses have been documented in *Erodium* (Adams et al., 2002), and mitochondrial genes sequenced from *Pelargonium* x *hortorum* had a drastic reduction in predicted or verified RNA editing sites compared to all other angiosperms examined (Parkinson et al., 2005). Third, genome-wide comparisons of nucleotide substitutions in plastid DNA indicated rapid rate acceleration in genes encoding ribosomal proteins, RNA polymerase, and ATP synthase subunits in some lineages. In the case of RNA polymerase genes there was evidence for positive selection (Guisinger et al., 2008; Weng et al., 2012). Fourth, plastid genomes of Geraniaceae are the most highly rearranged of any photosynthetic land plants examined (Palmer et al., 1987; Chumley et al., 2006; Blazier et al., 2011; Guisinger et al., 2011). Multiple and extreme contractions and expansions of the inverted repeat (IR) have resulted in genomes with both the largest IR (74, 571bp, Chumley et al., 2006) as well as the complete loss of this feature (Blazier et al., 2011; Guisinger et al., 2011). Considerable accumulation of dispersed repeats associated with changes in gene order has been documented along with disruption of highly conserved operons and repeated losses and duplications of genes (Guisinger et al., 2011). In *P.* x *hortorum* plastids, these genomic changes have generated several fragmented and highly divergent *rpoA*-like ORFs of questionable functionality (Palmer et al., 1987; Chumley et al., 2006; Guisinger et al., 2008, 2011; Blazier et al., 2011), despite the fact that *rpoA* encodes an essential component of the plastid-encoded RNA polymerase (PEP).

Because nuclear genes supply both organelles with the majority of their proteins, it is likely that the extensive organellar genomic upheaval in Geraniaceae will also influence the evolution of organelle-targeted genes in the nuclear genome. For example, given the drastic reduction of RNA editing in Geraniaceae mitochondrial transcripts, it is reasonable to expect a correlated reduction of nucleus-encoded pentatricopeptide repeat (PPR) proteins, many of which are critical for organellar RNA editing (Kotera et al., 2005; Okuda et al., 2007, 2009; Fujii and Small, 2011). In particular, the uncertain status of the *P.* x *hortorum* plastid-encoded *rpoA* gene is also likely to have nuclear consequences. If this plastid gene is not functional, then a functional copy might have been relocated to the nuclear genome. To date, this is known to have occurred only once in the evolution of land plants, and that was in mosses (Sugiura et al., 2003; Goffinet et al., 2005). Alternatively, it is possible that PEP has become nonfunctional in *P.* x *hortorum*, as observed in the holoparasite *Phelipanche aegyptiaca* (Wickett et al., 2011). In *P. aegyptiaca*, loss of all plastid-encoded PEP components (*rpoA*, *rpoB*, *rpoC1* and *rpoC2*) resulted in the parallel loss of the requisite nucleus-encoded components (sigma factors) that assemble with the plastid encoded proteins to form the core of the PEP holoenzyme (Wickett et al., 2011). In contrast, if the highly divergent plastid *rpoA* gene is still functional in *P.* x *hortorum*, then the typical set of sigma factors should be present in the nuclear genome.

One prerequisite to begin to address the effects of organellar genomic upheaval on the nuclear genome in Geraniaceae, is availability of nuclear sequence information. Transcriptome sequencing provides a tractable proxy for nuclear gene space. The use of next-generation sequencing (NGS) for transcriptome sequencing is widespread because volumes of data can be generated rapidly at a low cost relative to traditional Sanger sequencing (Kircher and Kelso, 2010). The assembly of reads into contigs may be executed using a *de novo* or a reference-based approach (Ward et al., 2012). In studies of non-model organisms, *de novo* assembly is more commonly used due to the absence of a closely related reference (Pepke et al., 2009; Wheat, 2010). A survey of recent transcriptome studies in comparative biology demonstrates that most sequencing projects

are focusing on non-model organisms where little or no genomic data is available (Ward et al., 2012; Der et al., 2011; MS Barker; Angeloni et al., 2011; Hou et al., 2011; Margam et al., 2011; Roberts et al., 2012; Savory et al., 2012). The lack of a reference genome makes the reconstruction and evaluation of the transcriptome assembly challenging. Several issues must be addressed when performing transcriptome sequencing of non-model organisms, including which NGS platform should be employed, how much sequence data is needed to provide a comprehensive transcriptome, which assembler should be utilized, and what tissues should be sampled.

This chapter provides a comprehensive comparison of the transcriptomes of two non-model plant species, *Pelargonium* x *hortorum* and *Geranium maderense*, from the two largest genera of Geraniaceae. There were three primary goals for the initial comparative transcriptome analysis in Geraniaceae: (1) What are the best sequencing platforms and assembly methods for generating a high-quality transcriptome that broadly covers gene space in the absence of a reference genome? (2) Does sequencing from multiple tissue types improve the breadth of transcriptome coverage? (3) Are there any losses of PPR proteins involved in RNA editing and sigma factors associated with PEP in Geraniaceae?

RESULTS

Ribosomal RNA content and Illumina library complexity

To assess the efficiency of ribosomal RNA (rRNA) depletion in Geraniaceae transcriptome libraries, rRNA contigs were identified using rRNA from *Arabidopsis thaliana* as a reference. All Illumina reads (146,690,142 reads for *Geranium maderense* and 148,749,374 reads for *Pelargonium* x *hortorum*) were mapped to rRNA contigs as described in Methods, and 0.7% and 2% of the reads of *G. maderense* and *P.* x *hortorum* were identified as rRNA reads, respectively. Library complexity was analyzed using Picard (http://picard.sourceforge.net/) and rRNA reads were eliminated prior to the analysis. The percentages of unique start sites were 42.7% and 46.1% for *G. maderense* and *P.* x *hortourm*, respectively. The values for rRNA content and library complexity

were comparable to other transcriptome analyses using similar approaches (Tariq et al., 2011; Levin et al., 2010).

Assessment of sequencing platforms and assemblers for transcriptome assembly

To determine the optimal sequencing and assembly strategy, the efficacy of five different assemblers was examined using two initial data sets generated by Roche/454 FLX and Illumina HiSeq 2000 platforms for *P.* x *hortorum*. The Illumina run produced approximately 40 times more sequence data than the 454 run, even though the cost of the 454 data was at least four times more than the Illumina data (Table 2.1). A comparison of basic assembly statistics (Table 2.2) showed that the Trinity assembler outperformed all other platform/software combinations in terms of number of contigs, number of assembled nucleotides, mean and maximum contig length, and N50. More generally, the Illumina assemblers consistently outperformed the 454 assemblers, although the MIRA and Newbler 454 assemblers produced longer maximal contigs than SOAPdenovo and SOAPtrans. To determine the amount of usable protein sequence information generated by each assembler, the assemblies were translated as described in Methods and compared (Table 2.3). Again, the Illumina assemblers outperformed the 454 assemblers in all metrics, with the Trinity assembler providing the most amino acids with the longest mean and maximal sequences. The length distribution of assembled nucleotides and translated amino acids further confirms that Trinity outperformed SOAPdenovo and SOAPtrans, and all three Illumina assemblers outperformed the 454 assemblers (Figure 2.1).

Two important considerations in assembly analysis are the breadth of gene space coverage and the degree of coverage fragmentation. A good assembler should generate high-quality assemblies that contain as many reference transcripts as possible, and each reference transcript should be covered as completely as possible with a single long contig rather than a combination of several short contigs. To assess assembly coverage and fragmentation, two published data bases were used, 357 ultra-conserved ortholog (UCO) coding sequences (Kozik et al., 2008) from *Arabidopsis* and 959 single copy nuclear genes shared between *Arabidopsis*, *Oryza*, *Populus*, and *Vitis* (Duarte et al., 2010).

Trinity and SOAPtrans outperformed all other assemblers in terms of the percentage of reference genes identified, completeness of coverage (i.e., fraction of reference gene coverage by one or more contigs), and contiguity of coverage (i.e., fraction of reference gene coverage by a single long contig), with Trinity performance slightly better than SOAPtrans at higher thresholds (Figures 2.2-2.3).

To examine whether the superior performance of Trinity and SOAPtrans was due to the much larger amount (40 times) of Illumina data than 454 data, the Illumina assemblers were re-anlayzed using a data set containing 1/40th of the Illumina reads (Appendix Figure 2.1). In terms of contiguity and completeness, the performance of Trinity using the reduced Illumina data set remained superior to the 454 programs (Newbler, MIRA) that used the entire 454 data sets. In contrast, the performance of SOAPdenovo and SOAPtrans were noticeably worse with the reduced Illumina data set than with the full data set, producing results that were generally worse than the original 454 assemblies.

Effect of sequencing depth on assembly coverage breadth and fragmentation

To determine how much sequence data is needed to assemble a high-quality transcriptome with broad coverage, 146,690,142 reads for *G. maderense* and 148,749,374 reads for *P.* x *hortorum* were generated on the Illumina HiSeq 2000 platform assembled using Trinity with different increments of reads from 5% to 100% of the total. While the number of contigs assembled continued to increase with increasing numbers of reads (Figure 2.4A), the percentage of reference genes recovered and their contiguity and completeness plateaued at approximately 40% of the total reads (Figure 2.4B-D). Including the remaining 60% of the reads increased contiguity and completeness by only 1% to 2% (Figure 2.4B-C). Although there were more translated contigs of *G. maderense* than *P.* x *hortorum*, the contiguity and completeness of both species were very similar.

Although increasing the number of reads beyond 10% contributed little to finding novel hits to the *Arabidopsis* database, increasing the amount of data helped extend the

existing contigs and generate longer alignments to reference genes. To evaluate this, the contiguity of all contigs relative to the two published databases was calculated at different contiguity thresholds up to 100% (Figure 2.5). The inclusion of more reads generated assemblies with higher contiguity, especially when contiguity thresholds were greater than 50%. To allow for the high level of sequence divergence between Geraniaceae and *Arabidopsis*, the number of contigs that had contiguity thresholds ≥ 80% was calculated. When 100% of the reads were used 4185 contigs and 4494 contigs were found in *G. maderense* and *P.* x *hortorum*, respectively. Reducing the read input to 40% reduced contiguity values by 7% (4163/4494) in *G. maderense* and 11% (3731/4185) in *P.* x *hortorum*.

Functional assessment of Geraniaceae nuclear transcriptomes

The assemblies generated using 100% of the reads for both Geraniaceae species were used for functional annotation. Assemblies were first aligned against the NCBI nr database and the alignment results were used to generate the gene ontology (GO) terms. Of the 114,762 contigs in *P.* x *hortorum*, 56,283 (49%) had blast hits; 42,506 (37%) were annotated and 222,765 GO terms were retrieved (Table 2.4). Of the 119,109 contigs in *G. maderense*, 76,332 (64%) had blast hits; 58,461 (49%) were annotated (Table 2.4) and 311,108 GO terms were retrieved. The annotation files are shown in Supplemental Data File 2.1. The distribution of gene ontology annotations was examined using GO-slim (plant) ontology to compare the transcriptomes of *G. maderense* and *P.* x *hortorum*. Although the number of annotated contigs differed substantially between the two transcriptomes (Table 2.4), the proportion of annotated contigs in all categories with >1% representation within the categories Cellular Component, Molecular Function, and Biological Process were very similar (Figure 2.6). This similarity persists even though only emergent leaves were sampled for *G. maderense* versus four tissue types (emergent and expanded leaves, roots and flowers) for *P.* x *hortorum*.

To more directly address the question whether sequencing from multiple tissue types improves the breadth of transcriptome coverage, orthologous genes between *G.*

maderense and *A. thaliana* and between *P.* x *hortorum* and *A. thaliana* were identified. Of the 35,386 protein sequences from *A. thaliana*, the *G. maderense* assembly had homologs to 11,131 sequences and the *P.* x *hortorum* assembly had homologs to 11,583 sequences. The comparable numbers of orthologous genes found for the two Geraniaceae species indicated that there was little improvement on the breadth of transcriptome coverage by sequencing from multiple tissue types (1 versus 4 tissues for *G. maderense* and *P.* x *hortorum*, respectively).

Identification of selected organelle targeted genes

Pentatricopeptide repeat proteins (PPRs) are a large family of RNA binding proteins encoded by over 400 genes in angiosperms; most are organelle targeted and involved in regulating organelle gene expression. The transcriptomes of *P.* x *hortorum* and *G. maderense* were annotated using 429 *Arabidopsis* PPR sequences as a reference database (Table 2.5). The overall number of PPR genes varied considerably between the two Geraniaceae and *Arabidopsis*, with PPR gene number reduced in *P.* x *hortorum*. The numbers of P class PPR genes were found to be similar in all three species, whereas many fewer PLS class genes were found in the Geraniaceae, especially in *P.* x *hortorum*.

Sigma factors are nuclear encoded, plastid targeted proteins that assemble with four plastid encoded proteins (*rpoA*, *rpoB*, *rpoC1* and *rpoC2*) to form the core of the PEP holoenzyme. At least one copy of each of the six *Arabidopsis* sigma factors was detected in both the *G. maderense* and *P.* x *hortourm* transcriptomes (Table 2.5). The nucleotide and amino acid sequence identities between *Arabidopsis*/*Geranium* and *Arabidopsis*/*Pelargonium* for all six sigma factors were very similar (Table 2.6). The four contigs from *G. maderense* that aligned to sigma factor 2 were similar to each other in nucleotide sequence identity (87%), suggesting that they may represent variant copies of the same gene. Two of the three contigs from *G. maderense* that aligned to sigma factor 5 were very similar to each other but less so to the third contig (98% versus 71% nucleotide sequence identity). Sigma factors 2 and 6 were each represented by two *P.* x *hortorum* contigs, however only one of the contigs for each sigma factor appeared

functional having start/stop codons at the 5' and 3' ends and lacking internal stop codons. Further experiments are needed to determine if the copies with internal stop codons are pseudogenes or assembly artifacts.

DISCUSSION

Strategies for de novo assembly of transcriptomes

The use of NGS platforms is widespread and is applied in many research fields as volumes of data can be generated rapidly at a low cost relative to traditional Sanger sequencing (Kircher and Kelso, 2010). RNA-seq, one popular NGS application, provides an efficient and cost-effective way of obtaining transcriptome data. There are a number of platforms available for generating NGS data (Harismendy et al., 2009; Metzker, 2010). Currently among the most popular are the Roche/454 FLX (http://www.roche.com) and the Illumina HiSeq 2000 (formerly Solexa; http://www.illumina.com) platforms. The Roche/454 FLX system is advantageous when longer reads are important (average read length 700 bp), whereas the Illumina system provides deeper sequencing coverage at a reduced cost per base, albeit with shorter read length (average length 100 bp).

For each platform various assemblers have emerged but during the past several years Roche 454 sequencing and the platform-specific assembler Newbler has been the most common approach for de novo assembly of transcriptome data (Weber et al., 2007; Novaes et al., 2008; Vega-Arreguín et al., 2009; Wall et al., 2009; Cantacessi et al., 2010). This may be attributed to the idea that longer reads are more likely to overcome the specific challenges of de novo transcriptome assembly. Illumina sequencing has been used mainly when a related organism's genome was available for reference-based assembly (Nagalakshmi et al., 2008; Rosenkranz et al., 2008), although due to recently increased read length it is becoming more common for use in de novo assembly as well (Birol et al., 2009; Wang et al., 2010). Several recent studies have compared the performance of different sequencing platforms and assembly methods (Kumar and Blaxter, 2010; Feldmeyer et al., 2011; Bräutigam et al., 2011) but none of these comparisons evaluated the level of completeness or contiguity of their assemblies. Nor

was the performance of the assemblers evaluated without known genome information, which is the situation for any project on non-model organisms.

Our comparisons of sequencing platforms and assemblers for the Geraniaceae clearly indicated that the Illumina platform with Trinity assembly delivered the best performance in assembling a more complete transcriptome in the absence of a reference genome. The Illumina assemblers (Trinity, SOAPdenovo, SOAPtrans) generated more contigs containing a greater total number of bases than the Roche/454 FLX assemblers (Newbler, MIRA). While the MIRA assembly generated many more long contigs (>6 kb) than SOAPdenovo, the Trinity assembly out-performed all others in delivering long contigs, suggesting that the Trinity assembly contained more useful information than any of the other assemblies analyzed. While the Roche/454 FLX assemblies and the Illumina SOAPdenovo assembly produced similar results with regard to completeness and contiguity, the Illumina Trinity and SOAPtrans assemblies obtained much higher values for both parameters indicating that these assemblies comprise many more nearly complete transcripts (Figures 2.2-2.3).

Functional annotation of Geraniaceae transcriptomes

A total of 58,461 (49%) and 42,506 (37%) contigs were annotated from *G. maderense* and *P.* x *hortorum*, respectively. The low percentage of annotated contigs is most likely due to the large number of total contigs assembled. The number of aligned and annotated contigs is comparable to nine other recently published transcriptomes (Angeloni et al., 2011; Garg et al., 2011; Kaur et al., 2011; Logacheva et al., 2011; Natarajan and Parani, 2011; Shi et al., 2011; Wenping et al., 2011; Ward et al., 2012). The number of annotated contigs in assemblies from both Geraniaceae species was very similar for the three major categories cellular component, molecular function, and biological process (Figure 2.9). This is encouraging since different tissues were sampled for the two species; only one tissue, emergent leaves for *Geranium* and four tissues, emergent leaves, expanded leaves, roots and flowers for *Pelargonium*. Particularly noteworthy is the detection of genes associated with flower and embryo development and

pollen-pistil interaction since flowers were not sampled for *Geranium*. Overall, this comparison indicates that there is no marked improvement in transcriptome breadth of coverage when sampling four tissues compared to only emergent leaves.

PPR proteins and sigma factors in Geraniaceae

PPRs are a large family of RNA binding proteins encoded by over 450 genes in sequenced angiosperms. Most are organelle targeted and involved in regulating organelle gene expression (Schmitz-Linneweber and Small, 2008). Of the two classes (P and PLS) within the PPR family, those from PLS class (E and DYW subclasses) have been reported to be involved in RNA editing (Kotera et al., 2005; Okuda et al., 2007; Chateigner-Boutin et al., 2008; Cai et al., 2009, 66; Hammani et al., 2009; Robbins et al., 2009; Yu et al., 2009, 2; Zhou et al., 2009; Okuda et al., 2009; Tseng et al., 2010). Previous studies have demonstrated correlated evolution of PLS genes and RNA editing sites in plants (Fujii and Small, 2011; Hayes et al., 2012). Consistent with these results, a reduction in PLS genes (Table 2.5) in Geraniaceae was detected, where reduced editing frequency was previously demonstrated (Parkinson et al., 2005). The reduced editing frequency and reduced PPR content in Geraniaceae is especially intriguing with respect to the increased mitochondrial substitution rate in this family. Although an inverse correlation between editing frequency and substitution rate has been noted previously in Geraniaceae and other taxa (Parkinson et al., 2005; Lynch et al., 2006; Cuenca et al., 2010; Sloan et al., 2010), the finding that PPR gene content is also reduced in Geraniaceae indicates that this family is ideally suited for future studies assessing the evolutionary dynamics of editing frequency, PPR content, and mitochondrial substitution rates.

One long-standing question regarding the plastid genomes in Geraniaceae is the putative loss of the *rpoA* gene from *P.* x *hortorum* (Palmer et al., 1990a, 1990b; Downie et al., 1994). The complete plastid genome sequence of this species revealed several *rpoA*-like open reading frames (ORFs) that are highly divergent relative to *rpoA* genes in other angiosperms or even other Geraniaceae (Chumley et al., 2006; Guisinger et al.,

2011). Two alternative explanations were suggested for these observations: (1) a copy of the gene in the nucleus had gained functionality; or (2) at least one of the highly divergent *rpoA*-like ORFs remains functional. Extensive evolutionary rate comparisons of plastid genes across the Geraniaceae revealed that the other three PEP subunits (*rpoB*, *rpoC1*, *rpoC2*) have significantly elevated nucleotide substitution rates and have likely experienced positive selection (Guisinger et al., 2008; Weng et al., 2012). Despite exhaustive searching of the nuclear transcriptome of *P.* x *hortorum* no copy of the *rpoA* gene was detected. However, intact copies of all six sigma factors, which are required for PEP to function (Lysenko, 2007), were identified in the transcriptome. The holoparasite *Phelipanche aegyptiaca* lacks a functional PEP and mining unigene files published in a recent transciptomic study of parasitic plants (Wickett et al., 2011) failed to uncover a single sigma factor suggesting that in species where PEP sequences are lost from the plastid the requisite sigma factors are also absent from the nuclear transcriptome. The identification of all six sigma factors in the *P.* x *hortorum* transcriptome supports the likelihood that PEP is active in *P.* x *hortorum* plastids.

CONCLUSIONS

With the widespread application of NGS techniques, the ability to process and analyze massive quantities of sequence data in a timely manner becomes imperative to a successful project. Regardless of the goals of a particular project, it is desirable to obtain data that are as accurate and complete as possible in a way that is cost effective as well as timely. In this study a cross-platform comparison of de novo transcriptome assembly was conducted using representative species from the two largest genera of Geraniaceae, *G. maderense* and *P.* x *hortorum*. As no reference genome is available for Geraniaceae, or any of its close relatives, this approach represents a truly de novo assembly allowing evaluation of efficacy among the platforms/assemblers that more closely resembles current NGS research. The assembly of Illumina HiSeq 2000 reads with Trinity or SOAPtrans was highly effective in reconstructing, as completely as currently feasible, the protein-coding transcripts of Geraniaceae. As for the differences between the two

assemblers, Trinity generated slightly more single contiguous contigs and reconstructed more reference genes with a combination of multiple contigs, while SOAPtrans ran much faster than Trinity. These differences in contiguity and completeness became more obvious with the reduced set of input data (1/40th in this case). These findings recommend the Illumina platform with Trinity assembly to obtain the most complete gene coverage by a single contig, especially when a small amount of reads are available. In instances where a large amount of data is available and there are limited computational resources, Illumina SOAPtrans assembly may be preferred as it generated a relatively complete assembly much more quickly than Trinity. Furthermore, evaluation of the amount of Illumina sequence data required for generating a complete transcriptome is approximately 60 million reads.

Geraniaceae organelle genomes have been shown to exhibit a number of unusual features relative to other angiosperms, including highly accelerated rates of nucleotide substitutions in both mitochondrial and plastid genes (Parkinson et al., 2005; Guisinger et al., 2008; Weng et al., 2012), reduced RNA editing in mitochondrial genomes (Parkinson et al., 2005) and highly rearranged plastid genomes (Palmer et al., 1987; Chumley et al., 2006; Chris Blazier et al., 2011; Guisinger et al., 2011). This comparative transcriptome analysis of *G. maderense* and *P.* x *hortorum* detected a reduction in PPR proteins associated with RNA editing, which corresponds with reduced RNA editing in the mitochondria. Examination of nuclear encoded, plastid targeted sigma factors required for PEP function supports the hypothesis that PEP is active in *P.* x *hortorum* plastids, possibly incorporating the product of at least one of the highly divergent *rpoA*-like ORFs in the plastid genome.

MATERIALS AND METHODS

RNA isolation

Plant tissues were collected from live plants grown in the University of Texas (UT) greenhouse and frozen in liquid nitrogen for two species from different genera of Geraniaceae, *Geranium maderense* and *Pelargonium* x *hortorum* cv ringo white. For

Pelargonium leaf and inflorescence samples were collected. Leaves were of two developmental stages, newly emerged and fully expanded. Entire inflorescences were harvested prior to anthesis. Root samples of *P.* x *hortorum* were harvested from specimens grown aseptically in agar media. For Geranium, only emergent leaves were collected. Total RNA was isolated separately from each sample type by grinding in liquid nitrogen followed by 30 min incubation at 65 oC 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 oC and total RNA was pelleted by centrifugation at 17, 000 x g for 20 min at 4 oC. 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. For Pelargonium, the four tissue types were pooled in equimolar ratio. All RNAs were treated with DNase I (Fermentas, Glen Burnie MD, USA) according to the product protocol. 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 2.5 volumes of cold absolute ethanol for 20 min at -80 oC. 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 and shipped on dry ice to the Beijing Genomics Institute (BGI) in Hong Kong or delivered to the Genome Sequencing Analysis Facility (GSAF) at UT. Confirmation of sample quality and concentration was conducted at each facility using the Agilent 2100 Bioanalyzer instrument (Agilent Technologies, Santa Clara CA, USA).

Illumina sequencing

Sample preparation for Illumina sequencing was performed at BGI according to Illumina's protocol (Part # 1004898 Rev. D). Total RNA was treated with the Ribo-Zero™ rRNA Removal Kit (Epicentre Biotechnolgies, Madison WI, USA) prior to fragmentation and priming with random hexamers for first strand cDNA synthesis using SuperScript® III Reverse Transcriptase (Invitrogen, Beijing, China). Second strand cDNA synthesis was carried out using RNase H (Invitrogen) and DNA polymerase I (New England BioLabs, Beijing, China). The resulting cDNA fragments were purified with QIAQuick® PCR extraction kit (Qiagen, Shanghai, China) and normalized with Duplex-Specific thermostable nuclease (DSN) enzyme from Kamchatka crab (Evrogen, Moscow, Russia) according to the protocol outlined by Invitrogen (Part # 15014673 Rev. C). End repair and adenylation of the normalized cDNA library was followed by ligation to the paired-end (PE) sequencing adapters. Following gel electrophoresis for size selection (180-220 bp) the library was PCR amplified for sequencing using the Illumina HiSeq™ 2000. The PE library was sequenced for 101 bp.

Roche/454 FLX sequencing

The method for cDNA library construction and normalization was based on that of Meyer et al. (Meyer et al., 2009). Briefly, total RNA was reverse-transcribed using oligo-dT coupled to a PCR-suppression primer. The reverse complement of this primer was incorporated at the 3' end of the first-strand cDNA using the template switching capability of the SuperScript II Reverse Transcriptase (Invitrogen). Duplex-specific nuclease was added to digest the abundant double-stranded cDNA. After purification, PCR was performed, and the product was purified and sheared by nebulization. The fragmented DNA was then end-repaired and ligated to Roche Rapid library adaptors using the NEBNext® Quick DNA Sample Prep Master Mix Set 2 and NEBNext® DNA Sample Prep Master Mix Set 2 (New England BioLabs). Final library size and concentration were measured on the Agilent BioAnalyzer and by qPCR before sequencing on the Roche/454 FLX sequencer.

Read pre-processing

Raw reads were preprocessed to eliminate contaminant and low quality sequences. Filtering of Illumina HiSeq 2000 reads included the removal of low quality bases, reads where (poly) adenosine constitutes more than 6% of bases, and reads containing specialized features such as adaptors and other artifacts arising from library construction. Roche/454 FLX reads were preprocessed by removing reads shorter than 50 bp and reads with artificial sequences based on a vector reference file. The complete data set is available at NCBI Sequence Read Archive (Accession numbers SRA059171 for Geranium and SRA053016.1 for Pelargonium).

Ribosomal RNA content and Illumina library complexity

Ribosomal RNA (rRNA) contigs were identified using reciprocal blast of rRNA from *Arabidopsis* (5.8S, 18S and 25S in nucleus, 5S, 16S and 23S in chloroplast, and 5S, 18S and 26S in mitochondria) as reference. The rRNA sequences from *Arabidopsis* were downloaded from TAIR (Lamesch et al., 2012). Ribosomal RNA reads were removed prior to the library complexity analysis. Due to a lack of nuclear genome sequence, the remaining reads were mapped back to the whole transcriptome data using bowtie2 (Langmead and Salzberg, 2012). The mapping results were sorted using samtools (Li et al., 2009a) and then analyzed with MarkDuplicates module of Picard (http://picard.sourceforge.net/).

Assembly

Transcritpome assemblies were initially performed on Pelargonium using a variety of assemblers to compare the efficacy of different platforms and assemblers. After these initial comparisons, all subsequent assembles were performed on both Geranium and Pelargonium using Trinity and Illumina data. For assembly of clean Illumina reads, Trinity (Grabherr et al., 2011), SOAPdenovo and SOAPtrans (http://soap.genomics.org.cn/SOAPdenovo-Trans.html) (Li et al., 2008, 2009b) were used. Trinity, released on 2011-08-20 (http://sourceforge.net/projects/trinityrnaseq/), was run with parameters "--seqType fq --CPU 10 --paired_fragment_length 200 --

run_butterfly" on a 24-core 3.33GHz linux work station with 1TB memory at the Texas Advanced Computing Center (TACC, http://www.tacc.utexas.edu/). The assembly was split into three steps according to the provided script trinity.pl released with the software. The split scripts run the corresponding three steps in Trinity: inchworm, chrysalis, and butterfly. The parameters were the same for each step, and each step picked up the previous step's output as input and processed it. The scripts will be provided by JZ upon request. The SOAPtrans assembly was run with the parameters "kmer=61, max_rd_length=100, avg_ins=200" on the same server as that of Trinity. For SOAPtrans kmer lengths from 23bp to 81bp were explored; 61bp was selected because it generated the best contiguity compared with other kmer values. The SOAPdenovo assembly was done at BGI on a 48-core 2.67GHz Linux workstation with 50GB memory with parameters "Kmer=41, insert size=200, overlap threshold=50" for assembly, and "Kmer+1" to fill the gaps. The generated fasta file was postprocessed by BGI to remove the sequences shorter than 150 bp. Assembly of Roche/454 FLX utilized MIRA (Chevreux et al., 2004) and Newbler (Margulies et al., 2005). MIRA 3.4.0 for a 64-bit linux system (http://sourceforge.net/projects/mira-assembler/files/MIRA/stable/) was released on 2011-08-21. MIRA was run with parameters "--job=denovo, est, accurate, 454 --fasta 454_SETTINGS" on a 12-core 3.33GHz linux work station with 24GB memory at TACC. Newbler 2.6 accompanies the Roche/454 FLX platform and assembly was conducted at UT GSAF on 24-core 2.40GHz linux work station with 64GB memory using the parameters "runAssembly -cpu 8 -urt -cdna –vt vector.fa".

Comparative analysis of assemblies

Trinity, SOAPdenovo and SOAPtrans assembly output comprised a single contig file each and these were used in the analyses. Unpadded fasta files were selected from the MIRA output and the isotig file was selected from the Newbler output for use in analyses.

The initial assembly quality was evaluated using the following metrics: number of assembled contigs, maximum, minimum and mean contig length, N50 and redundancy.

Initial assembly statistics and contig length distribution analysis was done by custom perl scripts and MATLAB version R2011b. Contig clustering and removal of redundant contig sequences was performed using CD-HIT (Li and Godzik, 2006). CD-HIT version 4.5.4 (downloaded from http://code.google.com/p/cdhit/downloads/list) was executed using parameters "cd-hit -c 1.0 -n 5 -T 12" for cDNA sequences and "cd-hit-est -c 1.0 -n 10 -T 12" for protein sequences. Redundancy was calculated from the difference between the number of contigs before and after clustering. Maximum, minimum, and mean contig length, N50 and total bases were calculated from the contigs after clustering and removal of those contigs < 200 bp.

The assemblies were aligned to two published reference databases: 357 ultra-conserved ortholog (UCO) coding sequence (Kozik et al., 2008) from *Arabidopsis* (sequences available at: http://compgenomics.ucdavis.edu/compositae_reference.php), and a list of 959 single copy nuclear genes shared between *Arabidopsis*, *Oryza*, *Populus*, and *Vitis* (Duarte et al., 2010) using BLASTX with evalue of 1 E-10. Contig alignment to the reference databases utilized the standalone BLAST+ (Camacho et al., 2009) program for 64-bit linux system (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/). The parameters for BLAST+ DNA alignment were "blastn -task blastn -evalue 1 E-10 -word_size 11 -outfmt 6 -num_threads 12". Parameters for protein alignment were "blastp -task blastp -num_threads 12 -outfmt 6". For blastp, two different e values were used, 1 E-10 and 1 E-20, in order to address the generality of the results. Multiple sequence alignment was done by muscle (Edgar, 2004). Muscle for 64-bit linux system (http://www.drive5.com/muscle/downloads.htm) was used with default parameters.

The local reference database for identifying the open reading frames contained four proteomes downloaded from Phytozyme (http://www.phytozome.net/search.php): *Citrus clementina*, *C. sinensis*, *Eucalyptus grandis* and *Arabidopsis thaliana*. Contigs were translated by alignment to the local database using blastx to identify open reading frames. The blastx parameter was "blastx -evalue 1e-6 -max_target_seqs 1 -num_threads 48 –outfmt'6 std qframe". The reading frame parameter was added to the output in order

to facilitate the following analysis. The aligned regions of contigs were translated, extracted, and then extended by translating the contigs in both directions according to standard codon usage until a stop codon was encountered. The translated contigs were clustered again using CD-HIT at a threshold of 100% and all other parameters used the default settings. Two parameters, contiguity and completeness as described by Martin and Wang (Martin and Wang, 2011) were used to evaluate the alignment results. Briefly, contiguity is defined as the percentage of the reference transcripts covered at some arbitrary coverage threshold by a single longest contig. Completeness is defined as the percentage of the reference transcripts covered at a threshold by multiple assembled contigs (Box 1 in Martin and Wang, 2011). In this study a range of thresholds up to 100% was evaluated, and 80% was selected as the threshold for both contiguity and completeness calculations. Both parameters were calculated with protein sequence alignment, and the alignment results were analyzed using custom perl scripts available from JZ upon request.

Evaluation of assemblies with different proportion of reads

To assess how much data (number of reads) is needed to construct the complete transcriptome, different proportions of sequencing data ranging from 5% to 100% were extracted for both species. The extracted reads were assembled with Trinity using the parameters described above. Extraction and assembly were repeated three times for each proportion except 100%, and the assembly statistics (contig number, contiguity and etc.) were averaged.

Basic statistics and assembly parameters such as contiguity and completeness were calculated using the same local database described above. To determine how well the assemblies cover a complete transcriptome, the custom *Arabidopsis* protein database was constructed by extracting all *Arabidopsis* proteins from Uniprot/Swissprot database (Boeckmann et al., 2003), and protein sequences with name "hypothetical" or "predicted" were discarded. The assemblies were aligned with the database using BLASTX with an E-value of 1 E-10.

Orthologous genes identification

Orthologous genes between transcriptomes of *G. maderense*, *P.* x *hortorum* and *A. thaliana* were identified with reciprocal blast with parameters "blastp -task blastp -num_threads 12 -max_target_seqs 1 -evalue 1e-10 -outfmt='6 std qlen slen". Blast results were analyzed with custom perl scripts.

Functional annotation

The assemblies were aligned with the NCBI nr database using BLASTX with an E-value of 1 E-6 and taking the best 10 hits for annotation. The blast results were used to annotate each sequence with gene ontology (GO) terms using Blast2GO (Conesa et al., 2005; Conesa and Götz, 2008; Gotz et al., 2008). To improve the efficiency of annotation, local blast2go database was downloaded (http://www.blast2go.com/b2glaunch/resources/35-localb2gdb). GO terms were mapped to the reduced GO-slim (plant) ontology to get a broader functional representation of the transcriptome.

Identification of selected organelle targeted genes

PPR proteins were searched for using HMMER (Eddy, 1998; Lurin et al., 2004) with previously established PPR motif alignment files (Small and Peeters, 2000). Transcript sequences with more than one PPR motif were considered PPR genes. Sigma factor protein sequences from *Arabidopsis* were downloaded from TAIR (Lamesch et al., 2012) and used as reference. Sigma factor structure and conserved domain information were obtained from previous studies (Helmann and Chamberlin, 1988; Isono et al., 1997; Hakimi et al., 2000). Putative transit peptides were predicted with targetP (Nielsen et al., 1997; Emanuelsson et al., 2000). Orthologs from two transcriptomes of *G. maderense* and *P.* x *hortorum* were identified by reciprocal blast at E-value 1 E-10.

Table 2.1. The *Pelargonium* x *hortorum* transcriptome dataset read statistics.

Technology	Number of trimmed reads	Number of trimmed bases	Max read length	Min read length
454	472,268	119,394,317	828	50
Illumina	46,475,742	4,674,574,200	100	100

Table 2.2. Basic assembly statistics for the *Pelargonium* x *hortorum* transcriptome.

	Newbler	MIRA	SOAPdenovo	Trinity	SOAPtrans
Number of nonredundant contigs	28,182	30,947	67,028	67,614	62,470
Total bases	12,972,883	15,326,277	39,088,184	58,210,111	33,057,051
Max contig length	8,147	12,431	6,616	16,017	7,574
Mean contig length	460	495	583	860	529
N50	478	525	782	1,319	678

Table 2.3. Translated contig statistics for *Pelargonium* x *hortorum*.

	Newbler	MIRA	SOAPdenovo	Trinity	SOAPtrans
Number of translated contigs	18,525	19,279	42,907	39,742	44,379
Total amino acids (AA)	2,413,770	2,575,430	8,363,275	11,058,408	7,697,127
Max translated AA length	902	1,086	1,902	2,618	2,520
Mean translated AA length	130	133	195	278	173
N50	145	145	278	387	230

Table 2.4. Statistics of transcriptome annotations for *Geranium maderense* (GMR) and *Pelargonium* x *hortorum* (PHR).

	GMR	**PHR**
Total contigs	119,217	114,762
Aligned contigs	76,332	56,283
Annotated contigs	58,461	42,506
Assigned GO terms	311,108	222,765
Assigned EC	25,533	19,354
Contigs with EC	20,337	15,252

GO = Gene Ontology; EC = Enzyme Code

Table 2.5. PPR protein[a] and sigma factor[b] distribution.

	Arabidopsis thaliana	*Geranium maderense*	*Pelargonium* x *hortorum*
PPR proteins	429	523	315
P class	238	387	262
PLS-E class	105	96	22
PLS-DYW class	86	40	31
Sigma factors	6	10	6
Sig 1	1	1	1
Sig 2	1	4	1
Sig 3	1	1	1
Sig 4	1	1	1
Sig 5	1	3	1
Sig 6	1	1	1

[a]PPR protein data of *Arabidopsis* are from Small and Peeters (Small and Peeters, 2000). The PPR class represents the number contigs longer than 150 aa, which is the minimum length of PPR proteins identified in *Arabidopsis*. [b]The number of total contigs and the number intact contigs aligned to the reference sigma factors are shown. Intact contigs are those with start/stop codons on 5' and 3' ends, and without any internal stop codons.

Table 2.6. Sequence identities between intact contigs in Geraniaceae and *Arabidopsis thaliana* sigma factors.

Arabidopsis thaliana	***Geranium maderense***		***Pelargonium* x *hortorum***	
Sequence identity (%)[a]	**nucleotide**	**amino acid**	**nucleotide**	**amino acid**
Sig 1	64.5	52.4	61.1	50.4
Sig 2	62.6	48.9	62.5	47.1
Sig 3	57.7	39.9	58.6	43.3
Sig 4	58.6	42.6	58.3	43.5
Sig 5	64.3	54.1	65.8	55.0
Sig 6	58.6	41.1	59.6	42.4

[a]In cases where there is more than one intact contig for a sigma factor, the one with highest sequence identity to *Arabidopsis* was selected for comparison.

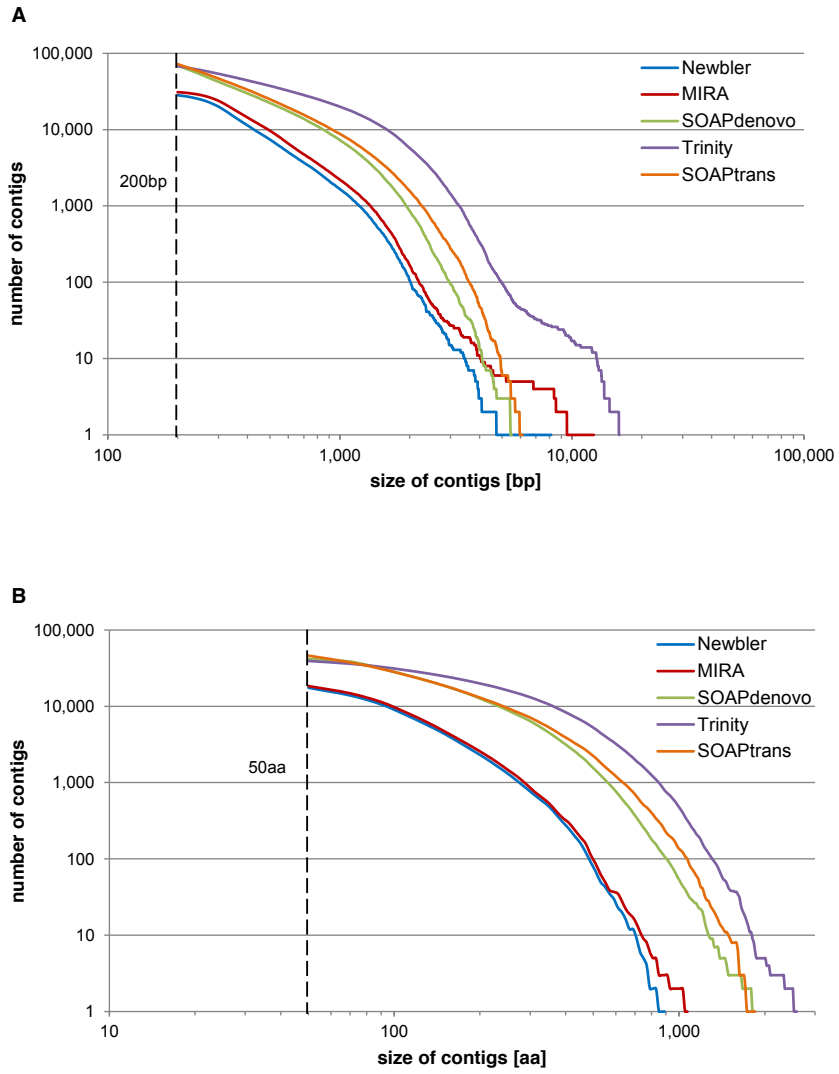


Figure 2.1. Contig length distribution.

The vertical dashed line shows (A) the arbitrary cutoff of 200 base pairs (bp) or (B) 50 amino acids (aa).

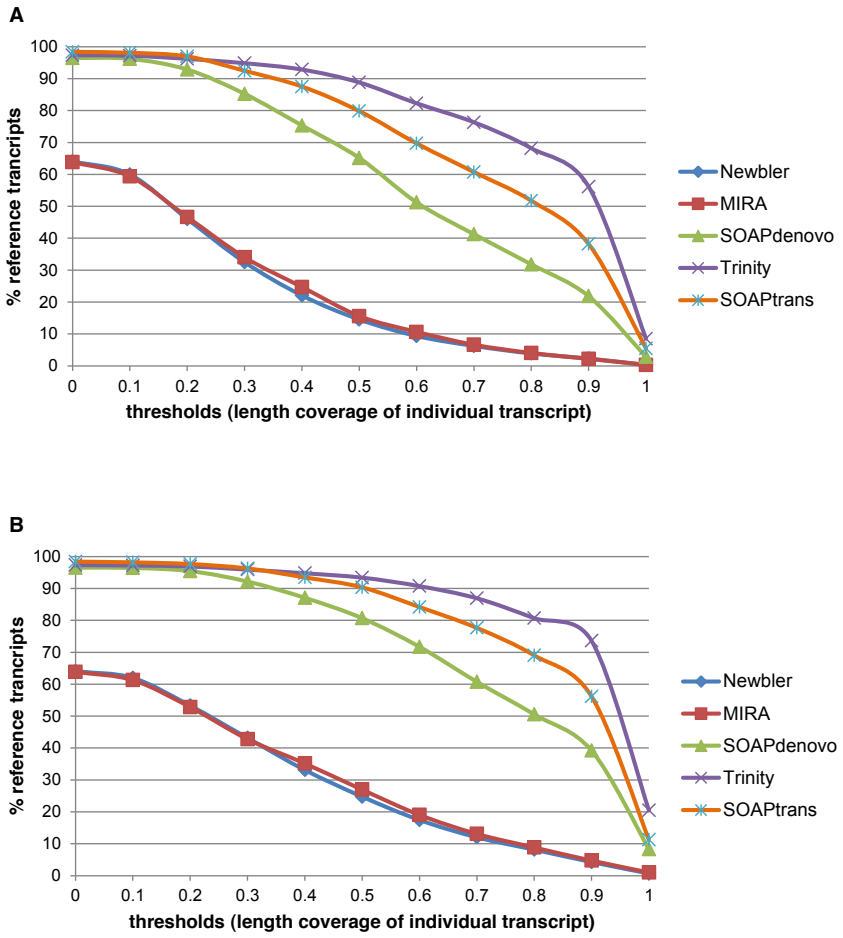


Figure 2.2. Contigstaitistics of different assemblers at different thresholds.

Contiguity (A) and completeness (B) of different assemblers at different thresholds. The assemblies were aligned with two published reference data bases: 357 ultra-conserved ortholog (UCO) coding sequence (Kozik et al., 2008) and 959 single copy nuclear genes (Duarte et al., 2010).

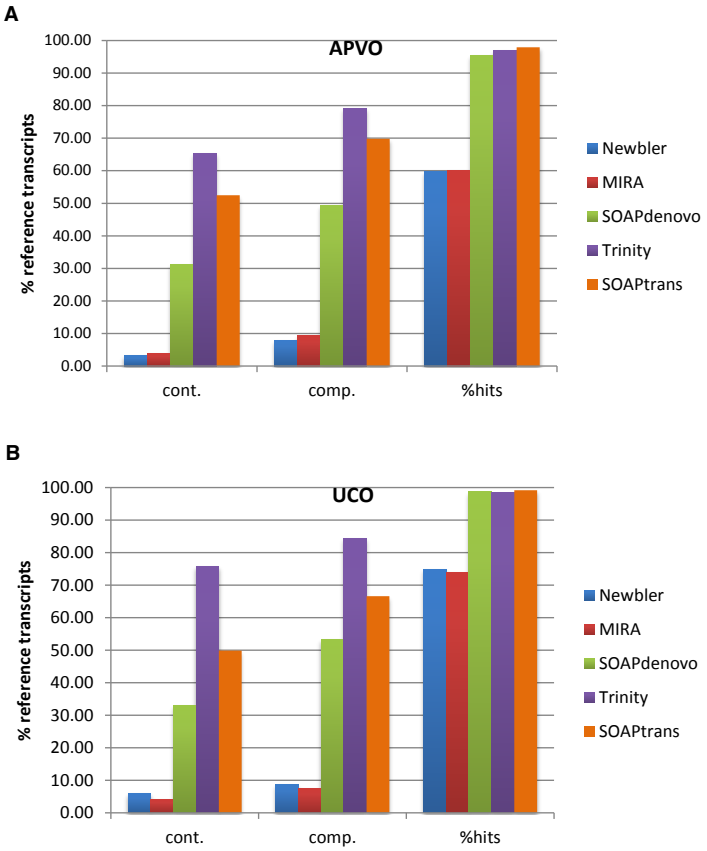


Figure 2.3. Completeness and contiguity results at threshold 80% using two published reference protein sets.

Data sets: 357 ultra-conserved ortholog (UCO) coding sequence (Kozik et al., 2008) and 959 single copy nuclear genes (Duarte et al., 2010). Cont = contiguity, comp = completeness, % hits = percentage of hits in reference transcriptome.

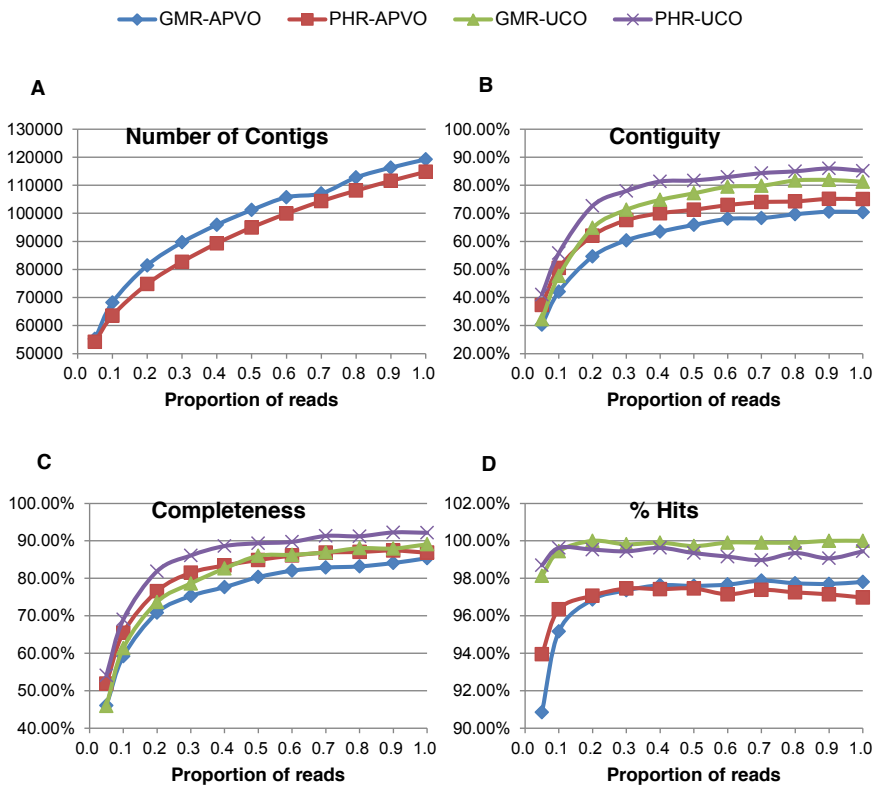


Figure 2.4. Comparisons of *Geranium maderense* and *Pelargonium* x *hortorum* for four assembly parameters using different percentages of sequencing reads.

(A) number of contigs, (B) contiguity, (C) completeness, and (D) percentage of hits. For completeness and contiguity two published reference protein sets were used (357 ultra-conserved ortholog (UCO) coding sequence (Kozik et al., 2008) and 959 single copy nuclear genes (Duarte et al., 2010)). Assemblies were aligned with the reference data sets using BLASTX with an E-value of 1 E-10.

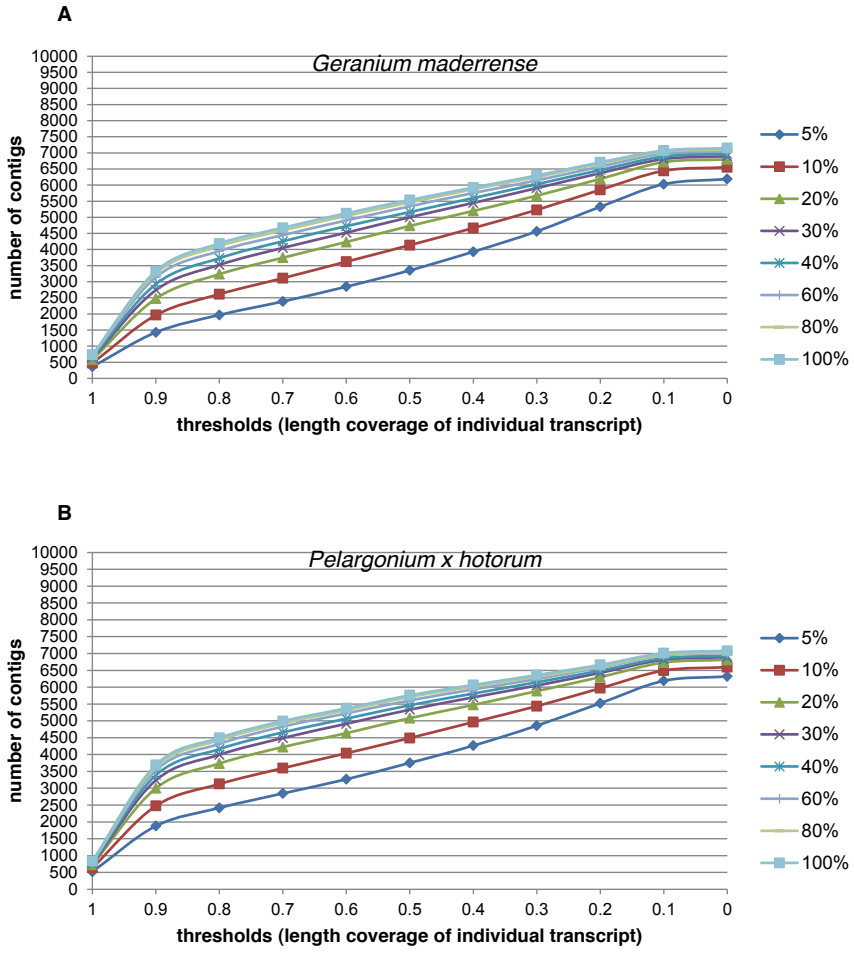


Figure 2.5. Contiguity of *Geranium maderense* and *Pelargonium* x *hortorum* at different threshold values with differ percentages of reads using all *Arabidopsis* proteins from Uniprot/Swissprot database (Boeckmann et al., 2003).

Assemblies were aligned with the database using BLASTX with an evalue of 1 E-10.

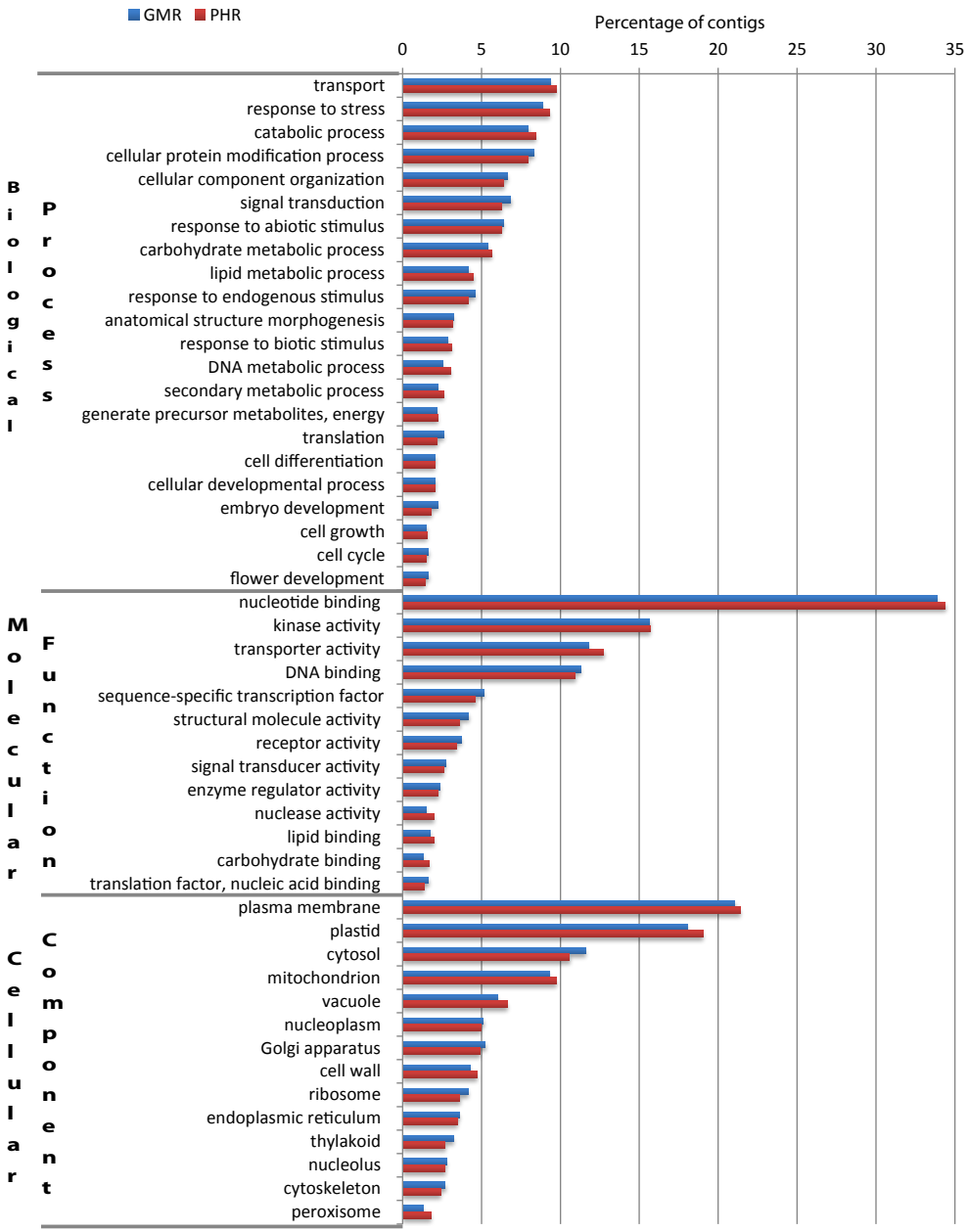


Figure 2.6. Gene ontology assignments for *Geranium maderense* (GMR) and *Pelargonium* x *hortorum*.

The proportion of annotated contigs in all categories with >1% representation within the ontology (GO) categories for cellular component, molecular function, and biological process.

Chapter 3: Coordinated rates of evolution between interacting plastid and nuclear genes in Geraniaceae

INTRODUCTION

Although coevolution of gene sequences is a widely recognized phenomenon in biological systems, it has rarely been studied between the plastid and nuclear genomes of plants within a well-established phylogenetic framework. Coevolution may be detected within a single organism, such as gene pairs with known physical interactions in *Escherichia coli* (Pazos and Valencia, 2001), or between organisms, such as the correlated change of sequences between viral and host genes (Lobo et al., 2009). The coevolution of genes from organellar and nuclear genomes may be considered an intermediate case, in which the genes of interest are within the same organism but are encoded in different cellular compartments. Given that there is an order of magnitude higher mutation rate in nuclear genomes compared to plastid genomes in plants (Wolfe et al., 1987; Drouin et al., 2008), the detection of correlation in evolutionary rates, and how that correlation is maintained, presents an interesting area of study.

As gene function is expressed in amino acid sequences, coevolution between two genes is usually reflected in the encoded polypeptides. If mutual selective pressure exists between two genes, changes to the amino acid sequences encoded in one gene would be expected to cause corresponding changes in the other gene to maintain normal biological activity (Pazos and Valencia, 2008). Similarly, coevolution between two genes can be evaluated based on the rate of nonsynonymous substitutions (*dN*), which are nucleotide mutations that cause a change in the amino acid sequences. However, *dN* is also affected by local rate heterogeneity or local background mutation rates, represented by the rate of synonymous substitutions (*dS*), which do not result in amino acid changes. The correlation of *dS* between two genes is more likely due to a shared mutation rate than an indicator of coevolution.

Several factors can contribute to correlation of evolutionary rates (Lovell and Robertson, 2010), such as obligate physical interaction of gene products (Mintseris and

Weng, 2005), shared functional constraint (Zhang and Broughton, 2013) or gene expression levels (Subramanian and Kumar, 2004). Because the evolutionary rate of the mammalian mitochondrial genome is much higher than that of the nuclear genome, studies of correlated evolution between organellar and nuclear genomes have focused on proteins of enzyme complexes with subunits encoded in each of these compartments. Using this approach, studies have shown that some nuclear genes that encode products that participate in mitochondrial-localized complexes have a corresponding higher evolutionary rate relative to cytosol targeted nuclear gene products (Willett and Burton, 2004; Osada and Akashi, 2012; Barreto and Burton, 2013; Zhang and Broughton, 2013).

The correlation of evolutionary rates between plastid and nuclear genomes has rarely been studied because plastid genome sequences are generally more highly conserved than those of the nuclear genome (Wolfe et al., 1987; Drouin et al., 2008), making it difficult to select appropriate taxa and genes for analyses of correlated rate acceleration. Studies in *Silene* (Sloan et al., 2014) identified elevated protein sequence divergence in organelle-targeted, but not cytosolic, ribosomal proteins in pairwise comparisons of species with rapidly evolving mitochondrial and plastid DNA, suggesting that coevolution occurs between different compartments. Like the *Silene* study, many investigations have adopted pairwise species comparisons, an approach that does not account for the effects of shared phylogeny on predictions of coevolution (Barreto and Burton, 2013).

Methods that incorporate a phylogenetic framework have proven more accurate in detecting coevolution among interacting proteins than pairwise comparisons (Clark and Aquadro, 2010). Various methods have been developed that incorporate the effects of phylogeny for detecting gene coevolution (Pazos and Valencia, 2008; de Juan et al., 2013; Rao et al., 2014). The mirror tree method (Pazos and Valencia, 2001) was originally introduced to predict protein-protein interactions, and it quantifies rate correlations by estimating the similarities of corresponding phylogenetic trees. For each gene tree, the evolutionary rates on each branch are extracted to form a rate vector, and Pearson correlation coefficients are calculated between the rate vectors of two genes.

Despite its popularity, the original mirror tree method does not effectively account for underlying phylogenetic histories. Different modifications of this method were developed to remove the effects of shared phylogeny by introducing a correction factor (Pazos et al., 2005; Sato et al., 2005). A more recent likelihood based approach evaluates the coevolution between genes using normalized *dN*, or *dN*/*dS* (Clark and Aquadro, 2010). In this method, the likelihoods of three models (null, correlated, free) are calculated and the correlation is quantified as the proportional improvement of the likelihood of the correlated model to the null model, with respect to the maximal possible improvement gained by the free model over the null model.

Studies of coevolution of amino acids adopt a different set of approaches, which assess coevolution between two sites by detecting similar amino acid frequencies or substitution patterns calculated from the multiple sequence alignment (Göbel et al., 1994; Neher, 1994; Taylor and Hatrick, 1994). Dutheil and Galtier (2007) developed an approach (CoMap) that examines the coevolution of given amino acid sites using the known phylogenetic history. In this method, the ancestral state of a given amino acid site is inferred from the phylogenetic history and the sequence of changes that occur across time (branches on a phylogenetic tree) form a substitution vector. Structurally-mediated coevolution of any amino acid site is then evaluated using the substitution vector and a cluster-based approach (Dutheil and Galtier, 2007). Another approach studies the coevolved amino acids by incorporating a continuous-time Markov process model (Yeang and Haussler, 2007). Both of these approaches agree well with experimental results; however, the latter approach is computationally demanding and therefore more feasible for studies of small protein domains.

In plants, the plastid-encoded RNA polymerase (PEP) is a multi-subunit enzyme complex (Shiina et al., 2005) containing subunits encoded by genes in both the plastid (RNAP: *rpoA*, *rpoB*, *rpoC1* and *rpoC2*) and nuclear genomes (SIG: *sigma factor 1-6*). Studies in Geraniaceae have revealed highly elevated evolutionary rates in the plastid genome, especially in *rpoB*, *rpoC1* and *rpoC2* (Guisinger et al., 2008; Weng et al., 2012), and highly divergent *rpoA* sequences in the genus *Pelargonium* (Chumley, 2006). The

interaction of SIG and RNAP gene products provides an attractive platform for the study of coevolution between the two genomes. Using transcriptomic and genomic data from 27 species with a well-established phylogenetic framework, the entire sigma factor gene family in Geraniaceae has been characterized and a systematic correlation analysis of evolutionary rates between plastid and nuclear genomes was conducted. Despite an order of magnitude difference in the mutation rate between these two genomes (Wolfe et al., 1987; Drouin et al., 2008), we detected a correlation of evolutionary rates among 27 species representing the entire family. Furthermore, analyses of interacting amino acid pairs suggest that structurally-mediated coevolution plays a minimal role in maintaining the coordination of evolutionary rates. The identification of rate correlations between RNAP and SIG genes suggests a plausible explanation for the observed plastome-genome incompatibility within *Pelargonium* and possibly other genera of flowering plants.

RESULTS

Transcriptome sequencing and assembly for 27 species was performed following Zhang et al. (2013). Sigma factor genes were extracted and accession numbers are provided in Appendix Table 3.1. An amino acid maximum likelihood (ML) tree was generated to infer phylogenetic relationships among the 178 complete sigma factor (SIG) sequences identified from the 27 species in Geraniales and *Arabidopsis thaliana* (Figure 3.1, see Supplemental Data File 3.1 for alignments). The ML tree (-lnL = -65001.9) topology parsed the 178 sequences into six major clades. Two additional alignment algorithms (see Methods) were utilized and resulted in the same six major groups of sigma factor genes (Appendix Figure 3.1, see Supplemental Data File 3.1 for alignments).

The copy number of individual SIG genes varied across different species (Appendix Figure 3.2, see Supplemental Data File 3.1 for alignments). A single copy of *sig1* and *sig2* was found in all species except for *Pelargonium transvaalense*, *P. tetragonum* and *Geranium maderense*, where two copies of *sig2* were identified. A complete s*ig3* sequence was identified in all species except for *P. tetragonum*, *P.*

myrrhifolium and *P. nanum*. The *sig4* sequence was detected in all *Pelargonium* and *Geranium* species, and, while a *sig4* pseudogene missing the start codon was detected in *Melianthus villosus*, *sig4* was not found in *Francoa sonchifolia*, *Erodium chrysanthum* and *E. gruinum*. Two copies of *sig5* and *sig6* were identified in various species (Appendix Figures 3.2E and F). Multiple copies of *sig5* were identified in species of *Geranium* and *Erodium* and in *California macrophylla* while two copies of *sig6* were found in *C. macrophylla* and species of *Erodium* and *Pelargonium*. The *sig6* gene of *Hypseocharis bilobata* contained multiple internal stop codons. RT-PCR confirmed 18 out of 21 bioinformatically identified gene duplication and pseudogenization events (Appendix Table 3.1). Among the SIG gene families, 21 gene duplications and 10 losses were inferred with Notung (Durand et al., 2006) (Appendix Figure 3.3, Appendix Table 3.2, see Supplemental Data File 3.1 for alignments) on the branches leading to the 27 species of Geraniaceae.

Evolutionary rates of each gene were estimated based on alignments from MAFFT (Katoh and Standley, 2013). To avoid biases of rates estimation specific to an alignment algorithm, rates based on alignments from two other tools, MUSCLE and ClustalW (Edgar, 2004; Larkin et al., 2007), were compared to MAFFT (see Supplemental Data File 3.1 for alignments). The agreement between rate estimates from the three alignment methods indicated that there was no or negligible bias due to the alignment method (Appendix Table 3.3). Thus MAFFT was used for all subsequent analyses.

Clade-specific rate acceleration was assessed for the four plastid RNA polymerase (RNAP) subunits, six nuclear encoded SIG genes and 20 control genes (Figure 3.2, Appendix Figures 3.4 and 3.5). Ten control genes from nuclear and plastid genomes were selected for all coevolution analyses, and two additional sets of ten nuclear control genes were randomly selected from the APVO database (see Methods) for the mirror tree analysis to reduce any bias of nuclear control gene sampling (see Supplemental Data File 3.2 for detailed control gene information).

Although *dN* for the *Pelargonium* C clade was accelerated in all four RNAP genes and two SIG genes (*rpoA*, *rpoB*, *rpoC1*, *rpoC2*, *sig1*, *sig2*) in an initial ranksum test, acceleration of rates of *rpoA*/*B*/*C1* only remain significant after correction for multi-hypothesis testing (Figure 3.2), while two control genes have significant acceleration of rates in the *Geranium* and *Erodium* clades. Significant acceleration of *dS* in the *Pelargonium* C clade was observed for the *rpoA* gene alone (Appendix Figure 3.4). Elevated *dS* in *sig6* and five nuclear control genes was observed in *Geranium*, with no acceleration detected within other clades (Appendix Figure 3.4, see Supplemental Data File 3.1 for alignments).

The values of *dN* and *dS* for each gene from all branches were used to analyze the rate correlation between gene pairs from RNAP, SIG and control genes. The highest average values for *dN* were found in SIG genes followed by RNAP genes (Figure 3.3A, Appendix Figure 3.5). Four plastid genes (*cemA*, *matK*, *rpl14*, *rps2)* and one nuclear gene (*rh22*) had similar average *dN* values to the RNAP genes, and the other nuclear genes had slightly lower values. The lowest average *dN* values were found in the remaining plastid genes, which represent ATP synthase and photosynthetic genes. The *dS* values were similar among genes from the same cellular compartment (Figure 3.3B). The average *dS* values of nuclear genes were much higher than those of plastid genes except for *rpoA*, which had the highest *dS* value among plastid genes.

Correlation of *dN* and *dS* was evaluated for each gene pair by three variations of the mirror tree method, each of which adopts a different approach for removing the effect of shared phylogeny prior to tree similarity estimation (Pazos et al., 1997; Pazos and Valencia, 2001; Pazos et al., 2005): average by all, average by separation and PCA (see Methods for detailed description). The sequences of *sig1*, *sig2* and *sig5* were grouped together for rate correlation analysis using complete sequences (Figure 3.4) or conserved domains (Appendix Figure 3.6). Due to their absence in different species, the genes *sig3*, *sig4* and *sig6* were analyzed separately using complete sequences (Appendix Figure 3.7) or conserved domains (Appendix Figure 3.8). In addition to the initial 10 plastid and

nuclear control genes, two additional sets of nuclear control genes were added to the *sig1*, *sig2* and *sig5* analyses (Appendix Figures 3.9 and 3.10).

A cutoff of 0.6 for the Pearson correlation coefficient was used as an indicator of strong rate correlation (Sato et al., 2005). After removing the effects of shared phylogeny (see Methods), correlation of *dN* was detected between RNAP and *sig1/2* genes (orange rectangle in Figures 3.4A and B), but not between RNAP/SIG and the control genes. The mirror tree methods did not detect a *dN* correlation between RNAP and *sig3*, *sig4* or *sig6* (Appendix Figure 3.7). The correlation of *dS* was sensitive to the average method used in the analyses (Figures 3.4A and B). Correlation of *dS* was identified between certain plastid or nuclear gene pairs but not between the two groups when the average by all method was employed (Figure 3.4A), and only one pair of nuclear genes had correlated *dS* when the average by separation method was used (Figure 3.4B). Application of the PCA method produced correlations of *dN* and *dS* that were similar to those from the average by all method except that correlation of *dN* was detected between *rpl14*/*rpoA* and *sig1* (Figure 3.4C). None of the three methods identified correlation of *dS* between RNAP and SIG genes (Figures 3.4A-C), suggesting that the correlation of *dN* was not due to the effects of background mutation rates. The number of gene pairs with positive rate correlations is shown in Table 3.1. Similar rate correlations were detected with the additional nuclear control genes (Appendix Figures 3.9 and 3.10).

The correlation of *dN* and *dS* between RNAP and SIG genes using conserved domains was similar to that seen using the entire sequences; however, more gene pairs of RNAP and SIG were identified as correlated for *dN* (Appendix Figures 3.6 and 3.8). The number of rate correlations of all gene pairs is shown in Appendix Table 3.4.

The rate correlation coefficient between each individual gene and the RNAP genes was compared (Table 3.2). The *dN* correlation coefficients of RNAP and RNAP/*sig1*/*sig2* genes were ranked significantly higher ($p < 0.05$ after correction for multi-hypothesis testing) than all other pairs by average by separation and PCA methods. Using the average by separation method, correlation coefficients of RNAP genes and *sig6* were also ranked significantly higher than other pairs. No significantly higher rank was

detected between RNAP and the plastid or nuclear control genes by any method. Synonymous substitution rate correlation coefficient ranking produced no significant result for any of the gene groups. The same tests were performed with rates calculated from the conserved domains of selected genes as described (Appendix Table 3.5). Similar to the results generated using the entire sequences, the *dN* correlation coefficients of RNAP and RNAP/*sig1*/sig2/sig6 were ranked significantly higher than any other pairs by average by separation and PCA methods. The correlation coefficients of *dN* for RNAP and *sig5* were ranked significantly higher using PCA method. The rank of *dS* correlation coefficients was the same as that using the entire sequences with no significant highly ranked gene groups detected.

Correlation of normalized *dN* (*dN*/*dS* ratio) was evaluated with the proportional improvement method (Clark and Aquadro, 2010). Since the proportional improvement *dN*/*dS* test is more appropriate when *dS* is unsaturated (Clark and Aquadro, 2010), saturation was tested for each of the genes examined. To examine saturation of synonymous sites, values of *dN* and *dS* were plotted and linear/quadratic models were used to fit the data. If *dS* is saturated, the quadratic model with a concave curve should fit the data better. The two models were compared with the improvement of sum of squares explained by these models (Weng et al., 2014; Fares and Wolfe, 2003). Of the 30 genes tested, only *rbcL* showed significant improvement ($p < 0.05$) of sum of squares (Appendix Table 3.6), however, *rbcL* is known to be a conservative gene with low *dS* values (*dS* < 0.15 for all branches) (Figure 3.3*B*).

Strong correlation (proportional improvement > 0.6) of *dN*/*dS* was identified between *rpoB*/*C1*/*C2* and *sig1*/*5*/*6* genes, between *rpoC1*/*C2* and *sig2* genes, and between *rpoB*/*C2* and *sig3* genes (Figure 3.5). Correlation of *dN*/*dS* was also identified among RNAP (between *rpoB* and *rpoC1*/*C2*; *rpoC1* and *rpoC2*) and among SIG (between *sig1* and *sig2*/*5*/*6*; *sig2* and *sig5*) genes (Figure 3.5). A correlation of RNAP/SIG and control genes was lacking between most interaction pairs except for *rpoB* and three nuclear control genes (*OXase*, *ppr* and *nprb7*) and *sig1*/*2* and *nprb7* genes (Figure 3.5). Compared to the mirror tree methods, more interaction pairs (proportional improvement:

13, average by all: 2, average by separation: 4, PCA: 6) were identified with strong rate correlation between RNAP and SIG genes, while fewer or comparable interaction pairs (proportional improvement, 5; average by all, 20; average by separation, 5; PCA, 2) were identified between RNAP/SIG and control genes.

To investigate the role of structurally-mediated coevolution in the correlation of evolutionary rates between RNAP and SIG genes, CoMap (Dutheil and Galtier, 2007) was used to predict coevolved amino acid pairs by comparing the substitution vectors, weighted by the different amino acid properties (volume, charge and polarity) at given positions (see Methods for more details). Since the β' subunit of the cyanobacterial ancestor was split in the lineage leading to plants (Shiina et al., 2005), the relevant residues from the β' and β'' subunits in plants were combined for comparison with the β' subunit in *E. coli*. The interaction sites between SIG and RNAP subunits were predicted by contact map analysis (Sobolev et al., 2005) and by estimation of the physical distance between two interacting residues. More interaction sites were predicted by contact map analysis than by distance estimation (Appendix Table 3.7); however, few (0 to 20%) of the predicted coevolving amino acid sites overlapped with interaction sites (Appendix Table 3.7). The analysis of distance distributions across the coevolved amino acid pairs suggested that among the 4223 residue pairs predicted to be involved in structurally-mediated coevolution by CoMap, only one pair had a distance of less than 5 Å. The analyses of structurally-mediated coevolution within amino acid pairs showed a minimal overlap between coevolved and interacting amino acid sites (Appendix Table 3.7), with few of the coevolved residues in close physical proximity (<5 Å, Appendix Figure 3.11).

DISCUSSION

Duplication and loss of sigma factor genes

Twenty-one duplicated SIG genes were identified across the Geraniaceae (Appendix Figures 3.2 and 3.3, Appendix Table 3.2). Duplications of SIG genes have been documented in several angiosperms, including *Zea mays*, *Oryza sativa* and *Populus trichocarpa* (Lerbs-Mache, 2011). These duplicate copies may be functionally

diversified to regulate gene expression at different developmental stages or under changing environmental conditions as seen for the duplicated *sig1* of *Zea mays* that is differentially expressed in etiolated leaves (Tan and Troxler, 1999; Lerbs-Mache, 2011). The pattern of SIG gene duplication in Geraniaceae could have arisen in several ways, including whole genome duplication followed by elimination of some copies. Polyploidy is widespread across Geraniaceae (Widler-Kiefer and Yeo, 1987; Yu and Horn, 1988; Touloumenidou et al., 2007) and has likely contributed to duplication of SIG genes. However an alternative explanation, that multiple, small scale gene duplications have occurred (Davis and Petrov, 2004; Li et al., 2006) was supported by the pattern of SIG gene duplications observed in Geraniaceae (Appendix Figures 3.2 and 3.3).

Multiple gene losses were detected in SIG gene family in Geraniales. While it is possible that these genes are so lowly expressed as to fall below the level of detection, the high depth of coverage in transcriptome sequencing (Zhang et al., 2013) and the fact that the same genes were identified in transcriptomes of closely related species make this unlikely. The Sig4 protein specifically recognizes the promoter of *ndhF*, which encodes a subunit of NADH dehydrogenase in the plastid (Endo et al., 1999; Favory et al., 2005). The lack of *sig4* transcripts in *Erodium chrysanthum* and *E. gruinum* is plausible given the loss of *ndh* genes from the plastid genomes of these species (Blazier et al., 2011). The identification of a *sig4* pseudogene in *Melianthus* also correlates with a recent loss of *ndh* genes in that species (Weng et al., 2014). The loss of both *sig*4 and *ndh* genes provides an explicit example of coevolution between the plastid and nuclear genomes.

Coevolution of plastid and nuclear genomes

Genome coevolution is expected to produce correlated evolutionary rate changes between different genes. Studies of coevolution usually focus on protein sequences from multi-subunit enzyme complexes. Correlated change of evolutionary rates has been widely observed in various organisms (Lovell and Robertson, 2010; Pazos and Valencia, 2008; Campo et al., 2008). A previous study showed that nuclear genes encoding

subunits of enzyme complexes that assemble in the mitochondria with subunits encoded in the organelle have significantly higher evolutionary rates than genes whose products are targeted to the cytosol (Barreto and Burton, 2013). Likewise, analyses of coevolution between organellar and nuclear genomes have mainly focused on genes from mitochondrial and nuclear genomes (Zhang and Broughton, 2013; Barreto and Burton, 2013). A recent study of *Silene* (Sloan et al., 2014) suggested that coevolution occurred between plastid and nuclear genomes based on the observation of elevated protein sequence divergence in organelle targeted, but not cytosolic, ribosomal proteins for species with fast evolving mitochondrial and plastid genomes. The unusually high substitution rates of genes in the plastid genomes of Geraniaceae (Guisinger et al., 2008; Weng et al., 2012; Chumley, 2006) provide an attractive system for the study of coevolution. Using both transcriptome and genome data and a well-characterized phylogenetic framework, this systematic analysis revealed the existence of correlation of evolutionary rates, evidence of coevolution between plastid and nuclear genomes at different levels (*dN* and *dN*/*dS*).

Correlation of *dN* but not *dS* was detected in gene pairs between RNAP and SIG genes, suggesting that the correlation of *dN* was not due to shared background mutation rates. The absence of correlation of *dN* between RNAP/SIG and control genes indicates that the rate correlation between RNAP and SIG genes is likely due to coevolution between plastid and nuclear genome or a local functional constraint acting on RNAP and SIG genes, rather than a global constraint on *dN* of all genes. The case of *dN* correlation between *rpoA* and *rpl14,* detected exclusively by the PCA method, may be a result of factors other than direct physical interaction, such as a common functional role in gene expression in plastids (transcription for *sig1* and translation for *rpl14*) (Chen and Dokholyan, 2006; Agrafioti et al., 2005).

Because correlation coefficients of gene pairs with known interactions are significantly higher than unrelated sequences (Clark et al., 2012), correlation coefficients between each SIG gene and the four RNAP genes should be higher than those between any control genes and the four RNAP genes. The significantly highly ranked correlation

coefficients of *dN* between RNAP and *sig1/2* by any method supports the conclusion that there is a strong rate correlation between RNAP and *sig1/2* genes (Table 3.2). The significantly highly ranked rate correlation detected between RNAP and *sig5/6* genes using conserved domain sequences in both average and PCA methods suggests that there might be correlation between RNAP and *sig5/6* genes, but that it is weaker than those between RNAP and *sig1/2* (Appendix Table 3.4).

A correlation analysis of normalized *dN* (*dN*/*dS* ratio) was performed (Clark and Aquadro, 2010) and this approach identified additional strong rate correlations between RNAP (*rpoB*/*C1*/*C2*) and *sig3*/*5*/*6*. The low number (5/400) of strong rate correlation pairs (Figure 3.5) between RNAP/SIG and control genes is the result of either weaker correlation or false discoveries.

Across all methods, strong rate correlations (*dN* and/or *dN*/*dS*) are present for RNAP and all SIG except *sig4*. Rate correlations among interacting genes are affected by several factors (Lovell and Robertson, 2010), such as physical interaction (Mintseris and Weng, 2005), functional constraint (Zhang and Broughton, 2013) or gene expression levels (Fraser et al., 2004). The *sig4* gene is involved in the transcription of *ndhF* (Favory et al., 2005). Knockout studies of *sig4* in Arabidopsis revealed no observable phenotypes (Lerbs-Mache, 2011) and the loss of it and the corresponding plastid encoded *ndh* genes (Blazier et al., 2011) in multiple species in Geraniaceae suggests that *sig4* is dispensable. Relaxed functional constraint may contribute to the absence of coevolution between RNAP and *sig4* genes.

Possible phenomena that could underlie the rate correlations between RNAP and SIG genes include: 1) a cause-and-effect relationship between rate variation of RNAP and SIG genes, or 2) a common factor affecting rates of both RNAP and SIG genes such as relaxed functional constraint acting on both gene sets (Subramanian and Kumar, 2004; Mintseris and Weng, 2005; Lovell and Robertson, 2010; Zhang and Broughton, 2013). If the first explanation were correct, rate correlations between RNAP and SIG genes could be due to structurally-mediated compensatory evolution. However results suggest that structurally-mediated coevolution plays a minor role in maintaining rate correlations

between SIG and RNAP subunits and other factors contributing to compensatory evolution could not be excluded. Shared functional constraint is another agent that may be maintaining the observed correlations. Additional study is required to further elucidate the contribution of functional constraints, gene expression or other factors in the correlation of evolutionary rates of PEP subunits in Geraniaceae.

Plastome-genome incompatibility, or PGI, which was first documented in *Pelargonium* species (Smith, 1915), is observed across flowering plants (Schmitz-Linneweber et al., 2005; Greiner et al., 2008; Weihe et al., 2009; Greiner et al., 2011). Various mechanisms have been proposed for PGI including impaired interactions between *cis* elements and their cognate nuclear factors involved in transcription and/or transcript stability. In *Oenothera,* perturbations of photosystem II activity, presumed to be caused by changes in transcription of the *psbB* gene, contributes to PGI (Greiner et al., 2008). Likewise, steady state RNA levels of three PEP-controlled genes were severely reduced in leaf sections taken from variegated, interspecific hybrids of *Zantedeschia* (Yao and Cohen, 2000). The two nuclear genes included in the *Zantedeschia* study also evidenced low levels of mRNA; expression of *cab* and *rbsS* is known to be regulated by retrograde plastid to nuclear signals and would therefore be susceptible to the PGI phenotype (Ruckle et al., 2007). Correlation of accelerated nucleotide substitution rates between SIG and RNAP genes provides a plausible explanation for PGI in Geraniaceae (Weihe et al., 2009; Greiner et al., 2011). Specifically, the high evolutionary rates and rate correlation between SIG and RNAP genes within species could lead to interspecific incompatibilities, and such incompatibilities would reduce efficiency or even cause dysfunction of the PEP holoenzyme, impairing the transcription of essential plastid genes. The role of sigma factors in transcription initiation through *cis* element binding and polymerase recruitment suggests similarity between the Geraniaceae, *Zantedeschia* and *Oenothera* PGI systems.

METHODS

RNA isolation, transcriptome sequencing and assembly

Total RNA was isolated from newly emerged leaves of 27 species in Geraniales (Appendix Figure 3.5), and four tissues (emergent and expanded leaves, roots and flowers) of *Pelargonium* x *hortorum* following the protocols described in Zhang et al. (2013). Transcriptome sequencing was performed on the HiSeq 2000 platform and the sequence data were preprocessed and assembled as described in Zhang et al. (2013).

Identification of sigma factors

Sigma factor (SIG) sequences were extracted from transcriptome assemblies with reciprocal BLAST as described in Zhang et al. (2013). The orthologous genes of each class of SIG were determined by reciprocal BLAST and single gene phylogeny. RT-PCR was performed to verify the problematic SIG gene sequences with internal stop codons or missing 5' or 3' ends. For any species in which multiple gene sequences were identified as the same class of sigma factor, RT-PCR was performed to verify the existence of all the gene sequences (for primers see Appendix Table 3.8). All RT-PCR products were subjected to Sanger sequencing to confirm the result.

Phylogenetic analysis

Multiple sequence alignments were done using MAFFT (Katoh and Standley, 2013), MUSCLE (Edgar, 2004) and ClustalW (Larkin et al., 2007) with default parameters in Geneious 6.0 (Biomatters, http://www.geneious.com/) (see Supplemental Data File 3.1 for alignments). Amino acid-based ML trees for all SIG genes were constructed by RAxML (Stamatakis, 2006) with parameters "raxmlHPC-PTHREADS-SSE3 -f a -x 12345 -p 12345 -T 12 -m PROTGAMMAJTT -N 100". A Perl script (http://sco.h-its.org/exelixis/web/software/raxml/hands_on.html) was used to examine all protein models and the model with the best likelihood score (JTT) was selected. ML trees of each class of SIG genes were constructed by RAxML (Stamatakis, 2006) with parameters "raxmlHPC-PTHREADS-SSE3 -f a -x 12345 -p 12345 -T 12 -m

GTRGAMMAI -N 100". Bootstrap values were generated using RAxML with 100 replicates and the above settings.

Evolutionary rate estimation

PAML's codeml (Yang, 2007) was used to estimate *dN*, *dS* and *dN*/*dS* using the codon frequencies model F3X4. Gapped regions were excluded with parameter "cleandata = 1". The constraint tree was generated by RAxML using 12 plastid genes (*atpA*, *atpB*, *atpI*, *ccsA*, *cemA*, *matK*, *petA*, *rbcL*, *rpoB*, *rpoC1*, *rpoC2*, *rps2*) with a total length of 21,500 bp. Bootstrap values were generated using RAxML with 100 replicates and the above settings. Ten plastid genes (*rbcL*, *cemA*, *atpA*, *atpB*, *matK*, *petB*, *psaA*, *psbC*, *rpl14*, *rps2*) from different functional groups and thirty nuclear genes with three different subcellular targeting sites (plastid, mitochondria and other), which are orthologous to genes in the APVO database (Duarte et al., 2010), were used as negative control groups. The APVO database was separated into three groups based on their subcellular locations (plastid, mitochondria and other), and an approximately equal number of genes were selected randomly from each group. The plastid genes were extracted from the annotated plastid genome assemblies as described in Weng et al. (2014). Thirty Arabidopsis nuclear genes were downloaded from TAIR (Lamesch et al., 2012) and the corresponding accession numbers are in Supplemental Data File 3.2.

Analysis of correlation of evolutionary rate

Rates along branches leading to and within *Pelargonium* A, B and C clades, *Monsonia*, *Geranium* and *Erodium* I and II clades were grouped separately for clade specific rate acceleration analysis. The Ranksum test was performed to test clade-specific rate accelerations and a P value of less than 0.05 was considered significant. Correction for multi-hypothesis testing was performed by adjusting the original P value with the FDR correction method using the built-in p.adjust function (method="fdr") in the R software package (http://www.r-project.org). After correction, the false discovery rate among the significantly accelerated clades within each gene is less than 5%.

Correlation coefficients of evolutionary rates *dN* and *dS* between each gene pair were estimated using modified mirror tree methods (Pazos et al., 1997; Pazos and Valencia, 2001; Pazos et al., 2005). Specifically, the evolutionary rates, *dN* or *dS,* on each branch of a given gene tree were collected to form a rate vector. The rate correlation was quantified using the Pearson correlation coefficient between the rate vectors of different genes. The rate vector of each gene pair was adjusted via vector projection by a correction vector representing the shared phylogenetic effects prior to the comparison. The correction vector was generated with different modifications of the mirror tree method, the average method and principle component analysis (Sato et al., 2005). In the average method, the correction vector was determined in two ways, the average by all and the average by separation, in which either the correction vector was defined as the average of rate vectors of all genes, or two different correction vectors for nuclear and plastid genes were defined separately, as the average of rate vectors of corresponding gene groups. In principle component analysis, the correction vector was calculated as the first principle component of the rate matrix formed by rate vectors of all genes.

The correlation of *dN*/*dS* ratio of most (447/450) interaction pairs, quantified as "proportional improvement" as described in (Clark and Aquadro, 2010), was analyzed using HYPHY (Pond et al., 2005) with batch scripts downloaded from http://mbg.cornell.edu/cals/mbg/research/aquadro-lab/software.cfm (Clark and Aquadro, 2010). Specifically, the evolutionary rates and the likelihood of three (correlated, null, free) models for the estimated rates of each gene pair were evaluated with HYPHY, and the proportional improvement method estimates correlation of *dN*/*dS* ratio by calculating the proportional improvement of likelihood of the correlated model over the null model, with respect to the maximal possible improvement gained by the free model over the null model (Clark and Aquadro, 2010). The remaining interaction pairs (3/450, *psbC* and *psaA*, *psbC*/*psaA* and *nprb7*) were analyzed using the same batch script template with modifications so that the likelihoods of the correlation model with different start points (-

0.8, 1, 1.3) were optimized separately rather than sequentially. The test for *dS* saturation was performed as described in Fares and Wolfe (2003) and Weng et al. (2014).

The conserved domains of RNAP and SIG genes were predicted by NCBI CDD (Marchler-Bauer et al., 2013). The predicted conserved domains were used for rate analysis, except for *rpoC1* and *rpoC2,* because conserved domains were predicted to comprise the entire sequence for both of these genes. Rate corrections were done by custom python scripts and are available as Supplemental Data File 3.3. The Pearson correlation coefficients were calculated using the built in function in the python scipy module. A correlation coefficient value of 0.6 or above was used to indicate a positive rate correlation (Sato et al., 2005). The rate correlation of each gene with itself was removed from all analyses. The one side Wilcoxon Rank Sum test and correction for multi-hypothesis testing was performed using R software package as described above.

Structurally-mediated coevolution within groups of amino acids were evaluated for 28 plant species (27 Geraniales from this study plus Arabidopsis) and *E. coli* using CoMap (Dutheil and Galtier, 2007). To evaluate structurally-mediated coevolution between amino acids from different genes, three different features (volume, charge, polarity) of amino acids were considered and amino acid substitutions at specific sites on each branch of the gene tree were quantified based on these features. For each gene, the changes of amino acids on all branches at each site were extracted to form a site-specific substitution vector. The site-specific substitution vectors from two different genes were compared to find structurally-mediated coevolved changes (i.e. volume increase at site X of gene A and volume decrease at site Y of gene B) of sites from different genes. The interacting amino acid pairs between SIG and RNAP subunits were predicted with CMA (Sobolev et al., 2005), and by mapping the amino acid pairs with distance between them of less than 5 Å (Hu and Yan, 2009) in RNA polymerase of *E. coli* to those of Geraniales using custom python scripts (Supplemental Data File 3.3). The distribution of distances of the coevolved amino acid pairs identified in the structurally-mediated evolutionary analysis, and the overlap between coevolved and the interacting amino acid pairs were executed with custom python scripts (Supplemental Data File 3.3).

Table 3.1. The number of interaction pairs with a rate coefficient of over 0.6 within corresponding genes estimated by three mirror tree methods.

RNAP contains *rpoA*, *rpoB*, *rpoC1* and *rpoC2*. Results of the average by all method (ρ_{ava}), average by separation method (ρ_{avs}), and PCA method (ρ_{pca}) are shown.

	dN			***dS***		
interactions[a]	ρ_{ava}	ρ_{avs}	ρ_{pca}	ρ_{ava}	ρ_{avs}	ρ_{pca}
RNAP – RNAP (6)	5	3	5	3	0	2
RNAP – *sig1* (4)	1	2	3	0	0	0
RNAP – *sig2* (4)	1	2	3	0	0	0
RNAP – *sig3* (4)	0	0	0	0	1	0
RNAP – *sig4* (4)	0	0	0	0	0	0
RNAP – *sig5* (4)	0	0	0	0	0	0
RNAP – *sig6* (4)	0	0	0	0	0	0
RNAP – control (80)	0	0	0	10	0	8
***sig1* – control (20)**	0	0	1	4	0	3
***sig2* – control (20)**	0	0	0	5	0	4
***sig3* – control (20)**	8	5	1	3	5	3
***sig4* – control (20)**	11	0	0	5	1	2
***sig5* – control (20)**	0	0	0	8	0	5
***sig6* – control (20)**	1	0	0	10	0	9

[a]The number in parentheses is the total number of interaction pairs within corresponding genes.

Table 3.2. Ranksum test of rate correlation coefficient.

The entire sequence of each gene was used in the analysis. Correlation coefficients ranked significantly higher among all interaction pairs are indicated with "+". Coefficients that are not ranked significantly higher are indicated with "-". Pt is the group of control genes from the plastid genome and nu is the group of control genes from the nuclear genome. Results of the average by all method (ρ_{ava}), average by separation method (ρ_{avs}) and PCA method (ρ_{pca}) are shown. Results from average method were estimated in two ways (all/separate, see methods).

	dN			***dS***		
interactions	ρ_{ava}	ρ_{avs}	ρ_{pca}	ρ_{ava}	ρ_{avs}	ρ_{pca}
RNAP - RNAP	+	+	+	-	-	-
RNAP - *sig1*	-	+	+	-	-	-
RNAP - *sig2*	-	+	+	-	-	-
RNAP - *sig3*	-	-	-	-	-	-
RNAP - *sig4*	-	-	-	-	-	-
RNAP - *sig5*	-	-	-	-	-	-
RNAP - *sig6*	-	+	-	-	-	-
RNAP - pt	-	-	-	-	-	-
RNAP - nu	-	-	-	-	-	-

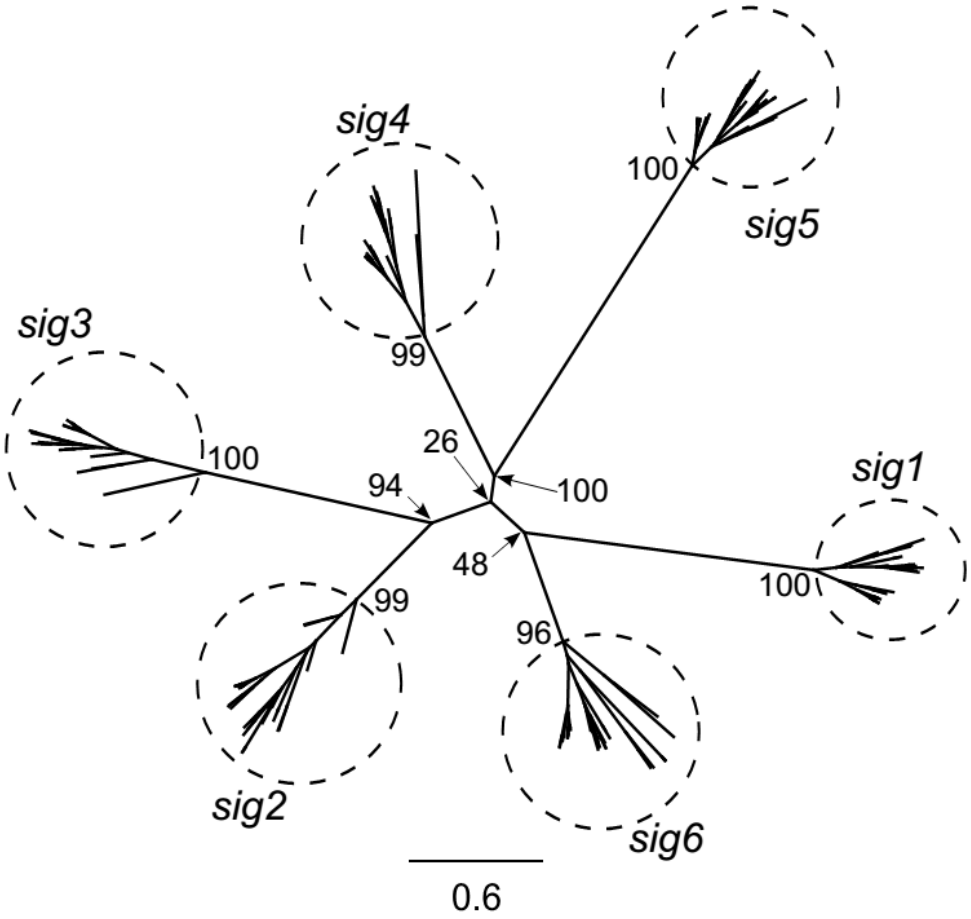
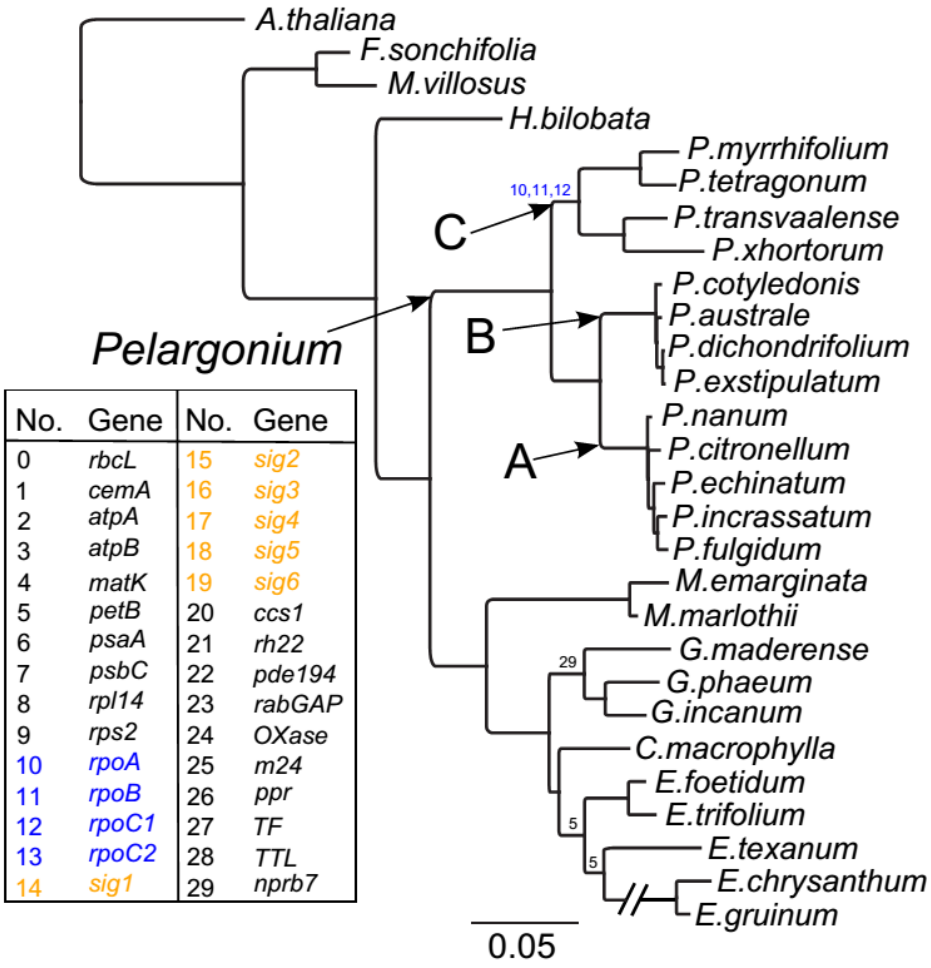


Figure 3.1. Six sigma factor families in Geraniales and Arabidopsis.

The unrooted amino acid-based maximum likelihood (ML) tree was generated using 178 complete sigma factor (SIG) sequences identified from 27 species of Geraniales and *A. thaliana*. The ML tree (-lnL = -65001.9) topology parsed the 178 sequences into six subgroups (enclosed in labeled circles). Scale bar represents the number of amino acid substitutions per site. Numbers at nodes are bootstrap support values.



No.	Gene	No.	Gene
0	*rbcL*	15	*sig2*
1	*cemA*	16	*sig3*
2	*atpA*	17	*sig4*
3	*atpB*	18	*sig5*
4	*matK*	19	*sig6*
5	*petB*	20	*ccs1*
6	*psaA*	21	*rh22*
7	*psbC*	22	*pde194*
8	*rpl14*	23	*rabGAP*
9	*rps2*	24	*OXase*
10	*rpoA*	25	*m24*
11	*rpoB*	26	*ppr*
12	*rpoC1*	27	*TF*
13	*rpoC2*	28	*TTL*
14	*sig1*	29	*nprb7*

Figure 3.2. Shared clade-specific nonsynonymous rate (*dN*) acceleration in Geraniaceae.

RNAP and SIG genes are highlighted in the key in blue and orange, respectively. Blue numerals on the constraint tree indicate shared *dN* acceleration in RNAP genes of the *Pelargonium* C clade. For a more detailed version of the constraint tree see Appendix Figure 3.5. Numbers at nodes indicate accelerated *dN* in corresponding gene from the key at left (0-14: plastid genes, 15-29: nuclear genes). Scale bar represents the number of nucleotide substitutions per codon. The long branch leading to *E. chrysanthum* and *E. gruinum* was interrupted for ease of visualization.

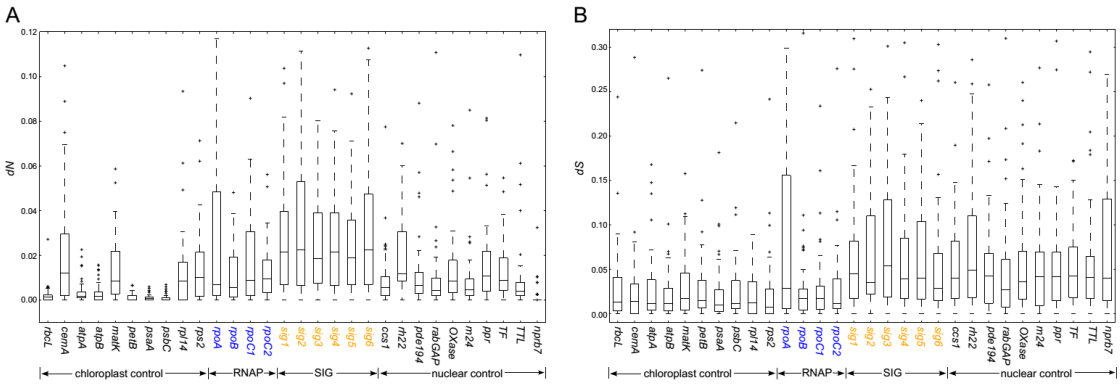


Figure 3.3. Nonsynonymous (*dN*) and synonymous (*dS*) substitution rates for individual genes.

Box plots represent the distribution of (A) *dN* or (B) *dS* value for each branch on the constraint tree (Appendix Figure 3.5). While the *dS* values (B) were similar among genes from the same cellular compartment, the highest average values for *dN* (A) were found in SIG genes (orange) followed by RNAP genes (blue). The scale for *dN* and *dS* is different to facilitate visualization of the results.

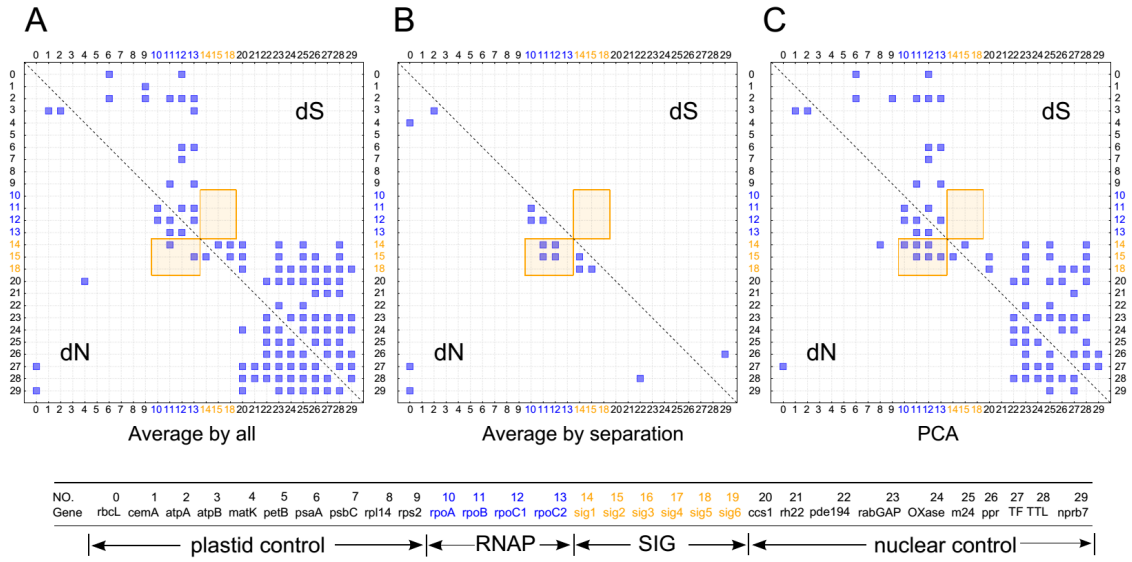


Figure 3.4. Strong correlation of nonsynonymous (*dN*) but not synonymous (*dS*) substitution rates between *sig1/2/5* and RNAP genes using three methods of analysis.

The entire sequence of each gene was used in this analysis (see Methods). The correlation of *dN* and *dS* values were calculated by modifications of the mirror tree method (A) average method (all), (B) average method (separation) and (C) PCA. All interaction pairs with a correlation coefficient of higher than 0.6 were considered significant and shown with a blue square. RNAP and SIG genes, highlighted in blue and orange fonts respectively, show strong correlation of *dN* but not *dS* (highlighted in orange shaded box). Little to no correlation of *dN* between RNAP/SIG genes and the control genes (in black font) was detected. Gene names and cellular locations corresponding to each number are given below the diagram.

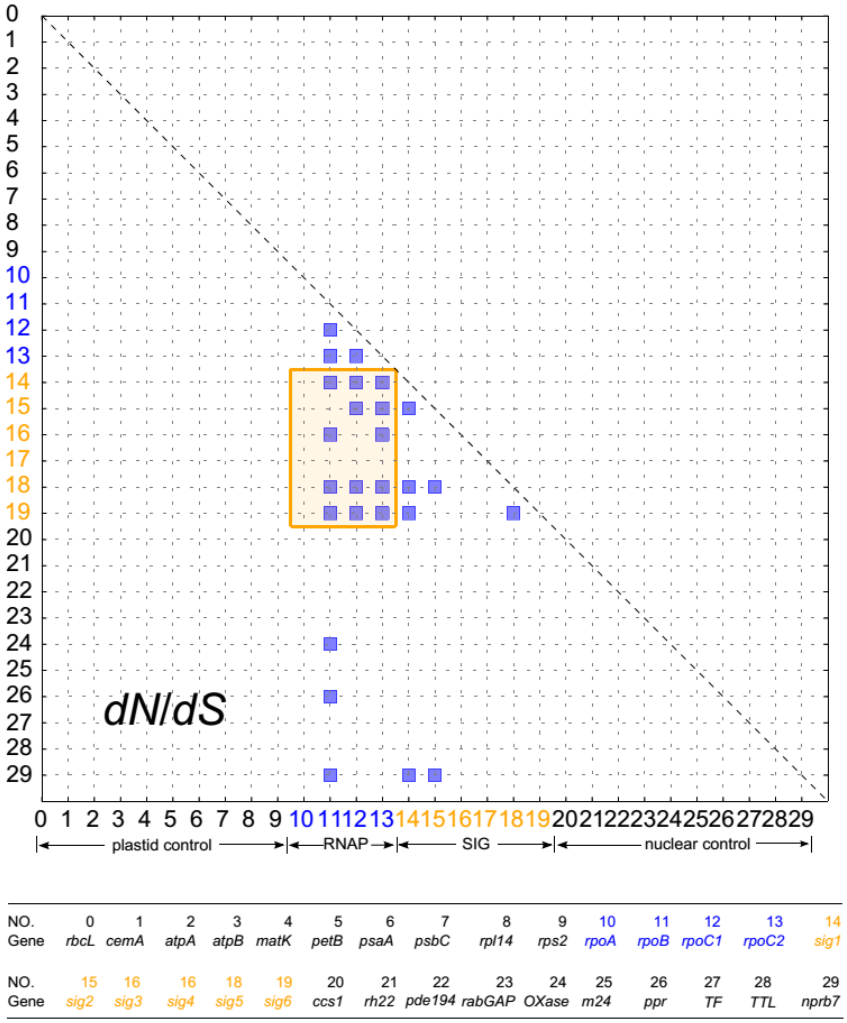


Figure 3.5. Strong correlation of *dN*/*dS* between RNAP and SIG genes.

The correlation of *dN*/*dS* was quantified using proportional improvement (Clark and Aquadro, 2010). Interaction pairs with a proportional improvement of higher than 0.6 were considered significant and shown with a blue square in the figure. RNAP and SIG genes, highlighted in blue and orange fonts, respectively, show strong correlation of *dN*/*dS* (highlighted in orange shaded box). Little to no correlation of *dN*/*dS* between RNAP/SIG genes and the control genes (in black font) was detected. Gene names corresponding to each number are given below the diagram.

Chapter 4: Coevolution between rates of nuclear encoded DNA RRR genes and plastid genome complexity

INTRODUCTION

Plastid genomes of angiosperms are generally highly conserved with a quadripartite structure that includes large single copy (LSC), small single copy (SSC) and two inverted repeat (IR) regions and a conserved gene content and order (Ruhlman and Jansen, 2014). However, several unrelated lineages have experienced substantial variation in genome organization, including Campanulaceae (Cosner et al., 2004; Haberle et al., 2008; Knox, 2014), Ericaceae (Fajardo et al., 2012; Martínez-Alberola et al., 2013), Geraniaceae (Chumley et al., 2006; Guisinger et al., 2008, 2011; Chris Blazier et al., 2011; Weng et al., 2014) and Fabaceae (Milligan et al., 1989; Perry et al., 2002; Cai et al., 2008; Sabir et al., 2014). In addition to highly rearranged plastid genomes, there is evidence for a correlation between rates of plastid genome rearrangements and nucleotide substitutions in several lineages (Jansen et al., 2007; Guisinger et al., 2008; Sloan et al., 2012; Weng et al., 2014). The cause of the correlation between genome complexity and evolutionary rates is not clear but it has been hypothesized that it may be due to alterations in DNA repair and recombination mechanisms.

Among angiosperms changes of plastid genome complexity in Geraniaceae are unprecedented in number and diversity. A large number of genome rearrangements have been detected in multiple genera, including *Hypseocharis*, *Pelargonium*, *Monsonia*, *Geranium* and *Erodium* (Chumley, 2006; Blazier et al., 2011; Guisinger et al., 2011; Weng et al., 2014). Two clade-specific IR loss events occur in *Erodium* and *Monsonia* (Blazier et al., 2011; Guisinger et al., 2011), and an order of magnitude difference in IR size (7kb to 75kb) is present among the remaining species in the family (Guisinger et al., 2011). Highly elevated nonsynonymous (*dN*) substitution rates are known in plastid encoded genes in Geraniaceae (Guisinger et al., 2008), and a high incidence of repetitive DNAs is prevalent in highly rearranged genomes (Chumley et al., 2006; Guisinger et al., 2011; Weng et al., 2014). While dysfunction of DNA replication, recombination and

repair (DNA RRR) systems has been suggested as a potential cause for these unusual phenomena, this hypothesis has not been tested (Jansen et al., 2007; Guisinger et al., 2008; Sloan et al., 2012; Weng et al., 2014).

Plastid genomes of angiosperms do not encode any genes for DNA RRR proteins to maintain genome stability (Bock, 2007; Ruhlman and Jansen, 2014). Genes for these proteins are encoded in nuclear genome and imported into plastids. Various DNA RRR genes have been verified experimentally (Day and Madesis, 2007; Maréchal and Brisson, 2010; Boesch et al., 2011). For plastid DNA replication, two plastid targeted DNA polymerases, *AtDNA I-like DNA polymerase* A (*AtPolI-like A*) and *AtPolI-like B* have been characterized in *Arabidopsis* (Mori et al., 2005). *AtPolI-like B* is upregulated after DNA damage caused by H_2O_2, suggesting a potential role in DNA repair (Mori et al., 2005). Replication protein A (RPA), a protein complex composed of three subunits (RPA70, RPA32, RPA14), is involved in plastid DNA replication (Hübscher et al., 1996; Kornberg and Baker, 2005). Multiple types of RPA are present in rice and the type-A RPA complex (OsRPA70a-OsRPA32-2-OsRPA14) is targeted to plastids (Ishibashi et al., 2006). Gyrase, previously characterized to be involved in untangling plastid DNA after replication/recombination, is critical for plastid DNA maintenance (Ye and Sayre, 1990; Cho et al., 2004). Two types of gyrase (*GYRA* and *GYRB*) are known in angiosperms, and plastid targeted genes for both types are present in *Arabidopsis* and *Nicotiana benthamiana* (Cho et al., 2004; Wall et al., 2004). For DNA recombination, *chloroplast mutator* (*CHM*/*MSH1* or *MSH1*), a homolog of the bacteria gene involved in DNA recombination and mismatch repair (MMR), is known to regulate DNA recombination and genome stability (Xu et al., 2011). Two homologs of *MSH1* (MMR-Muts and MMR-Muts-type2) with plastid subcellular localizaton were identified experimentally (Carrie et al., 2009). *RecA* is another well studied protein family essential for various pathways involved in homologous recombination (Cox, 2007). Three copies of *RecA* genes are present in *Arabidopsis* and two of them (*RecA1* and *RecA2*) are targeted to plastids (Shedge et al., 2007). Recent studies in *Physcomitrella patens* identified a homolog of bacterial *RecG* helicase, and revealed that *RecG* is critical in

maintaining organelle genome stability (Odahara et al., 2015). Two proteins *DRT111* (DNA-damage-repair-toleration) and *DRT112* from *Arabidopsis* were shown to partially restore DNA recombination proficiency and damage-resistance in *E. coli* mutants, and both of them were predicted to be targeted to plastids, suggesting a potential role in DNA recombination/repair in plastid genomes (Pang et al., 1993). Knockout studies in *Arabidopsis* also suggested that whirly protein families are involved in maintaining plastid genome stability by suppressing DNA recombination within short repeats, and two proteins from this family (*why1* and *why3*) are targeted to plastids (Maréchal et al., 2009). Mutants with reduced expression of *RAD51-1* and *RAD52-2* (or *ODB1* and *ODB2*) showed decreased levels of intrachromosomal recombination, and only *RAD52-2* is targeted to plastid (Samach et al., 2011). Deletion of *OSB1* (organellar single-stranded DNA binding protein) has been shown to increase recombination between repeats in the mitochondrial genome, and two homologs *OSB2* (or *PTAC9*) and *OSB3* are targeted to plastids (Zaegel et al., 2006; Maréchal and Brisson, 2010). *NTH1* and *NTH2*, which encode DNA glycolases, are involved in base excision repair in the plastid genome (Gutman and Niyogi, 2009). Two plastid targeted homologs of *RecQ* helicase, which is involved in maintaining genome stability, have been identified in rice (Saotome et al., 2006).

Plastid-targeted DNA RRR genes provide candidates for investigating the correlation between alterations in DNA RRR systems and the increase of plastid genome complexity in Geraniaceae. Previous studies (Jansen et al., 2007; Guisinger et al., 2008, 2011; Weng et al., 2014) have hypothesized that the highly rearranged plastid genomes and accelerated *dN* in this family are due to aberrant DNA RRR mechanisms. In this study, the complexity of 27 Geraniales plastid genomes was estimated using several independent metrics. Transcriptomes for the same 27 species (Zhang et al., 2015) were mined to extract nuclear encoded, plastid targeted DNA RRR genes and genes with different subcellular locations (plastid, mitochondrial and other) as negative controls. An analysis of correlation between nucleotide substitution rates of DNA RRR, control genes and plastid genome complexity was performed. The identified correlation between DNA

RRR genes and genome complexity, and the connection between functions of the DNA RRR genes and genome complexity metrics supported the hypothesis. Correlations identified between some control genes and genome complexity were unexpected. Comparison of Geraniales and Brassicales nuclear genes with different subcellular locations identified significant acceleration of *dN* in the DNA RRR and NUCP genes in Geraniales. The significant acceleration of *dN* of NUCP genes in Geraniales provides a possible explanation for the correlation between *dN* of NUCP genes and genome complexity. Possible causes of the acceleration of *dN* of NUCP genes are discussed.

RESULTS

Plastid Genome Sequencing

Illumina paired end and PacBio reads were generated and assembled for two species of *Monsonia* (See Appendix Table 4.1 for read information). The inverted repeat (IR) was absent from the plastid genomes of *M. emarginata* (134,416bp) and *M. marlothii* (156,877bp). Both genomes contained 107 different genes, including 74 protein coding genes, 29 transfer RNA genes and four ribosomal RNA genes. Introns were detected in 14 different genes in both genomes. The GC content for *M. emarginata* and *M. marlothii* were 40.2% and 39.3%, respectively.

Genome Complexity

Four measures of genome complexity (genome rearrangements, repeat content, evolutionary rates and insertions and deletions (indels)) were estimated by comparing the plastid genomes of 27 Geraniales species (Appendix Figure 4.1) to the *Arabidopsis* reference genome (Table 4.1), which has the same gene order as the ancestral genome of Geraniales (Weng et al., 2014). For estimating genome rearrangements, locally collinear blocks (LCBs) were identified by either Mauve for multiple genome alignments of 27 species of Geraniales and *Arabidopsis* or by pairwise genome alignment between each of the Geraniales species and *Arabidopsis*. An additional measure of genome rearrangement

was based on synteny of 63 shared genes across the 27 species of Geraniales and *Arabidopsis*.

Two metrics of genome rearrangement, break point (BP) and inversion (IV) distance, were estimated based on the order of LCBs or the synteny of shared genes (Table 4.1, Appendix Figure 4.2). Within the same species, the estimated BP and IV distance was similar. The BP and IV distances of Geraniaceae species ranged from 3 - 17, while the outgroup species (*Francoa sonchifolia*, *Melianthus villosus*), which have similar plastid genome organization to *Arabidopsis*, had much smaller BP and IV distances (0 - 2, Table 4.1, Appendix Figure 4.2). The largest IV distance was 17 between *Geranium maderense* and *Arabidopsis*, and the largest BP distance was 17 between *G. maderense* and *Arabidopsis* and between *Erodium chrysanthum* and *Arabidopsis*. Multiple clade-specific increases of BP and IV distances were observed in Geraniaceae (*Hypseocharis*, *Geranium*, *Monsonia*, *Pelargonium* C2 clade and *Erodium* clade I, Appendix Figure 4.1).

Two classes of repeats, small dispersed repeats (SDR) and tandem repeats, were identified. The greatest numbers of SDRs (752) and tandem repeats (107) were both found in *G. incanum* (Table 4.1, Appendix Figure 4.3A). The fewest repeats were found in the outgroup species for both SDR *(F. sonchifolia,* 36; *M. villosus,* 44) and tandem repeats (*F. sonchifolia,* 19; *M. villosus,* 15). Clade-specific increases in repeat content were observed in Geraniaceae (*Geranium*, *Monsonia*, *Pelargonium* C2 clade and *Erodium* clade I, Table 4.1, Appendix Figures 4.1 and 4.3A).

The number of nucleotide insertions and deletions (indels) were estimated in three different regions of the plastid genome, protein coding and ribosomal RNA (CDR), intron and intergenic (IG) regions. Among the 27 species of Geraniales and *Arabidopsis*, 63 CDRs, 8 intron regions and 24 IGs were identified as shared and analyzed (Appendix Table 4.2). The greatest number of IG indels (364) was identified in *G. phaeum* and *G. incanum* (Table 4.1, Appendix Figure 4.3B). The greatest number of CDR indels (123) was found in *M. emarginata*, and *E. chrysanthum* had the most intron indels (185, Table 4.1, Appendix Figure 4.3B). Fewer CDR indels (33 to 37) were identified in the

outgroup species compared to Geraniaceae species (73 to 123) and the number of intron and IG indels was similar across all Geraniaceae species (Table 4.1, Appendix Figure 4.3B).

Nucleotide substitution rates

Synonymous (*dS*) and nonsynonymous (*dN*) substitution rates in the plastid genome were estimated using the concatenated alignment of 59 shared plastid encoded protein coding genes (Table 4.1). Lower *dN* (0.039 to 0.041) and *dS* (0.35 to 0.37) were detected in the outgroup species compared with Geraniaceae (*dN*, 0.065 to 0.094; *dS* 0.43 to 0.62). The highest *dN* (0.094) was detected in *E. chrysanthum* and the highest *dS* (0.63) was detected in both *E. chrysanthum* and *E. gruinum* (Table 4.1). Multiple clade-specific accelerations of *dN* were detected in the *Pelargonium* C clade, and in *Erodium* clade I (Appendix Figure 4.1). Clade-specific acceleration of *dS* was only observed in *Erodium* clade I, while other Geraniaceae species had a similar *dS* value (0.5).

The value of *dS* and *dN* for 12 nuclear encoded DNA RRR genes, 90 nuclear encoded control genes and 59 plastid encoded genes were estimated based on the MAFFT alignments (See Supplemental Data File 4.1 for alignments). The 90 control genes were divided into three groups based on their subcellular localization (plastid 30, mitochondrial 30, other 30, see Supplemental Data File 4.2). Of the five gene groups, DNA RRR genes, plastid encoded genes, nuclear encoded plastid targeted genes(NUCP), nuclear encoded mitochondrial targeted genes (NUMT) and other nuclear encoded genes (NUOT), NUMT had the highest median value of *dN* (0.19) and *dS* (2.14, Figure 4.1). NUCP had the lowest *dN* (0.17), while NUOT had the lowest *dS* values (2.00). Both *dN* (0.061) and *dS* (0.52) of plastid encoded genes were much lower than nuclear genes (Figure 4.1). The average *dN* of nuclear genes (0.18) was approximately three times higher than plastid encoded genes (0.061), and *dS* was about four times higher (2.09 versus 0.52).

Correlation of Genome Complexity and Nucleotide Substitution Rates

To evaluate the correlation between measures of genome complexity and evolutionary rates, each metric was calculated for 27 species as a vector. Prior to the correlation test between genome complexity and substitution rates, correlation tests among genome complexity measures were performed to eliminate those that were highly correlated. Analysis of the six measures of genome rearrangement (Appendix Table 4.3) indicated that four of the six metrics were highly correlated (correlation coefficient >0.95). This resulted in the elimination of all but two genome rearrangement metrics, IV distance based on either LCBs from pairwise genome alignment or synteny of shared genes. Correlations among all remaining measures of genome complexity were calculated (Appendix Table 4.4). The highest correlation was observed between CDR indels and *dN* (0.92). High correlation was also observed among measures of genome rearrangement (0.88), and *dN* and *dS* (0.83). No other strong correlations were detected among the remaining measures (from -0.24 to 0.77).

The correlation between nine measures of plastid genome complexity and evolutionary rates of 12 nuclear encoded DNA RRR genes was evaluated and 90 nuclear encoded genes with different subcellular locations were included as negative controls (Figure 4.2). Significant correlation between *dN* and genome complexity was identified for various genes of NUCP, NUMT and DNA RRR gene groups (Figure 4.2A). Significant correlation (Pearson correlation test, P value < 0.05 after multi-hypothesis correction) of *dS* and genome complexity was detected in two pairs; *dS* of *recb* and Intron indels, and *dS* of *mmr2* and *dS* CP (Figure 4.2B). The number of genes with *dN* significantly correlated to genome complexity measures is summarized in Figure 4.3. Three genes were identified with significant correlation to genome complexity in both the NUMT and RRR gene groups (Figure 4.3). The greatest number of genes (10 out of 30) showing significant correlation was in the NUCP group (Figure 4.3). Detailed information on the gene and genome complexity pairs showing significant correlation is shown in Appendix Table 4.5, and results of the correlation analysis for all comparisons are shown in Supplemental Data Set 4.3. No significant correlations were identified

using measures of genome rearrangement or repeat content. Of the measures of genome complexity showing significant correlation, the highest number (8) was in CDR indels. Of the genes showing a significant correlation, the greatest number was identified for *rh22* and *uvrB*. Both genes were positively correlated with the same complexity measures: CDR and intron indels, and plastid genome *dN* (Appendix Table 4.5). Although three NUMT genes showed a significant correlation with various measures of genome complexity, two of them are targeted to both plastids and mitochondria (Appendix Table 4.5, Supplemental Data Set 4.2).

Significant correlations between *dN* and genome complexity were identified not only in DNA RRR genes, but also in NUCP control genes (Figures 4.2 and 4.3). This result was not expected, therefore another angiosperm family was investigated to compare nucleotide substitution rates in plastid and nuclear encoded genes. The same nuclear and plastid gene sets from 10 species of Brassicales were assembled from published data (Appendix Table 4.6). Five DNA RRR, 28 NUCP, 19 NUMT, 20 NUOT and 59 plastid encoded genes common to both Geraniales and Brassicales were identified (Appendix Table 4.7). Because the Geraniales data set was larger both in terms of numbers of species and genes, a subset of taxa and genes was utilized for rate comparisons of different gene groups (see METHODS). Rate acceleration of *dN* was observed in all gene groups in Geraniales compared to Brassicales (Figure 4.4A), and significant acceleration (student t test, $P < 0.05$) was observed in NUCP, RRR and plastid encoded genes. Significant acceleration of *dS* in Geraniales compared to Brassicales was identified in plastid encoded genes but not in any nuclear encoded genes (Figure 4.4B).

DISCUSSION

Nucleotide Substitution Rates in Geraniaceae

Synonymous (*dS*) and nonsynonymous (*dN*) substitution rates of nuclear and plastid encoded genes were estimated in Geraniaceae. The ratio of *dS* of nuclear and plastid encoded protein coding genes was approximately four to one (Figure 4.1), which

is in the range of previous studies from 4:1 to 6:1 (Wolfe et al., 1987; Gaut, 1998; Drouin et al., 2008). A recent study in *Arabidopsis* compared *dS* of 12 nuclear encoded protein coding genes and three plastid noncoding regions, and a similar ratio (5:1) between *dS* of nuclear and plastid DNA was detected (Huang et al., 2012). The ratio of *dN* for nuclear and plastid encoded genes in Geraniaceae was approximately 3:1 (Figure 4.1), which falls in the range of previous estimates in angiosperms of 1:1 to 5:1 (Gaut, 1998; Drouin et al., 2008). A similar ratio between *dN* (4.5:1) of nuclear and plastid encoded genes was also observed in diatoms (Sorhannus and Fox, 1999). Across angiosperms the ratio of *dN* between nuclear and plastid genome fluctuates more (1:1 to 5:1) than that of *dS* (4:1 to 6:1). Lineage specific effects could be one factor causing *dN* variation in different plant families, as previous studies suggested that nonsynonymous substitution rate is more likely to be affected by this phenomenon (Wolfe et al., 1987; Wolf, 2012).

Rates of nuclear genes with different subcellular locations, plastid (NUCP), mitochondrion (NUMT) and other (NUOT) were very similar (*dN* 1.0:1.0:1.1, *dS* 1.1:1.1:1.0, Figure 4.1) in Geraniaceae. To the best of our knowledge, this is the first family-wide analysis comparing rates of nuclear encoded genes that are targeted to different subcellular locations. Studies in *Silene* compared rates of ribosomal proteins with different subcellular locations (plastid, mitochondrial, cytosolic) using three species pairs (Sloan et al., 2014). Similar results in *dS* (1.0:1.0:1.0) but not *dN* (14.0:12.0:1.0) were detected. Unlike Geraniaceae, the *dN* of cytosolic targeted ribosomal proteins was much lower than their organellar targeted counterparts in *Silene*. There are two possible explanations for this difference. First, since NUOT genes could be targeted anywhere except to the plastid and mitochondrion (e.g. endoplasmic reticulum), a much higher *dN* in the non-cytosolic targeted NUOT genes, could elevate the average *dN* of that group such that they appeared to be similar to that of other gene groups in Geraniaceae. Second, assuming the ratio of *dN* of different gene groups in Geraniaceae (1.0:1.0:1.1) is the ancestral state for both families, the ratio of *dN* of the three gene groups (14:12:1) in *Silene* could be caused by lineage specific acceleration of *dN* in both plastid- and mitochondrial-targeted gene groups. The *Silene* species involved in rate comparisons

have fast-evolving plastid and mitochondrial encoded genes, which could be the factor causing the increase of rates of the nuclear encoded organellar targeted ribosomal proteins (Sloan et al., 2014). Correlated acceleration of *dN* but not *dS* of genes encoding interacting subunits from different subcellular compartments has been documented in several studies including Geraniaceae and *Silene* (Barreto and Burton, 2013; Sloan et al., 2014; Zhang et al., 2015), making the second explanation more plausible.

Correlation of Genome Complexity and Nucleotide Substitution Rates

Previous studies in Geraniaceae revealed highly elevated nucleotide substitution rates and extensive genome rearrangements (Chumley, 2006; Blazier et al., 2011; Guisinger et al., 2011; Weng et al., 2014). One possible explanation for the high levels of genome complexity is dysfunction of the nuclear encoded plastid targeted DNA replication, recombination and repair (RRR) genes (Guisinger et al., 2008; Weng et al., 2014). If this were the case, correlated changes in substitution rates of DNA RRR genes and genome complexity would be predicted. Of the 12 plastid targeted DNA RRR genes investigated, *dN* of three genes (*gyra*, *uvrB* and *why1*) show a significant correlation with the number of indels of protein and ribosomal RNA coding genes (CDR) and introns, and *dN* of the plastid genome (*dN* CP) (Figure 4.2 and Appendix Table 4.5). Of the three DNA RRR genes, *uvrB*, known to be involved in the plastid nucleotide excision repair system (Hsu et al., 1995), was significantly correlated to both indels of CDR/introns and *dN* CP (Figure 4.2), suggesting that changes in the nucleotide excision repair system may be involved in the observed complexity of Geraniaceae plastid genomes. The detection of a correlation between *why1* and indels of CDR, and between *gyra* and *dN* CP is unexpected because both of the genes are involved in DNA replication/recombination directly, while both indels of CDR and *dN* CP are complexity measures based more on nucleotides. *Why1* is known to affect plastid genome stability (Maréchal et al., 2009), and *gyra* has been suggested to be involved in untangling DNA after replication and recombination (Cho et al., 2004). There are other possible explanations for the correlations between *dN* of DNA RRR genes and genome complexity. First, a common

factor affected both the increased rates of nucleotide substitutions of the DNA RRR genes (*uvrB*, *why1*, *gyra*) and plastid genome complexity. Second, the dysfunction of DNA RRR genes (*why1*, *gyra*) led to the increase in genomic rearrangements, indels and *dN* CP. The first scenario does not explain why the common factor only affects certain measures of genome complexity, i.e., *dN* CP but not *dS* CP (Appendix Table 4.5). On the other hand, the underlying connection between function of DNA RRR genes and plastid genome complexity provides a biological basis for the correlation, making the second explanation more likely.

Estimates of correlation between *dN* of DNA RRR genes and plastid genome complexity included 90 nuclear encoded genes with three different subcellular locations as negative controls. Unexpected significant correlations between genome complexity in both NUCP and NUMT genes were detected (Figure 4.3). Two out of the three NUMT genes that showed a correlation are dually targeted to plastids, suggesting that this result is likely caused by the plastid component for two of the genes. Significant correlations with NUCP genes involved genes with a wide diversity functions (i.e. RNA binding, cytochrome c biogenesis and cell cycle control, Supplemental Data File 4.2), arguing against the possibility that shared functional constraints maintain the correlation for the control genes. None of the DNA RRR or nuclear control genes show both correlated *dN* and *dS* with the genome complexity, suggesting that the correlations are not due to the effect of a global background mutation rates (Figure 4.2B). To investigate the possible cause of the correlation between the control genes and genome complexity, we compared rates of the same genes that are targeted to different subcellular locations (NUCP, NUMT, NUOT, RRR, and plastid) between Geraniales and Brasscicales. Brassicales have highly conserved plastid genomes with only a single genomic rearrangement (loss or *rps16*, Figure 4.3 in Ruhlman and Jansen, 2014). Significant acceleration of *dN* for Geraniales was observed only in NUCP, DNA RRR and plastid encoded gene groups, relative to Brassicales (Figure 4.4A). The acceleration of *dN* in NUCP and DNA RRR genes provides an explanation for the observation of correlation of the *dN* of these gene groups with genome complexity. The lack of significant acceleration of *dS* for NUCP

and DNA RRR genes (Figure 4.4B) argues against the explanation that the correlation of *dN* is due to differences in background mutation rates between the orders.

Due to the unusually high levels genome rearrangement and elevated rates of nucleotide substitutions (Chumley et al., 2006; Guisinger et al., 2008; Blazier et al., 2011; Guisinger et al., 2011; Weng et al., 2014), Geraniaceae are an attractive system to study nuclear-organelle genome coevolution. Previous studies hypothesized that these phenomena may be the result of alterations in DNA repair and recombination mechanisms (Guisinger et al., 2008, 2011; Weng et al., 2014). The identification of a significant correlation between rates of DNA RRR genes and some measures of genome complexity, and the connection between the functions of DNA RRR genes and specific metrics of genome complexity support this hypothesis. The identification of a correlation between NUCP and NUMT control genes and genome complexity is unexpected. Two possible explanations for the increase in *dN* of these control genes include: that they are caused by the increase of plastid genome complexity, which was caused by dysfunction of DNA RRR systems; and that the increase of *dN* of the control genes is caused by a factor other than the plastid genome complexity, and even possibly the same factor that causes the increase of *dN* of DNA RRR genes. Further study is needed to disentangle these scenarios. Rate comparisons between Geraniales and Brassicales indicated that the acceleration of *dN* of the NUCP control genes in Geraniales may explain the correlation with the genome complexity.

METHODS

DNA isolation, whole genomic sequencing and assembly

Total genomic DNA of *Monsonia emarginata* and *M. marlothii* was isolated from young leaf tissues and was sequenced with Illumina HiSeq 2000 at the University of Texas Genomic Sequencing and Analysis Facility (GSAF) as described in Weng et al., (2014). For each species, approximately 60 million 100 bp, paired end reads were generated from 800 bp insert libraries. A 10 kb SMRT cell library was constructed with

PacBio RS II sequencing and one cell of sequence data was generated at the University of Florida Interdisciplinary Center for Biotechnology Research. All PacBio reads were corrected with the long read error correction tool (LSC) (http://www.healthcare.uiowa.edu/labs/au/LSC/LSC_manual.html) using Illumina paired end reads.

De novo assembly of the Illumina paired end reads was performed with Velvet v 1.2.07 (Zerbino and Birney, 2008) with various parameters (kmer 79 to 95 bp and coverage 200, 500 and 1000) settings to optimize contig length. Corrected PacBio reads (Appendix Table 4.1) were used to join Velvet assemblies and create scaffolds. Both Illumina and PacBio reads were mapped back to the scaffolds using Geneious (Biomatters, http://www.geneious.com/) to fill gaps.

RNA isolation and transcriptome sequencing and assembly

Total RNA was extracted from newly emergent leaves of 27 species of Geraniales, and four different tissue types (emergent and expanded leaves, roots and flowers) of *Pelargonium* x *hortorum* as described in Zhang et al. (2015). Transcriptome sequencing was performed with HiSeq 2000 at UT GSAF. Sequence data was preprocessed and assembled as described in Zhang et al. (2013, 2015).

Plastid genome complexity analysis

Plastid genomes of 27 species of Geraniales and *Arabidopsis* were used for the genome complexity analysis, 26 of which were downloaded from NCBI (*Arabidopsis thaliana* NC_000932, *Francoa sonchifolia* NC_021101, *Melianthus villosus* NC_023256, *Hypseocharis bilobata* NC_023260, *Pelargonium nanum* KM527896, *P. citronellum* KM527888, *P. echinatum* KM527891, *P. incrassatum* KM527894, *P. fulgidum* KM527893, *P. cotyledonis* KM459516, *P. australe* KM459517, *P. dichondrifolium* KM459515, *P. exstipulatum* KM527892, *P. myrrhifolium* KM527895, *P. tetragonum* KM527899, *P. transvaalense* KM527900, *P.* x *hortorum* NC_008454, *Geranium maderense*, *G. phaeum*, *G. incanum*, *California macrophylla* JQ031013, *Erodium texanum* NC_014569, *E. chrysanthum*, *E. gruinum* NC_025907, *E. foetidum*, *E. trifolium*

NC_024635). Plastid genomes of the remaining two species (*Monsonia emarginata*, *M. marlothii*) were assembled in this study as described. Different measures of genome complexity (genome rearrangements, repeat content and indels, described below) were estimated by comparing the plastid genomes of 27 species of Geraniales to *Arabidopsis*.

Multiple genome alignment of the 27 species of Geraniales and *Arabidopsis*, and pairwise genome alignment between each species of Geraniales and *Arabidopsis* was performed using the progressive Mauve algorithm (Darling et al., 2010) in Geneious. Shared genes across the 28 species were identified by a custom Python script. The locally collinear blocks (LCBs) identified by Mauve alignment and the order of the shared genes identified by a custom Python script were numbered for genome rearrangement estimation. Two genome rearrangement measures, inversions and break point distances, were estimated by comparing the numbered LCBs and gene order between 27 species of Geraniales with *Arabidopsis*, using Grimm (Tesler, 2002a, 2002b) and the online web server Common Interval Rearrangement Explorer respectively (Bernt et al., 2005).

Each plastid genome was blasted against itself with NCBI-BLAST (BLAST 2.2.28+) using default parameters. One IR was removed from the plastid genomes where two copies were present. Blast results were parsed with a custom Python script to identify small dispersed repeats. Tandem repeats were identified using Tandem Repeat Finder (v 4.07b, Benson, 1999) with default parameters. Variation in repeat content was estimated by subtracting the number of repeats in *Arabidopsis* from those identified in the other 27 species.

Shared protein coding genes, intron regions and intergenic regions among the 28 species were identified and sequences were aligned with MAFFT (Katoh and Standley, 2013) in Geneious. Indels within these regions were calculated by comparing the aligned regions of 27 species in Geraniales to *Arabidopsis* using a custom Python script. Alignments of the shared protein coding genes were concatenated for plastid genome rate estimation.

Evolutionary rate estimation

PAML's codeml (Yang, 2007) was used to estimate synonymous (*dS*) and nonsynonymous (*dN*) substitution rates using the codon frequencies model F3X4. Gapped regions were excluded with parameter "cleandata = 1". Pairwise rates were estimated with parameter "runmode = -2". Twelve nuclear encoded, plastid targeted DNA replication, recombination and repair (RRR) genes were analyzed. Ninety nuclear control genes with different subcellular locations (plastid 30, mitochondrial 30, other 30), which are orthologus to genes in APVO database (Duarte et al., 2010), were used as negative control groups. Fifty-nine plastid encoded genes were extracted from the plastid genome annotations as described in Weng et al. (2014). The accession numbers and descriptions of corresponding genes in *Arabidopsis thaliana* are shown in Supplemental Data File 4.1.

Analysis of correlation between evolutionary rate and genome complexity

Correlation between *dN* or *dS* of each gene and the measures of genome complexity was performed using the original mirror tree method as described in (Pazos et al., 1997; Pazos and Valencia, 2001; Pazos et al., 2005). The evolutionary rates (*dN* and *dS*) of each gene, and each genome complexity measure for the 27 species of Geraniales were collected as a rate or genome complexity vector, respectively. Any correlation between the rate vector and genome complexity vector was estimated with the Pearson correlation test using built in function pearsonr in scipy module of Python. The resulting P values were adjusted using Bonferroni correction to remove the effect of multi-hypothesis testing.

Rate comparisons between Geraniaceae and Brassicaceae

Based on previously published phylogenies of Brassicaceae (Kagale et al., 2014), nine species of Brassicacae (*Arabidopsis lyrata*, *Arabidopsis thaliana*, *Arabis alpina*, *Barbarea verna*, *Brassica napus*, *Brassica rapa*, *Capsella bursa-pastoris*, *Pachycladon cheesemanii* and *Raphanus sativus*) and one outgroup species (Caricaceae *Carica papaya*) were selected. For all species included, either complete or partial plastid

genome and transcriptome information was available. Reciprocal blast searches were used to identify orthologs of DNA RRR and nuclear control genes. Rates for those genes present in both Geraniaceae and Brassicaceae datasets were compared. To avoid bias in the rate comparisons, a similar number (11) of representative species of Geraniales (*Melianthus villosus*, *Hypseocharis bilobata*, *Pelargonium nanum*, *P. exstipulatum*, *P. myrrhifolium*, *P.* x *hortorum*, *Monsonia emarginata*, *Geranium maderense*, *Erodium chrysanthum*, *E. foetidum*) were selected. Genes were separated into five groups based on their functions and subcellular locations (DNA RRR, nuclear-encoded-plastid-targeted control, nuclear-encoded-mitochondrial-targeted control, other nuclear-encoded control, plastid encoded). Rates of the same gene groups from the two families were compared using Student's t test.

Table 4.1. Measures of genome complexity among 27 Geraniales species.

Species[1]	LCBs - p[2]	Gene order[2]	SDR[2]	Tandem[2]	CDR[2]	Intron[2]	IG[2]	*dN* CP[2]	*dS* CP[2]
F. sonchifolia	0	0	36	19	33	132	283	0.039	0.35
Melianthus villosus	2	0	44	15	37	150	299	0.041	0.37
H. bilobata	12	11	88	55	73	158	292	0.065	0.43
P. nanum	5	3	99	47	109	152	331	0.086	0.50
P. citronellum	5	3	84	40	108	153	336	0.087	0.50
P. enchinatum	5	3	71	33	110	155	326	0.088	0.50
P. incrassatum	5	3	125	56	106	152	339	0.088	0.50
P. fulgidum	5	3	116	43	108	151	340	0.087	0.50
P. cotyledonis	6	4	87	28	106	156	337	0.087	0.50
P. australe	6	4	90	42	105	155	338	0.086	0.50
P. dichondrifolium	7	4	105	42	104	154	330	0.087	0.50
P. exstipulatum	6	4	120	47	106	159	342	0.087	0.50
P. myrrhifolium	8	4	79	31	115	154	332	0.091	0.50
P. tetragonum	7	3	72	27	115	158	329	0.090	0.50
P. transvaalense	12	8	490	41	107	156	324	0.087	0.50
P. x *hortorum*	11	12	171	25	120	156	325	0.093	0.51
M. emarginata	9	13	191	58	123	169	333	0.088	0.52
M. marlothii	12	14	160	85	117	173	336	0.085	0.50
G. maderense	13	17	196	38	114	171	310	0.081	0.52
G. phaeum	11	10	378	64	120	174	364	0.079	0.51
G. incanum	14	11	752	107	122	175	364	0.078	0.50
C. macrophylla	3	4	58	40	79	165	311	0.071	0.49
E. texanum	12	12	132	27	110	180	300	0.083	0.52
E. chrysanthum	15	11	122	33	110	185	326	0.094	0.62
E. gruinum	14	9	72	35	111	166	319	0.093	0.62
E. foetidum	4	3	58	46	88	172	306	0.076	0.50
E. trifolium	6	5	60	35	95	170	320	0.077	0.51

[1]Generic name abbreviations: F, *Francoa*; H, *Hypseocharis*; P, *Pelargonium*; M, *Monsonia*; G, *Geranium*; C, *California*; E, *Erodium*.
[2]Genome complexity abbreviations: LCBs - p, IV distance estimated from local collinear blocks; Gene order, IV distance estimated from gene order; SDR, small dispersal repeats; Tandem, repeats estimated from Tandem Repeat Finder (Benson, 1999); CDR, number of indels in coding and ribosomal RNA regions; Intron, number of indels in intron regions; IG, number of indels in intergenic regions; *dN* CP, nonsynonymous substitution rates of the plastid genome; *dS* CP, synonymous substitution rates of the plastid genome.

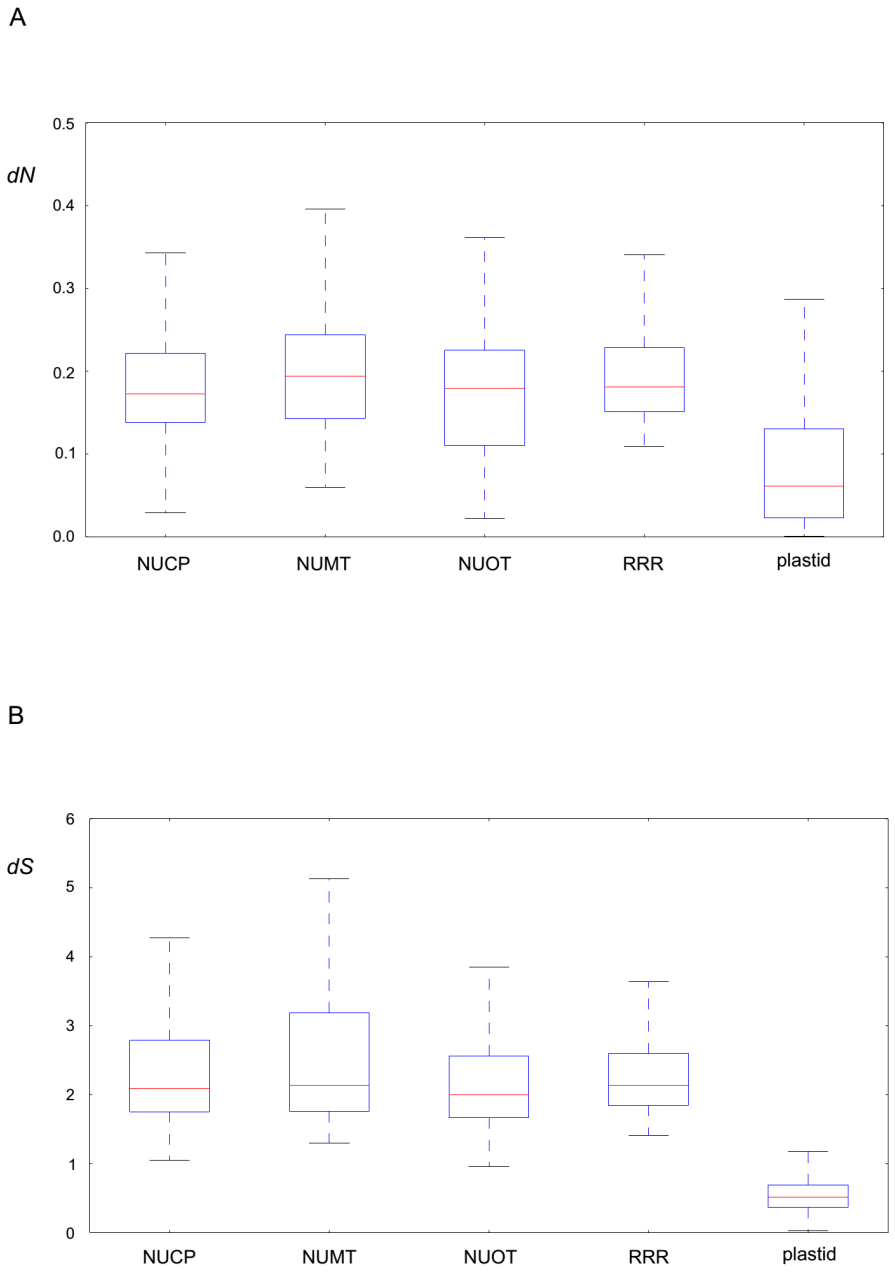


Figure 4.1. Evolutionary rates of DNA RRR, nuclear control and plastid encoded genes.

(A) Nonsynonymous (*dN*) and (B) synonymous (*dS*) substitution rates of different groups are shown. NUCP, nuclear encoded plastid targeted control genes; NUMT, nuclear encoded mitochondrial targeted control genes; NUOT, other nuclear control genes; RRR, DNA replication, recombination and repair genes; plastid, plastid encoded genes.

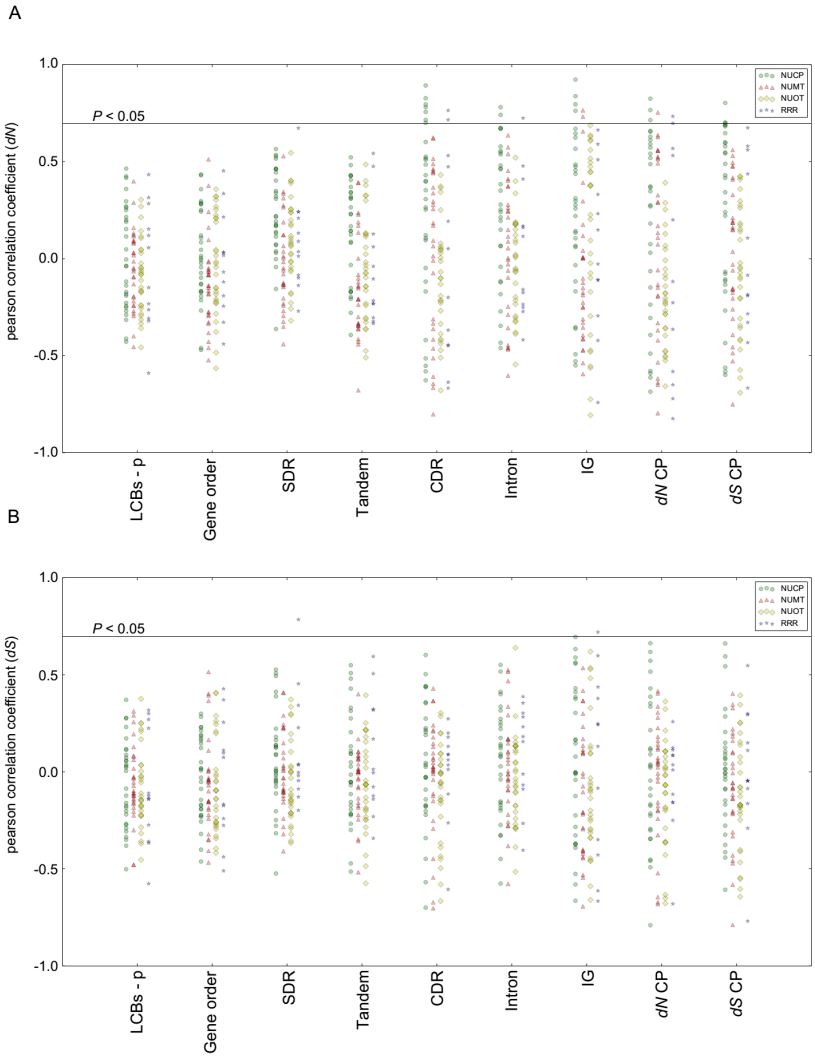


Figure 4.2. Pearson correlation coefficient between gene evolutionary rates and genome complexity.

Correlation of (A) Nonsynnonymous (*dN*) and (B) synonymous (*dS*) substitution rates of genes and genome complexity are shown. NUCP, nuclear encoded plastid targeted control genes; NUMT, nuclear encoded mitochondrial targeted control genes; NUOT, other nuclear control genes; RRR, DNA replication, recombination and repair genes. LCBs - p, IV distance estimated from local collinear blocks; Gene order, IV distance estimated from gene order; SDR, small dispersal repeats; Tandem, repeats estimated from Tandem Repeat Finder (Benson, 1999); CDR, number of indels in coding and ribosomal RNA regions; Intron, number of indels in intron regions; IG, number of indels in intergenic regions; *dN* CP, nonsynonymous substitution rates of the plastid genome; *dS* CP, synonymous substitution rates of the plastid genome.

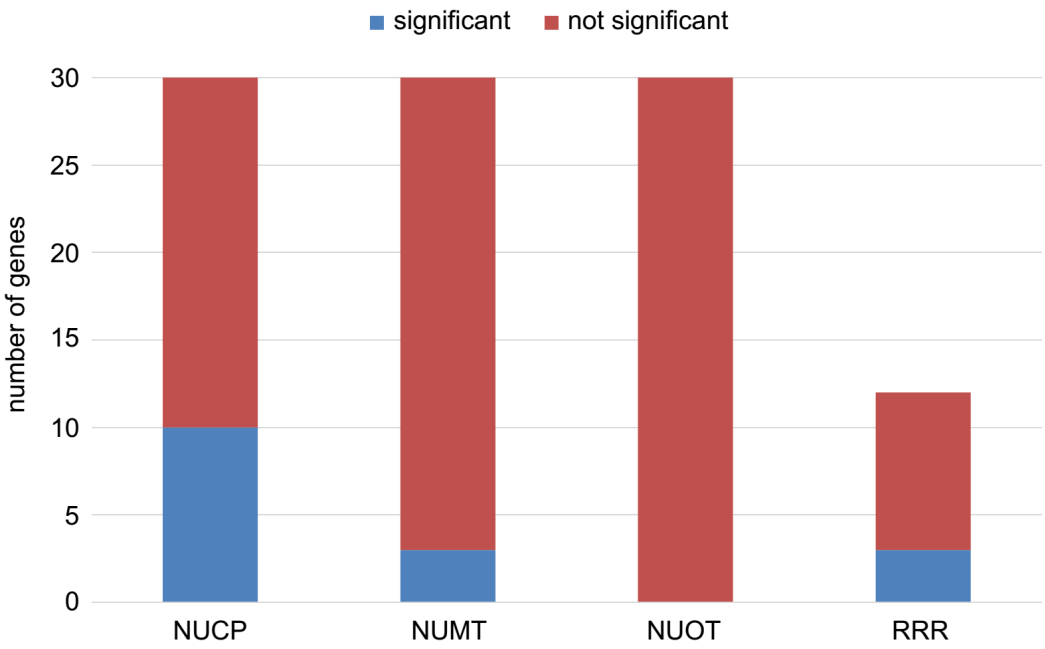


Figure 4.3. Significant correlation of evolutionary rates (*dN*) and genome complexity are identified in NUCP, NUMT and RRR but not NUOT genes.

NUCP, nuclear encoded plastid targeted control genes; NUMT, nuclear encoded mitochondrial targeted control genes; NUOT, other nuclear control genes; RRR, DNA replication, recombination and repair genes.

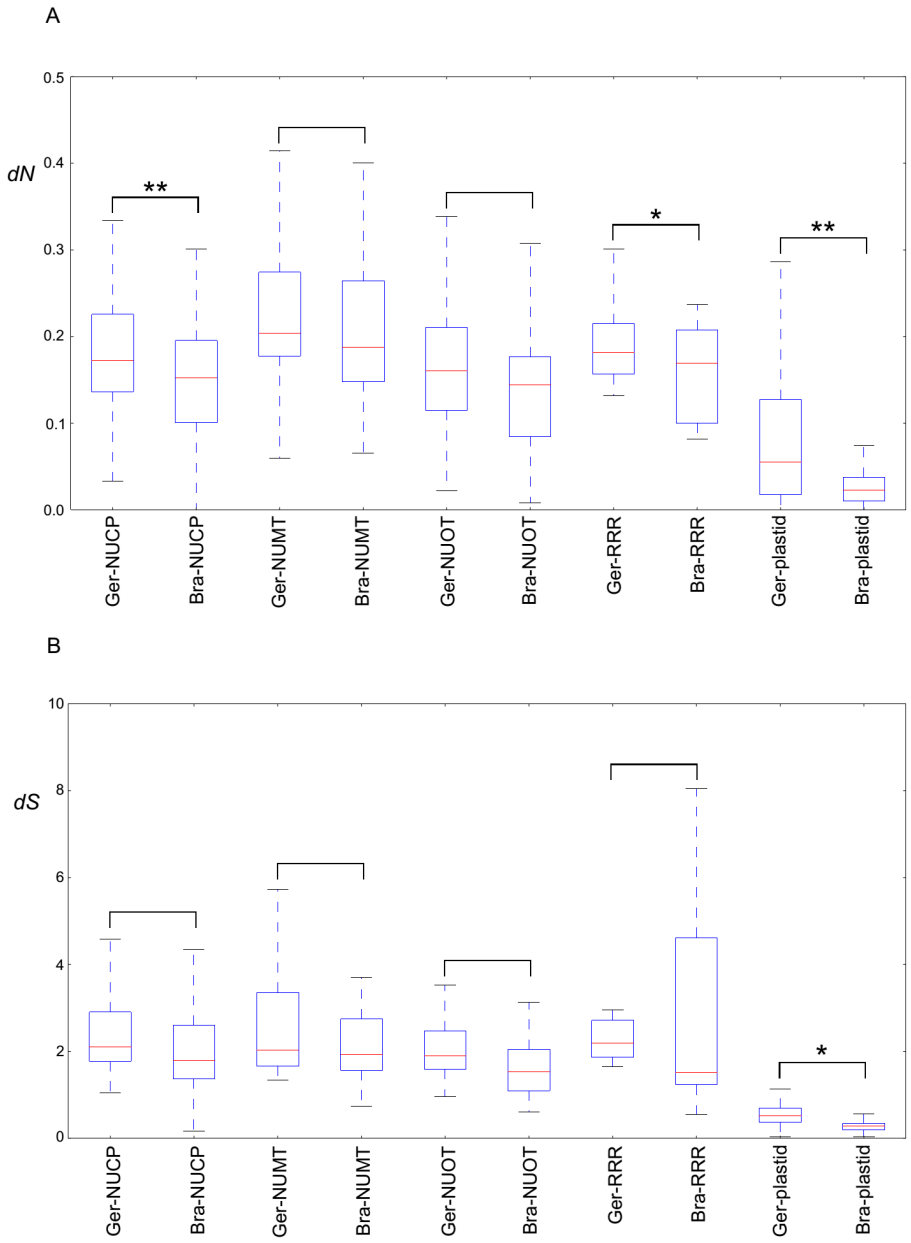
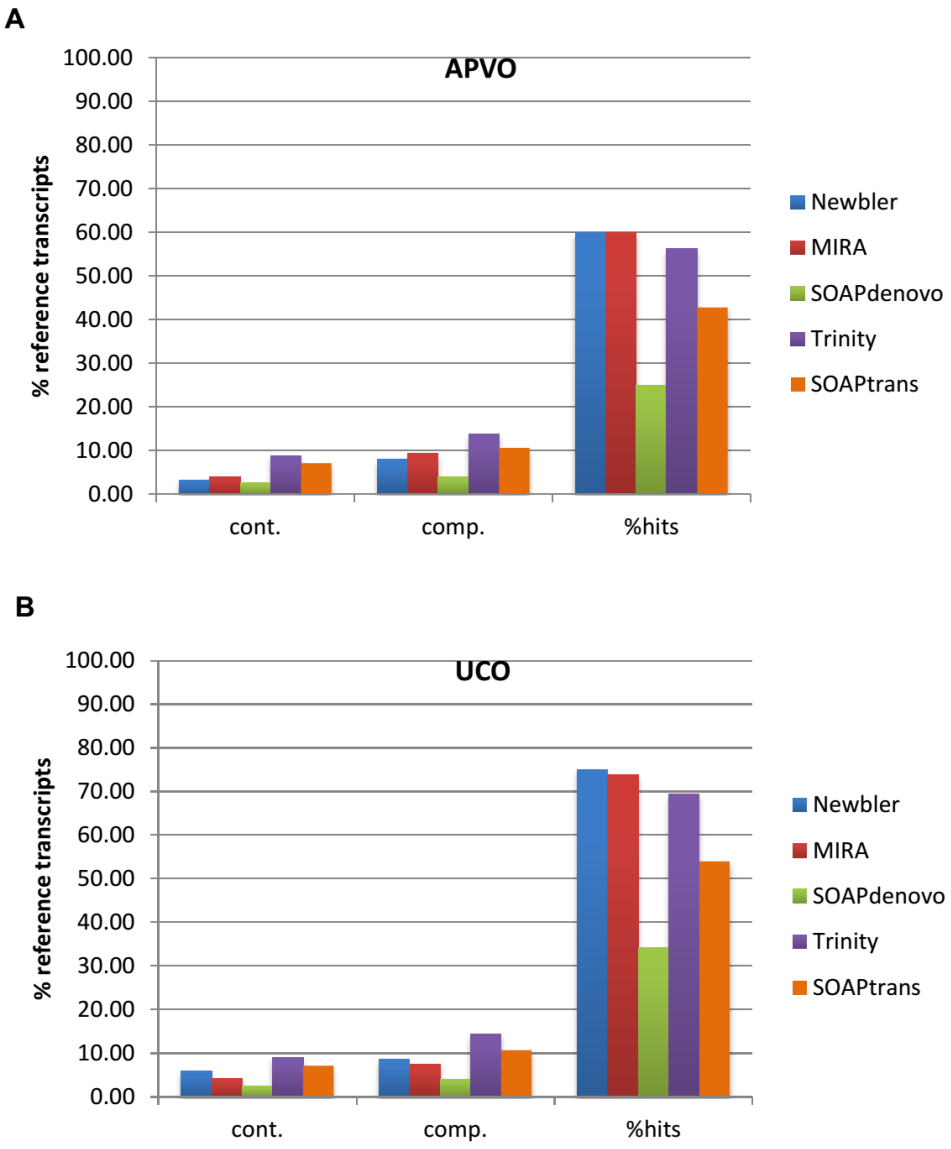


Figure 4.4. Comparison of evolutionary rates between gene groups of Geraniales and Brassicales.

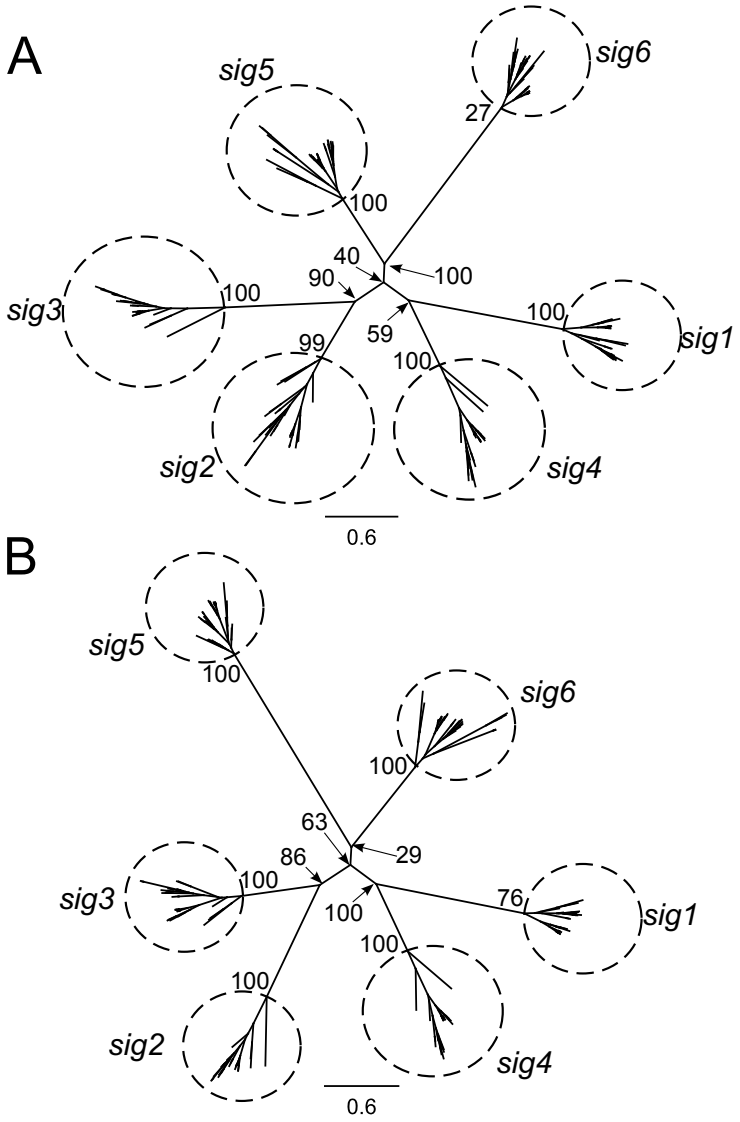
Ger, gene groups are from Geraniales. Bra, gene groups are from Brassicales. NUCP, nuclear encoded plastid targeted control genes; NUMT, nuclear encoded mitochondrial targeted control genes; NUOT, other nuclear control genes; RRR, DNA replication, recombination and repair genes; plastid, plastid encoded genes. Asterisks indicate $p < 0.05$ (*) and $p < 0.001$ (**).

Appendix Data

CHAPTER 2

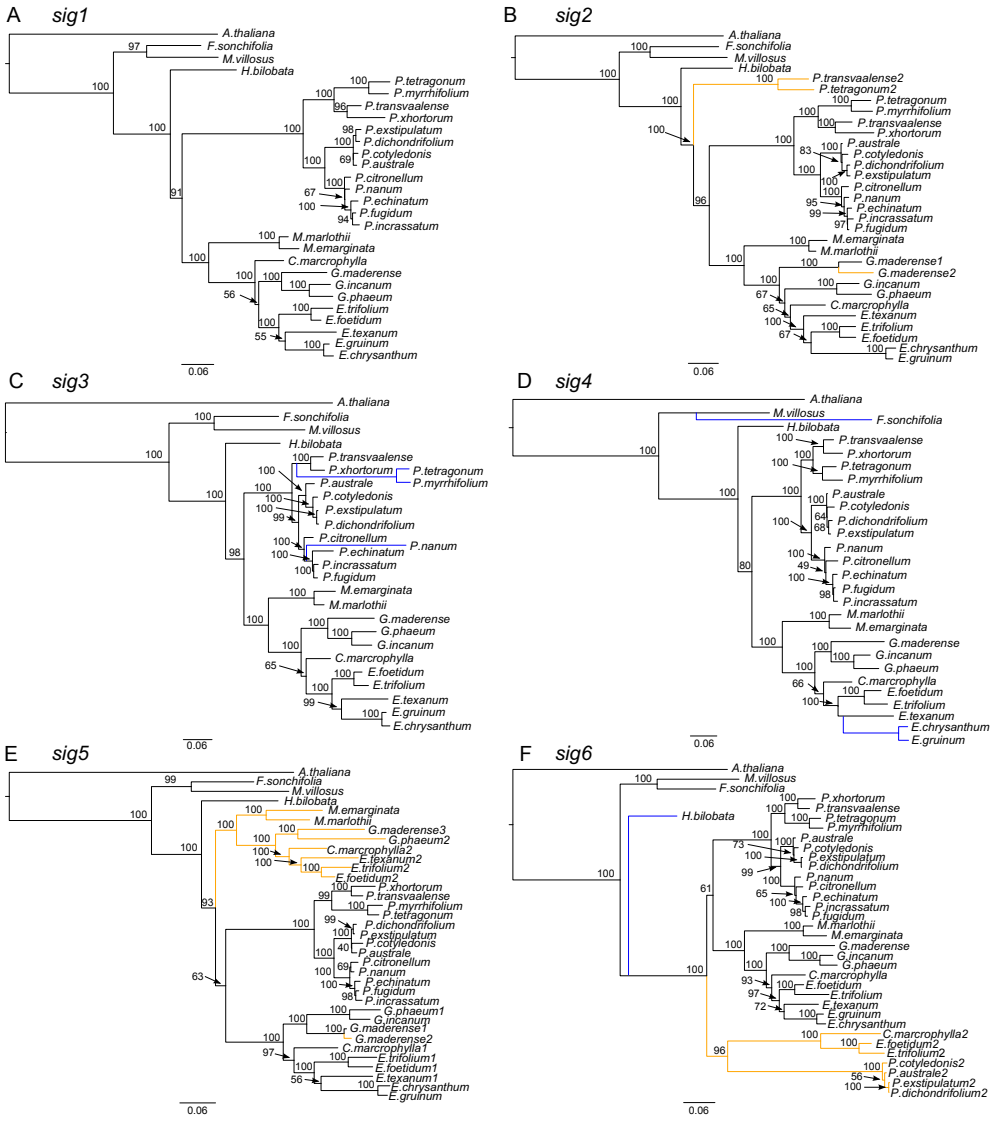


Appendix Figure 2.1 Contiguity and completeness of different protein data sets at E-value 1 E-10 (1/40th) of the Illumina data was used by Trinity).



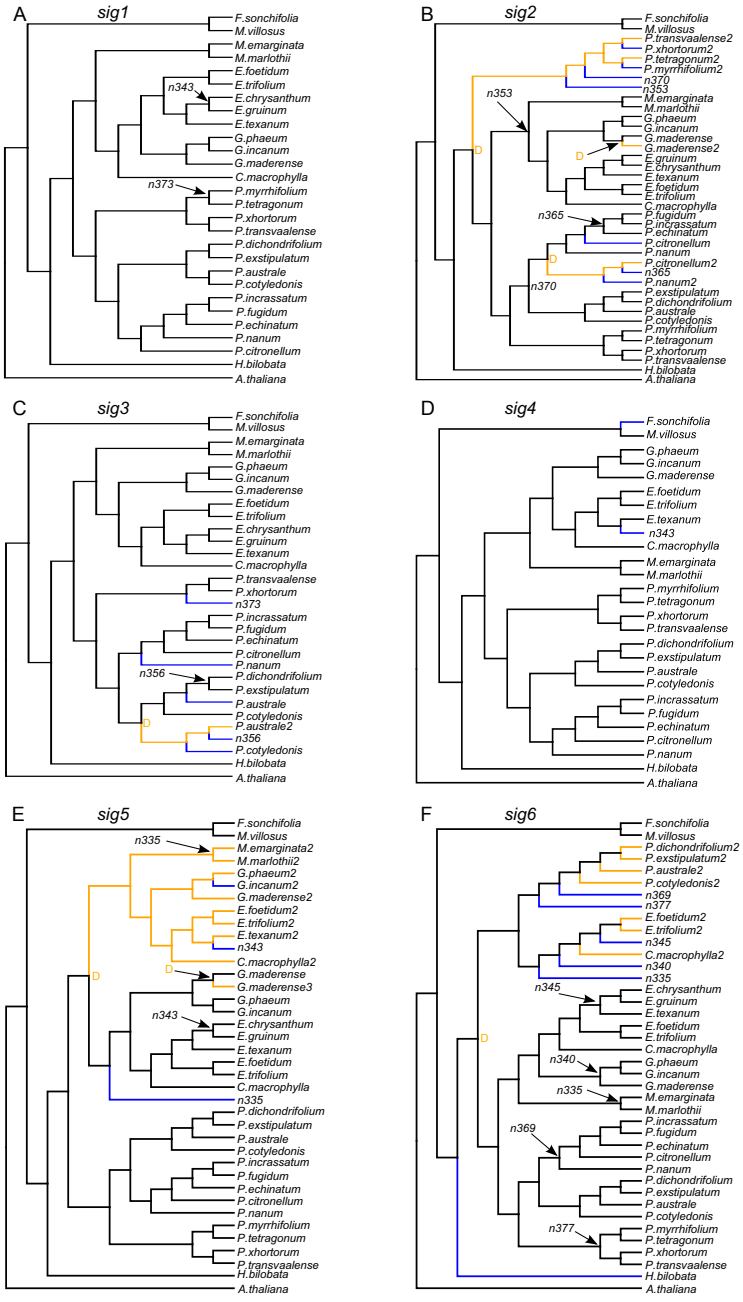
Appendix Figure 3.1. Phylogeny of the sigma factor families in Geraniales and Arabidopsis.

The unrooted amino acid-based maximum likelihood (ML) tree was generated based on alignments of 178 complete sigma factor (SIG) sequences from 27 species of Geraniales. (A) The ML tree (-lnL= -75246.8) was generated based on alignment from MUSCLE. (B) The ML tree (-lnL= -66362.0) was generated based on alignments from ClustalW. Both methods parsed the 178 sequences into the same six subgroups (enclosed in labeled circles). Numbers at nodes are bootstrap support values. Scale bar represents number of amino acid substitutions per site.



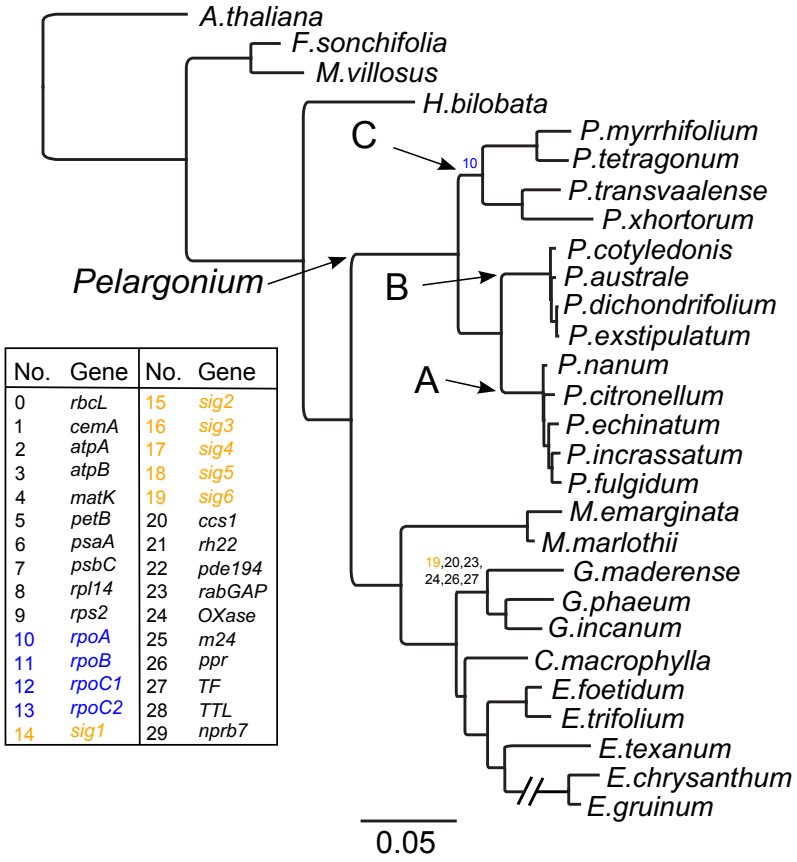
Appendix Figure 3.2. Copy number of the six SIG genes varies across Geraniales.

(A) *sig1*, (B) *sig2*, (C) *sig3*, (D) *sig4*, (E) *sig5* and (F) *sig6*. Blue branches indicate gene losses and orange branches indicate duplicated gene copies. Duplicate copies of genes are shown with a number following the species name. Numbers at nodes represent bootstrap support values. Scale bar represents number of nucleotide substitutions per codon.



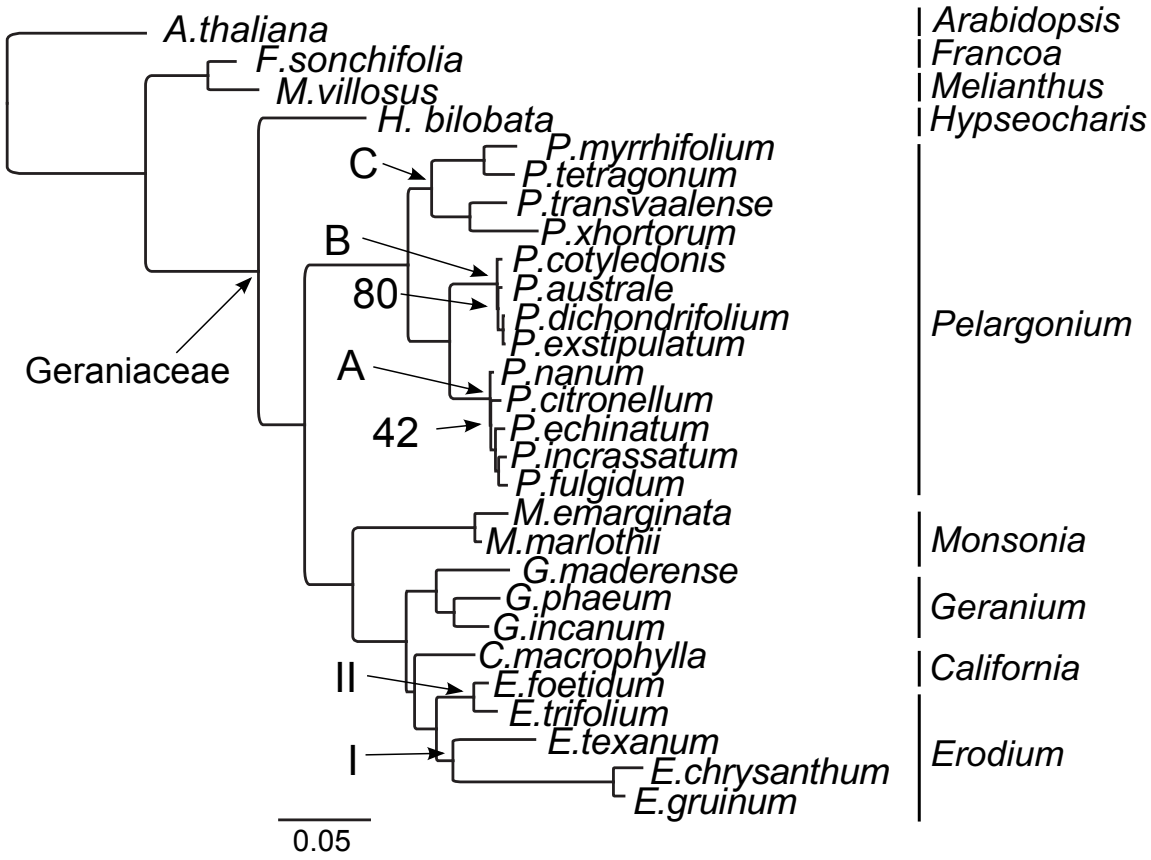
Appendix Figure 3.3. Multiple gene duplication and loss events in Geraniales.

Cladograms inferred by Notung show gene duplication and loss events of (A) *sig1*, (B) *sig2*, (C) *sig3*, (D) *sig4*, (E) *sig5* and (F) *sig6*. Blue branches indicate gene losses and orange branches indicate gene duplications. The duplicated nodes are marked with orange letter "D". The ancestral node involved in a loss event is named as "n" plus the node number (e.g. n353 in cladogram B).



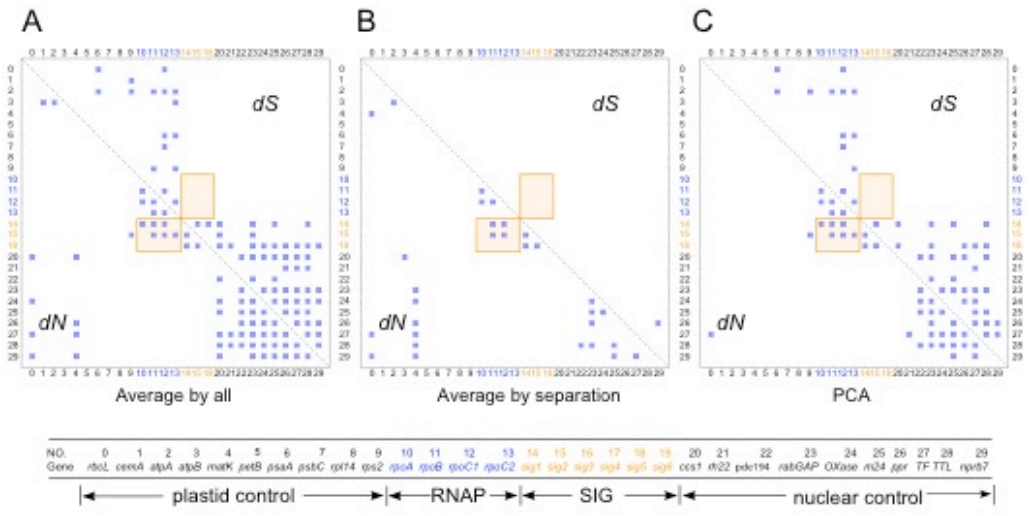
Appendix Figure 3.4. Shared clade-specific synonymous rate (*dS*) acceleration in Geraniaceae.

RNAP and SIG genes are highlighted in blue and orange on the constraint tree (Appendix Figure 3.11), respectively. Numbers at nodes indicate accelerated *dS* in corresponding gene from the key at left. Scale bar represents the number of nucleotide substitutions per codon. The branch leading to species *E. chrysanthum* and *E. gruinum* was interrupted for ease of visualization.



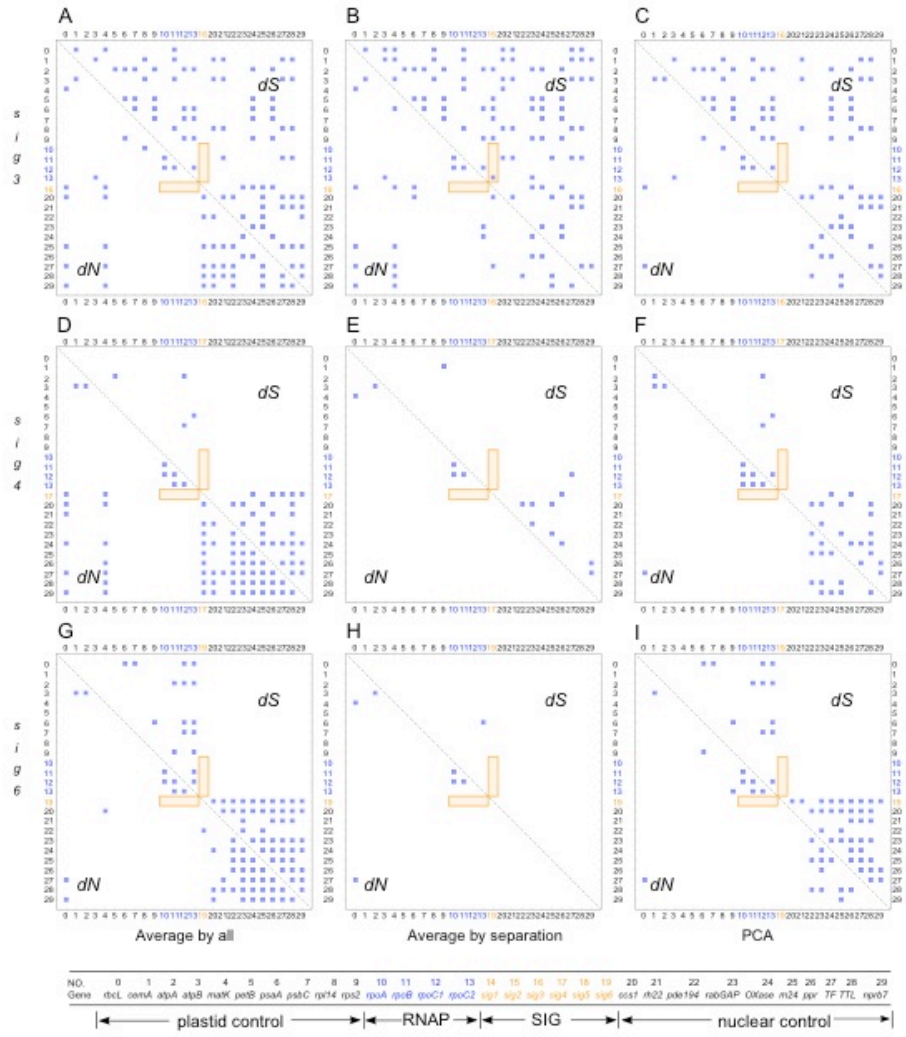
Appendix Figure 3.5. Maximum likelihood tree of 27 species from Geraniales and Arabidopsis.

Twelve plastid genes were used to construct the constraint tree (see Methods). All nodes have bootstrap values of 100 except those shown at nodes on the tree. Generic names are shown on the right and clade designations within each genus are indicated at nodes in the tree. Scale bar indicates numbers of nucleotide substitutions per codon.



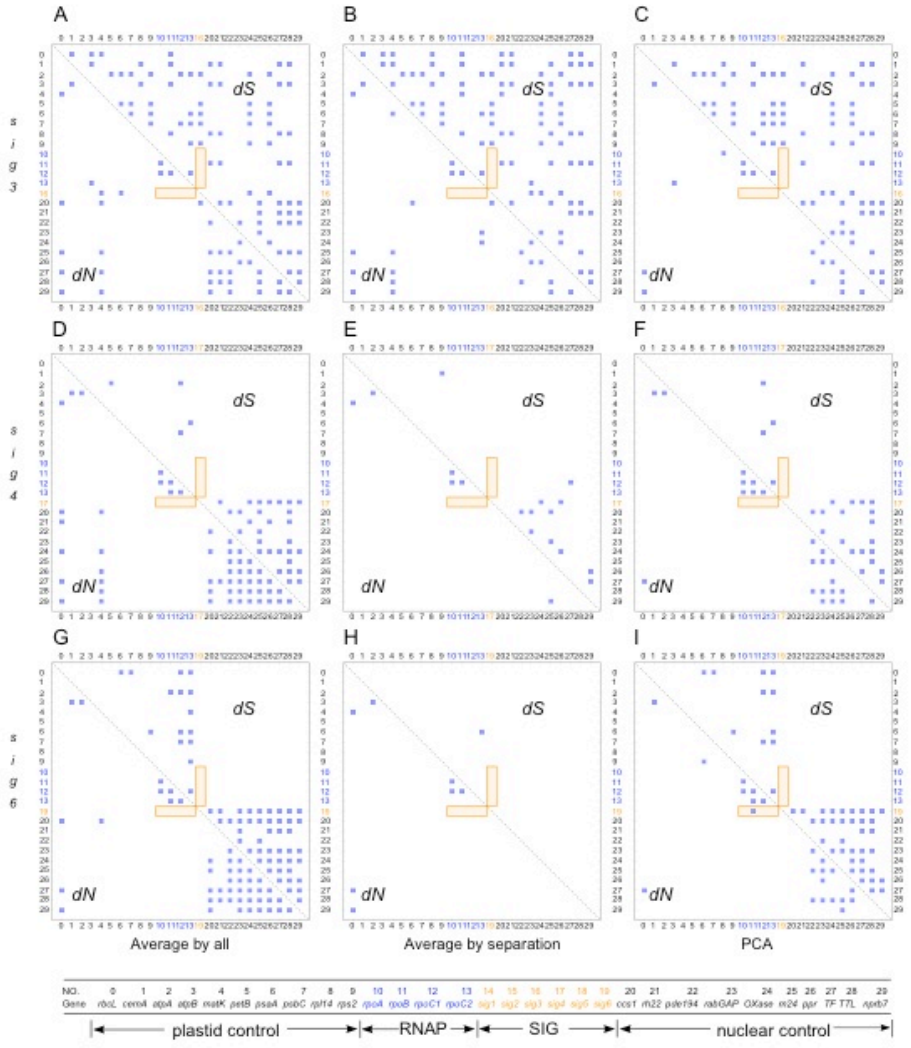
Appendix Figure 3.6. Strong correlation of nonsynonymous (*dN*) but not synonymous (*dS*) substitution rates between *sig1/2/5* and RNAP genes by three modifications of mirror tree methods using conserved domains.

The conserved domains of *rpoA/B* and SIG genes, and the entire sequences of *rpoC1/C2* and control genes were used in the analyses. The correlation of *dN* and *dS* values were calculated by (A) average method (all), (B) average method (separation) and (C) PCA. All interaction pairs with a correlation coefficient higher than 0.6 were considered significant and are indicated with a blue square. The rate correlations between RNAP and SIG genes are highlighted with an orange rectangle. RNAP and SIG genes, highlighted in blue and orange fonts respectively, show strong correlation of *dN* but not *dS* (highlighted in orange shaded box), and no or few correlation of *dN* between RNAP/SIG genes and the control genes (in black font) was detected. Gene names and cellular location corresponding to each number are given below the diagram.



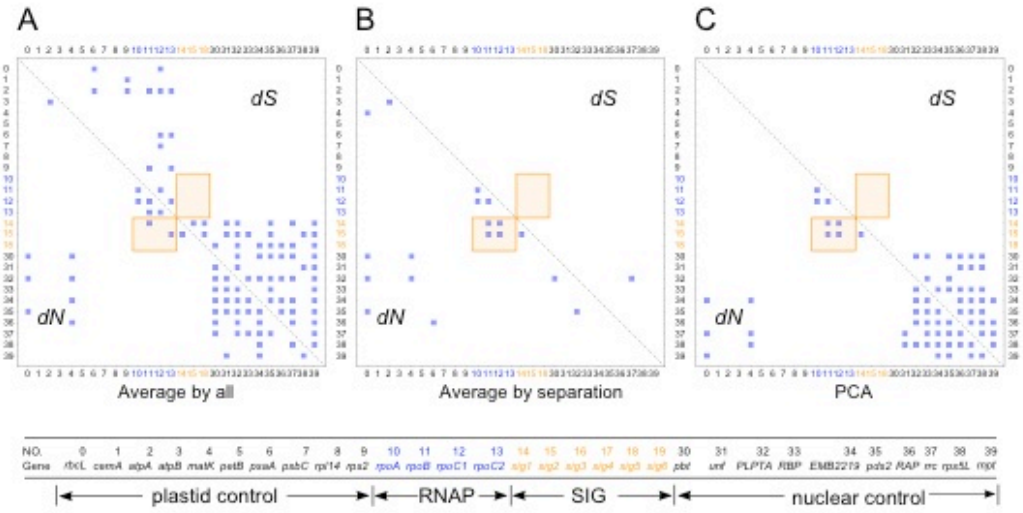
Appendix Figure 3.7. Little to no correlation of nonsynonymous (*dN*) or synonymous (*dS*) substitution rates between *sig3*/*4*/*6* and RNAP genes by three mirror tree methods.

The entire sequence of each gene was used in the analyses. Substitution rates (*dN*, lower; *dS*, upper) were calculated for (A-C) *sig3*, (D-F) *sig4* and (G-I) *sig6* using three methods: (A,D,G) average by all, (B,E,H) average by separation and (C,F,I) PCA. Interaction pairs with a correlation coefficient higher than 0.6 were considered significant and are indicated with a blue square. The rate correlations between RNAP and SIG genes are highlighted with an orange rectangle. A single correlation between *dS for sig3* and *rpoC2* was detected in (B). Gene names and cellular location corresponding to each number are given below the diagram.



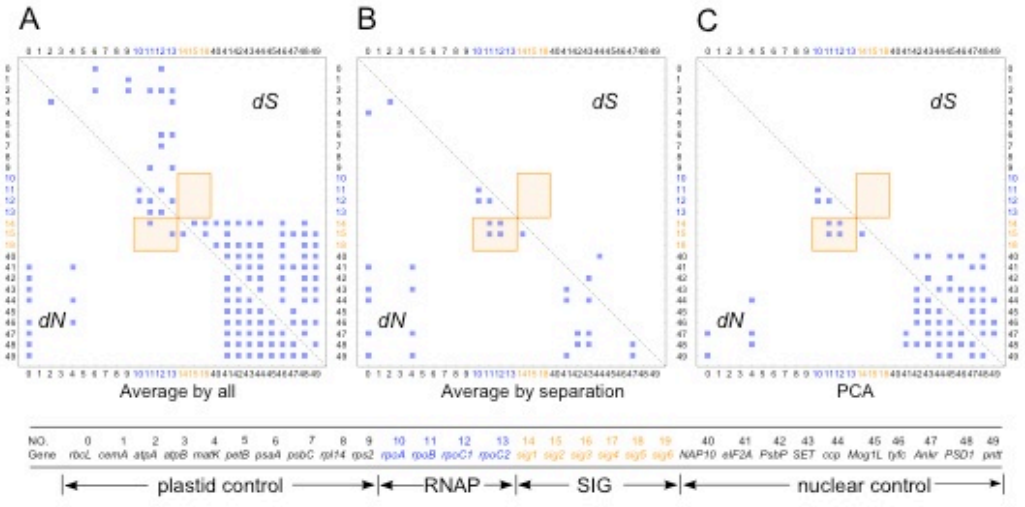
Appendix Figure 3.8. Little to no correlation of nonsynonymous (*dN*) or synonymous (*dS*) substitution rates between *sig3/4/6* and RNAP genes by three mirror tree methods using conserved domains.

The conserved domains of rpoA/B and SIG genes, and the entire sequences of rpoC1/C2 and control genes were used in the analyses. Substitution rates (dN, lower; dS, upper) were calculated for (A-C) sig3, (D-F) sig4 and (G-I) sig6 using three methods: (A,D,G) average by all, (B,E,H) average by separation and (C,F,I) PCA. Interaction pairs with a correlation coefficient higher than 0.6 were considered significant and are indicated with a blue square. The rate correlations between RNAP and SIG genes were highlighted with an orange shaded rectangle. A single correlation between dN for sig6 and rpoC1 was detected in (I). Gene names and cellular location corresponding to each number are given below the diagram.



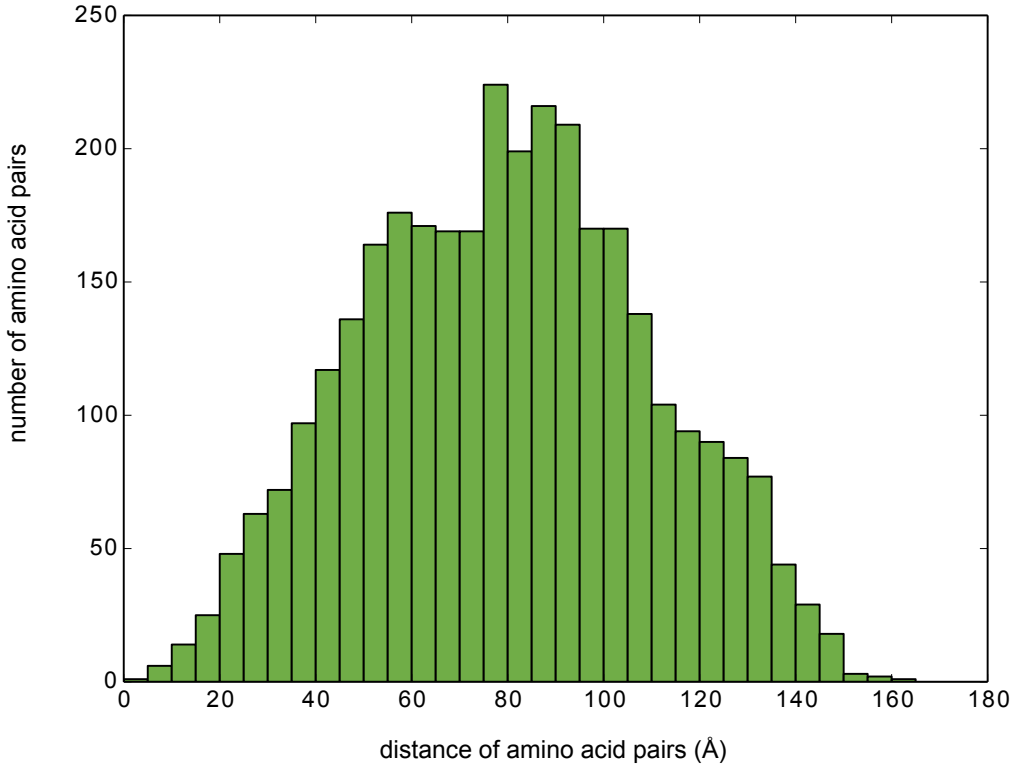
Appendix Figure 3.9. Strong correlation of nonsynonymous (*dN*) but not synonymous (*dS*) substitution rates between *sig1/2/5* and RNAP genes by three mirror tree methods using the entire sequences and a second set of 10 nuclear control genes.

The correlation of *dN* and *dS* values between SIG/RNAP genes and the additional control gene set (see Methods) was calculated by (A) average method (all), (B) average method (separate) and (C) PCA. All interaction pairs with a correlation coefficient higher than 0.6 were considered significant and are indicated with a blue square. The rate correlations between RNAP and SIG genes are highlighted with an orange shaded rectangle. RNAP and SIG genes, highlighted in blue and orange fonts respectively, show strong correlation of *dN* but not *dS* and no or few correlations of *dN* were detected between RNAP/SIG genes and the control genes (in black font). Gene names and cellular location corresponding to each number are given below the diagram.



Appendix Figure 3.10. Strong correlation of nonsynonymous (*dN*) but not synonymous (*dS*) substitution rates between *sig1/2/5* and RNAP genes by three mirror tree methods using the entire sequences with a third set of 10 nuclear control genes.

The entire sequence of each gene (except for *rpoA/B*, SIG) was used in the analyses. The correlation of *dN* and *dS* values between SIG/RNAP genes and the additional control gene set (see Methods) were calculated by (A) average method (all), (B) average method (separate) and (C) PCA. All interaction pairs with a correlation coefficient higher than 0.6 were considered significant and are indicated with a blue square. The rate correlations between RNAP and SIG genes are highlighted with an orange rectangle. RNAP and SIG genes, highlighted in blue and orange fonts respectively, show strong correlation of *dN* but not *dS*, and no or few correlations of *dN* were detected between RNAP/SIG genes and the control genes (in black font). Gene names and cellular location corresponding to each number are given below the diagram.



Appendix Figure 3.11. The distribution of distances between amino acid pairs predicted to be involved in structurally-mediated coevolution.

Among the 4223 residue pairs predicted to be involved in compensatory evolution by CoMap (see Methods), only one pair had a distance less than 5 Å.

Appendix Table 3.1. Summary of accession numbers, RT-PCR results and voucher information for all species examined.

Voucher information with collector and collection numbers used for both genome and transcriptome sequencing are deposited at TEX-LL. CB = Chris Blazier, Weng = Mao-Lun Weng, JZ = Jin Zhang.

Species	***sig1***	***sig2***	***sig3***	***sig4***	***sig5***	***sig6***	**Voucher**
California macrophylla	KJ916853	KJ916880	KJ916910	KJ916934	KJ916958* KJ916959*	KJ916992* KJ916993*	CB G1030
Erodium chrysanthum	KJ916854	KJ916881	KJ916911	KJ916935	KJ916960	KJ916994	CB G1039
Erodium cheilanthifolium	KJ916855	KJ916882	KJ916912	-	KJ916961* KJ916962*	KJ916995* KJ916996*	Weng G1032
Erodium gruinum	KJ916856	KJ916883	KJ916913	-	KJ916963	KJ916997	CB G1042
Erodium texanum	KJ916857	KJ916884	KJ916914	KJ916936	KJ916964* KJ916965*	KJ916998	CB G1018
Erodium trifolium	KJ916858	KJ916885	KJ916915	KJ916937	KJ916966 KJ916967	KJ916999* KJ917000*	Weng G1042
Francoa sonchifolia	KJ916859	KJ916886	KJ916916	-	KJ916968	KJ917001	Weng G1034
Geranium incanum	KJ916860	KJ916887	KJ916917	KJ916938	KJ916969	KJ917002	JZ G1001
Geranium maderense	KJ916861	KJ916888 KJ916889	KJ916918	KJ916939	KJ916970* KJ916971* KJ916972*	KJ917003	Weng G1046
Geranium phaeum	KJ916862	KJ916890	KJ916919	KJ916940	KJ916973* KJ916974*	KJ917004	Weng G1036
Hypseocharis bilobata	KJ916863	KJ916891	KJ916920	KJ916941	KJ916975	-	Weng G1041
Melianthus villosus	KJ916864	KJ916892	KJ916921	KJ916942	KJ916976	KJ917005	CB G1032
Monsonia emarginata	KJ916865	KJ916893	KJ916922	KJ916943	KJ916977	KJ917006	Weng G1033
Monsonia marlothii	KJ916866	KJ916894	KJ916923	KJ916944	KJ916978	KJ917007	Weng G1075
Pelargonium australe	KJ916867	KJ916895	KJ916924	KJ916945	KJ916979	KJ917008* KJ917009*	Weng G1002
Pelargonium citronellum	KJ916868	KJ916896	KJ916925	KJ916946	KJ916980	KJ917010	Weng G1006
Pelargonium cotyledonis	KJ916869	KJ916897	KJ916926	KJ916947	KJ916981	KJ917011* KJ917012*	Weng G1004
Pelargonium dichondrifolium	KJ916870	KJ916898	KJ916927	KJ916948	KJ916982	KJ917013* KJ917014*	Weng G1010
Pelargonium echinatum	KJ916871	KJ916899	KJ916928	KJ916949	KJ916983	KJ917015	Weng G1018

Pelargonium exstipulatum	KJ916872	KJ916900	KJ916929	KJ916950	KJ916984	KJ917016* KJ917017*	Weng G1020
Pelargonium fulgidum	KJ916873	KJ916901	KJ916930	KJ916951	KJ916985	KJ917018	Weng G1026
Pelargonium incrassatum	KJ916874	KJ916902	KJ916931	KJ916952	KJ916986	KJ917019	Weng G1009
Pelargonium myrrhifolium	KJ916875	KJ916903	-	KJ916953	KJ916987	KJ917020	CB G1006
Pelargonium nanum	KJ916876	KJ916904	-	KJ916954	KJ916988	KJ917021	CB G1015
Pelargonium tetragonum	KJ916877	KJ916905* KJ916906*	-	KJ916955	KJ916989	KJ917022	Weng G1007
Pelargonium transvalenese	KJ916878	KJ916907* KJ916908*	KJ916932	KJ916956	KJ916990	KJ917023	CB G1005
Pelargonium* x *hortorum	KJ916879	KJ916909	KJ916933	KJ916957	KJ916991	KJ917024	Weng G1003

* This sequence has been verified by RT-PCR.
- This gene is either missing or pseudogenized.

Appendix Table 3.2. Summary of SIG gene duplication and loss events in Geraniales.

Findings from Notung analysis (Durand et al., 2006). "D" indicates gene duplication, "L" indicates gene loss events and "-" indicates absence of any events. See Appendix Figure 3.3 for location of these events in phylogenetic trees.

	sig1	***sig2***	***sig3***	***sig4***	***sig5***	***sig6***
California macrophylla	-	-	-	-	D	D
Erodium chrysanthum	-	-	-	-	-	-
Erodium cheilanthifolium	-	-	-	-	D	D
Erodium gruinum	-	-	-	-	-	-
Erodium texanum	-	-	-	-	D	-
Erodium trifolium	-	-	-	-	D	D
Francoa sonchifolia	-	-	-	L	-	-
Geranium incanum	-	-	-	-	L	-
Geranium maderense	-	D	-	-	D	-
Geranium phaeum	-	-	-	-	D	-
Hypseocharis bilobata	-	-	-	-	-	L
Melianthus villosus	-	-	-	-	-	-
Monsonia emarginata	-	-	-	-	D	-
Monsonia marlothii	-	-	-	-	D	-
Pelargonium australe	-	-	D/L	-	-	D
Pelargonium citronellum	-	D/L	-	-	-	-
Pelargonium cotyledonis	-	-	L	-	-	D
Pelargonium dichondrifolium	-	-	-	-	-	D
Pelargonium echinatum	-	-	-	-	-	-
Pelargonium exstipulatum	-	-	-	-	-	D
Pelargonium fulgidum	-	-	-	-	-	-
Pelargonium incrassatum	-	-	-	-	-	-
Pelargonium myrrhifolium	-	L	-	-	-	-
Pelargonium nanum	-	L	L	-	-	-
Pelargonium tetragonum	-	D	-	-	-	-
Pelargonium transvalenese	-	D	-	-	-	-
Pelargonium x *hortorum*	-	L	-	-	-	-

Appendix Table 3.3. Pairwise comparison of evolutionary rates from different alignment methods.

Alignments are available in Supplementary Data Set 2.

Cosine similarity	*dN*			*dS*		
	MAFFT – MUSCLE	**MAFFT – ClustalW**	**MUCLE – ClustalW**	**MAFFT – MUSCLE**	**MAFFT – ClustalW**	**MUCLE – ClustalW**
atpA	1	1	1	1	1	1
atpB	1	1	1	1	1	1
cemA	1	1	1	1	0.99	0.99
matK	1	1	1	1	1	1
petB	1	1	1	1	1	1
psaA	1	0.99	1	1	1	1
psbC	1	0.99	0.99	1	1	1
rbcL	1	1	1	1	1	1
rpl14	1	0.99	0.99	1	1	1
rps2	1	1	1	0.99	1	1
rpoA	0.99	0.99	1	0.99	0.98	0.99
rpoB	1	1	1	1	1	1
rpoC1	1	1	1	1	1	1
rpoC2	1	1	1	1	1	1
sig1	1	0.99	1	1	1	1
sig2	1	1	1	1	1	1
sig3	1	0.61	0.60	1	0.81	0.75
sig4	1	1	1	1	1	1
sig5	1	1	1	1	1	1
sig6	1	1	1	1	1	1
ccs1	1	1	1	1	1	1
rh22	1	1	1	1	1	1
pde194	1	1	1	1	1	1
rabGAP	1	1	1	1	1	1
OXase	1	1	1	1	1	1
m24	1	1	1	1	1	1
ppr	1	1	1	1	1	1
TF	1	1	1	1	1	1
TTL	1	1	1	1	1	1
nprb7	1	1	1	1	1	1

Appendix Table 3.4. The number of interaction pairs with a rate coefficient of over 0.6 within corresponding genes.

RNAP contains *rpoA*, *rpoB*, *rpoC1* and *rpoC2*. Rates of *rpoA*, *rpoB* and SIG genes were calculated using the conserved domains and rates of *rpoC1*, *rpoC2* were calculated using the entire sequences (see Methods). Results of the average by all method (ρ_{ava}), average by separation method (ρ_{avs}) and PCA method (ρ_{pca}) are shown.

	dN			***dS***		
interactions*	ρ_{ava}	ρ_{avs}	ρ_{pca}	ρ_{ava}	ρ_{avs}	ρ_{pca}
RNAP – RNAP (6)	5	3	5	2	0	3
RNAP – *sig1* (4)	3	2	3	0	0	0
RNAP – *sig2* (4)	3	2	3	0	0	0
RNAP – *sig3* (4)	0	0	0	0	0	0
RNAP – *sig4* (4)	0	0	0	0	0	0
RNAP – *sig5* (4)	0	0	0	0	0	0
RNAP – *sig6* (4)	0	0	1	0	0	0
RNAP – control (80)	0	0	0	10	0	8
***sig1* – control (20)**	0	0	0	4	0	4
***sig2* – control (20)**	1	0	1	2	0	0
***sig3* – control (20)**	3	1	0	7	8	8
***sig4* – control (20)**	0	0	0	7	2	4
***sig5* – control (20)**	0	0	0	9	0	5
***sig6* – control (20)**	0	0	0	9	0	9

* The number in parentheses is the total number of interaction pairs within corresponding genes.

Appendix Table 3.5. Ranksum test of rate correlation coefficient using conserved domains.

Rates of *rpoA*, *rpoB* and SIG were calculated using the conserved domains and rates of *rpoC1* and *rpoC2* were calculated using the entire sequences (see Methods). Correlation coefficients ranked significantly higher in the overall interaction pairs are indicated with "+". Coefficients that are not ranked significantly higher are indicated with "-". Pt is the group of control genes from the plastid genome and nu is the group of control genes from the nuclear genome. Results of the average by all tree method (ρ_{ava}), average by separation method (ρ_{avs}) and PCA method (ρ_{pca}) are shown.

	dN			*dS*		
interactions	ρ_{ava}	ρ_{avs}	ρ_{pca}	ρ_{ava}	ρ_{avs}	ρ_{pca}
RNAP - RNAP	+	+	+	-	-	-
RNAP - *sig1*	-	+	+	-	-	-
RNAP - *sig2*	-	+	+	-	-	-
RNAP - *sig3*	-	-	-	-	-	-
RNAP - *sig4*	-	-	-	-	-	-
RNAP - *sig5*	-	-	+	-	-	-
RNAP - *sig6*	-	+	+	-	-	-
RNAP - pt	-	-	-	-	-	-
RNAP - nu	-	-	-	-	-	-

Appendix Table 3.6. Test of *dS* saturation of 30 genes.

The P value indicates the significance of the improvement of sum of squares from the linear model to the quadratic model

gene	P value	gene	P value	gene	P value
rbcL	<0.05*	***rpoA***	<0.05	***ccs1***	<0.05
cemA	<0.05	***rpoB***	<0.05	***rh22***	<0.05
atpA	<0.05	***rpoC1***	<0.05	***ppd194***	<0.05
atpB	<0.05	***rpoC2***	<0.05	***rabGAP***	<0.05
matK	>0.05*	***sig1***	<0.05	***OXase***	<0.05
petB	>0.05	***sig2***	<0.05	***m24***	<0.05
psaA	<0.05	***sig3***	<0.05	***ppr***	<0.05
psbC	>0.05	***sig4***	<0.05	***TTR***	>0.05*
rpl14	>0.05	***sig5***	<0.05	***TF***	<0.05
rps2	<0.05	***sig6***	<0.05	***nprb7***	>0.05*

* The best fit quadratic model has a concave shape.

Appendix Table 3.7. Analysis of interaction sites and overlap between the coevolving and interaction sites.

The number of sites involved in both coevolution and interaction is shown.

	SIG	*rpoA*	*rpoB*	*rpoC*
CMA*	96	14	62	100
sig1	17	0	8	18
sig2	16	0	3	9
sig3	17	0	6	14
sig4	14	0	4	12
sig5	26	0	8	13
sig6	26	0	7	17
Distance**	118	9	47	82
sig1	21	0	8	15
sig2	21	0	1	6
sig3	15	0	6	13
sig4	17	0	4	11
sig5	26	0	6	9
sig6	31	0	6	13

* The number of interaction sites predicted by contact map analysis (CMA).
** The number of interaction sites predicted by direct distance estimation.

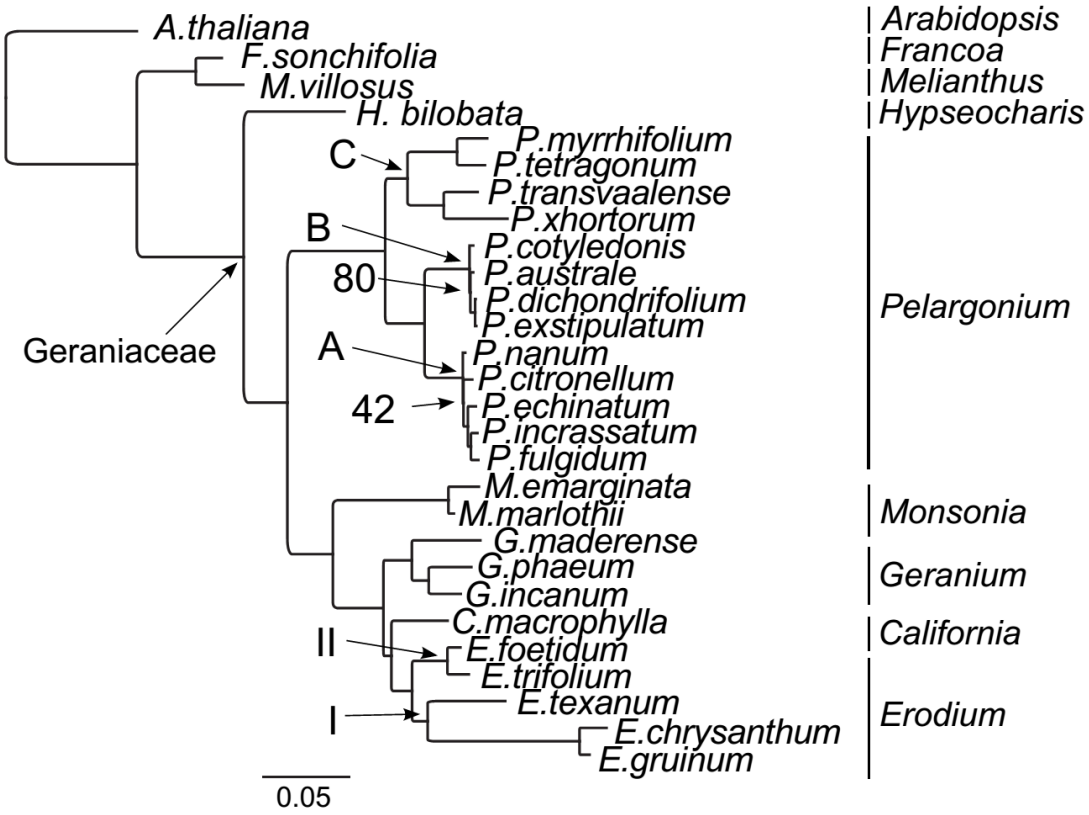
Appendix Table 3.8. Primer pairs used for amplification of sigma factor genes.

NO.	forward primer	reverse primer	position* (from/ to)	species	template
1	GACAGAGGCTGGTTTCTGCA	TTCAGACAGACTGTGCGGTC	860 to 1,521	*H. bilobata*	*sig6*
2	TCATCGTCGTCGTCGTCATC	CTGATCGGTGGGCTCGTATC	184 to 705	*P. myrrhifolium*	*sig1*
3	TCGGGAAACGCTTGTTACGA	GTGAGCTCCTCCTCTAGCCT	288 to 932	*P. myrrhifolium*	*sig2*
4	AGGGCCTGTTTATTTGCGGA	GCTGCACTGGAAAGTTGCTC	81 to 647	*P. transvalenese*	*sig2*-1
5	TGCTGCAAAAGGTCCACTGA	TACACGTCGCAATGGATCGT	225 to 774	*P. transvalenese*	*sig2*-2
6	CCTCCGGTTCTTGGTTCTTCA	GGGTGATCGGGTTCTTTGCA	652 to 1,264	*P. transvalenese*	*sig5*
7	GCTCAACCTCAAGTGGGACA	GAGCCCCAAGCCTTTGTACT	11 to 526	*P. exstipulatum*	*sig3*
8	TCCGCCGTTTTCTTCGAGAA	AACGTCGCTCTTGAAGTCGT	33 to 625	*P. exstipulatum*	*sig6*-1
9	CATGTCCTCCCAGAACCCAC	CTGGTTGCTCGTTTTGAGGC	38 to 491	*P. exstipulatum*	*sig6*-2
10	ACTGCGCTTCTTCTGCTTCT	CACCTTTTCTTGGAACGCCG	65 to 617	*P. exstipulatum*	*sig6*-2
11	AGCTGTGTATGATGATCCAGGA	GGGGCCATTGTCTGTGACTA	386 to 969	*P. dichondrifolium*	*sig4*
12	GCTCACTTCACTTCTCTTCACC	TGCTCATTCCCGTAGCACTG	-66 to 502	*P. dichondrifolium*	*sig4*
13	TGAGGCAATTGAGACAAGAAATTGT	TCATCAAGAGCAAGCCTGAGA	800 to 1,349	*P. dichondrifolium*	*sig5*
14	CTAGTTCATCTGCGGCACCA	CGCCATTTCCTTAACTCCCCT	211 to 820	*P. cotyledonis*	*sig3*
15	TGGACATTGTTGCTCTTCCCA	GAGCCCCAAGCCTTTGTACT	-52 to 476	*P. cotyledonis*	*sig3*
16	CTCCGCCATTTCCTTCGAGA	ACAACGGCTGAAAGCGATCA	32 to 751	*P. cotyledonis*	*sig6*-1

17	CGGCGTTCCAAGAAAAGGTG	GAGACCCCCACATCCTACCT	595 to 1,047	*P. cotyledonis*	*sig6*-2
18	ACAGTTAGAGGCAAGGAGGC	GCGACAGCGAAGAGATTTTCT	87 to 702	*P. australe*	*sig3*
19	GCAGTGTCGGGATGAGCTAA	GGTCGATGATCCTGAAGCCC	962 to 1,536	*P. australe*	*sig3*
20	TGAACAAATGGGGGTGCCTT	CCCGTAAGCTTGCTCGAGAA	1,020 to 1,594	*P. australe*	*sig3*
21	CACTCTCGTTCCTTCCCCAC	TCCAGAGCGGACAACTCTTT	39 to 633	*P. australe*	*sig3*
22	TACAGCAGCCGGGAAAAGTT	CTCCCTTGGCTTGAGAGTGG	943 to 1,503	*P. australe*	*sig6*-1
23	GGCTTCTTTCAATGGCGCAA	AACACCCTCTCCCTTCTTGA	114 to 668	*P. nanum*	*sig3*
24	CAAGAAGGGAGAGGGTGTTAGA	GCTCGGCTTAACGTGTAAGG	650 to 1,217	*P. nanum*	*sig3*
25	GGCGAGAACAGTCACAGTCA	TCTTCTTCCCAGAACGGACA	168 to 640	*P. fulgidum*	*sig3*
26	TGGACATTGTTGCTCTTCCCA	TGAAGTCTCGGTGGTCCTTG	-52 to 543	*P. fulgidum*	*sig3*
27	TAAAGCCTGGCCGTTTCACT	TGTAGGCGACCGCTGTAATC	350 to 869	*P. fulgidum*	*sig4*
28	GCTCTCCCACTCTCCCTACA	CATGCTCCCCACCTGAATGA	36 to 647	*P. fulgidum*	*sig4*
29	GGCGAGAACAGTCACAGTCA	ATGTGACGCTCGTTTTGCAG	180 to 789	*P. incrassatum*	*sig3*
30	ACTTAATGGATATTGTTGCTCTTCCC	TGAAGTCTCGGTGGTCCTTG	-59 to 561	*P. incrassatum*	*sig3*
31	TGAACCAGAAAGAGGGTCGC	CTTGCGCACCATTTCTCTCG	1,011 to 1,503	*G. maderrense*	*sig5*-1
32	TGCCAAGCAGGTATTCAGGG	CACCAAGCTCTCCTTCGGTT	934 to 1,389	*G. maderrense*	*sig5*-2
33	CTGTTTCTAGCTCAGCCGCT	ACCATGCTTGCTCCGTGTAA	14 to 688	*G. incanum*	*sig5*
34	TCATACAATGGCCGACGGTC	TGCCGACCAAACTGCAATTG	265 to 878	*G. phaeum*	*sig3*
35	CAAACTTGAACCCGCCACAG	ACAAACAGTACCAGGCGGAG	318 to 887	*G. phaeum*	*sig5*-1

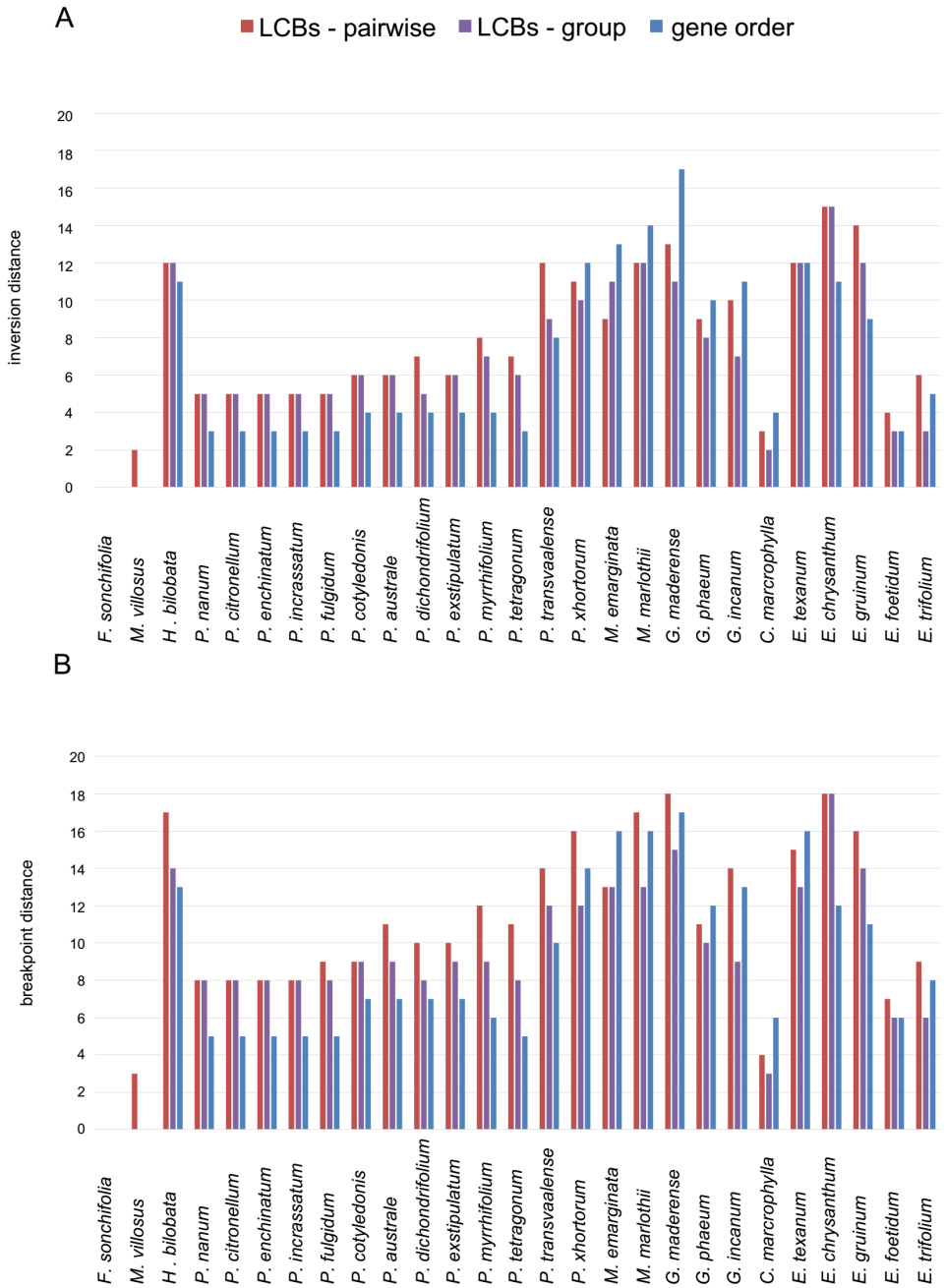
36	CCCGGCAAAGAAAGTCAAGC	AAAGCGATCAACGGCTCTGA	306 to 828	*G. phaeum*	*sig5*-2
37	TCTTCGGAGCAAGCATGGTT	GCTCTCAGCACATCACGGTA	664 to 1,217	*C. macrophylla*	*sig5*-1
38	TCGAGCAGCAACAAACAAGC	GCCTTTTCCATCCAGTCCGA	828 to 1,428	*C. macrophylla*	*sig5*-2
39	CCCTGGCTATTCCGAGCAAT	ATTCAACCAGGGTCGGTTCC	270 to 886	*C. macrophylla*	*sig6*-1
40	CTGCACATGGTTGCCGATTT	GCCTTAGCCTTTGCATTCCG	1,076 to 1,600	*C. macrophylla*	*sig6*-2
41	GGAACCGACCCTGGTTGAAT	CGTGGTAAGTGACACCTCCC	852 to 1,369	*E. trifolium*	*sig6*-1
42	ACGGCGCCGCTAAATTATTC	AGCCATGGATCTTCGTCATCA	599 to 1,277	*E. trifolium*	*sig6*-2
43	ACTTTAGTTGGAGCGCACGA	CGCTCTACCAACGTCCAGTT	127 to 741	*E. cheilanthifolium*	*sig5*-2
44	TGGATGGCCCCAATTTGGAA	TTCCCTTTTCCATCCAGCCC	863 to 1,385	*E. cheilanthifolium*	*sig5*-1
45	GCTCTGCTTCTTCTCCAGCA	CACCTTTGAGCTTTGGCACC	65 to 652	*E. cheilanthifolium*	*sig6*-1
46	CTGCGCCCTATGCACTATGA	TTTTCTTTGAACGCCGCTCG	-49 to 505	*E. cheilanthifolium*	*sig6*-2
47	TCCCACTTCACCGTTCAAGC	TTCGGCTTCTAGACTGCGAC	84 to 706	*E. texanum*	*sig3*
48	AGTTGATCGCTTCGAACCCA	ATGTTCAAGTTCCCGGCGAT	876 to 1,364	*E. texanum*	*sig5*-1
49	GCTGCTGCCAAACTCGAAAA	GCTTGTTTGTTGCAGCTCGA	289 to 829	*E. texanum*	*sig5*-2
50	ACTCTCATCCTCTCCCCACC	CCTTTGCCATCTCCACCTGA	41 to 667	*E. chrysanthum*	*sig3*

* The position of the first nucleotide in the start codon is defined as 1



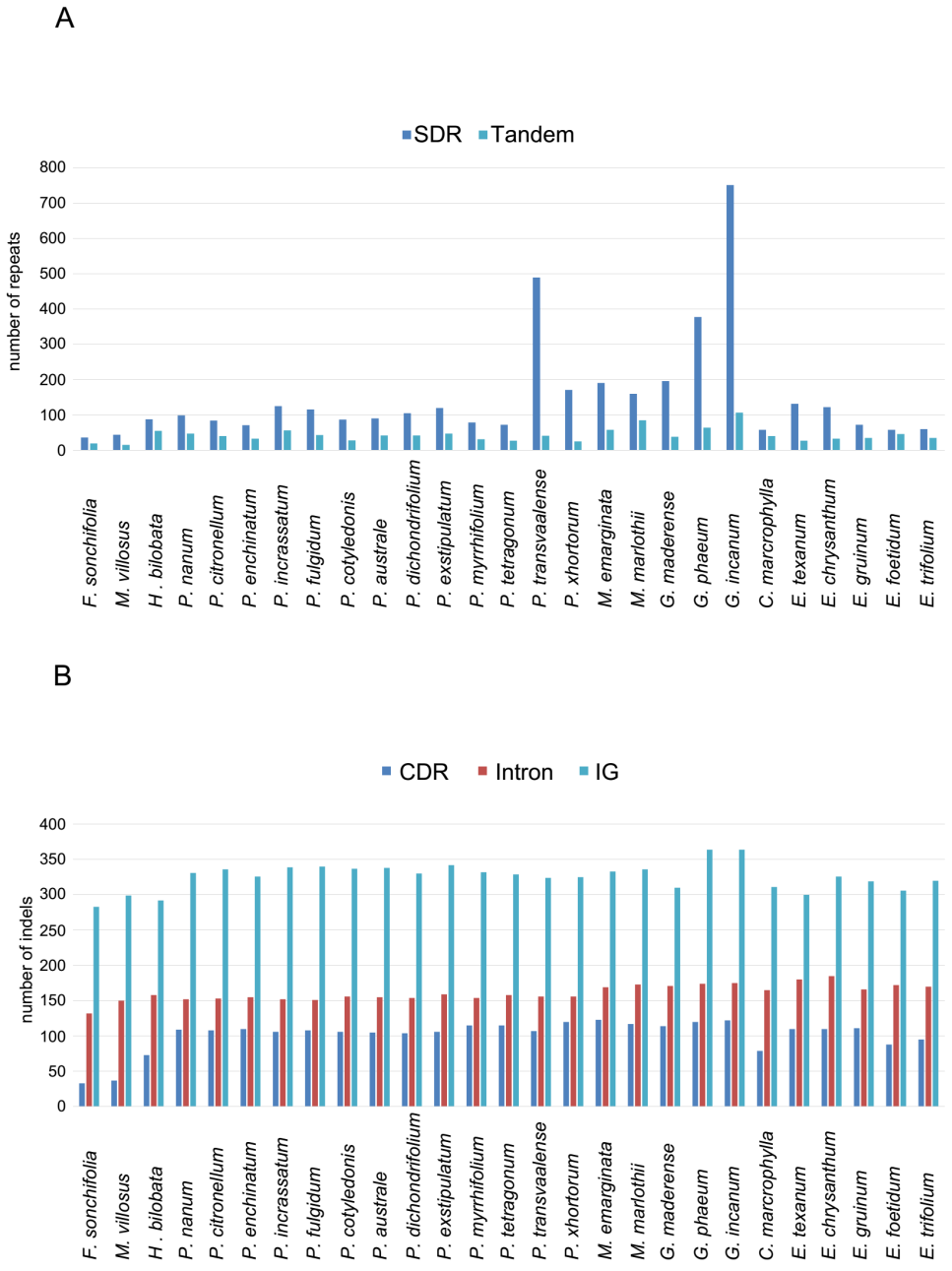
Appendix Figure 4.1. Maximum likelihood tree of 27 species from Geraniales and Arabidopsis.

Twelve plastid genes were used to construct the constraint tree (see Methods). All nodes have bootstrap values of 100 except those shown at nodes on the tree. Generic names are shown on the right and clade designations within each genus are indicated at nodes in the tree. Scale bar indicates numbers of nucleotide substitutions per codon.



Appendix Figure 4.2. Measures of genome rearrangement of 27 Geraniales species.

(A) Inversion and (B) breakpoint distances were estimated based on order of Local Collinear Blocks (LCBs) generated by pairwise or multiple genome alignments using Mauve, or on synteny of shared genes.



Appendix Figure 4.3. Enumeration of repeats and insertions/deletions (indels) in 27 Geraniales species.

(A) Number of small dispersal repeats (SDR) and tandem repeats, and (B) insertion and deletions (indels) of coding and ribosomal RNA (CDR), intergenic (IG) and intron regions.

Appendix Table 4.1. Read statistics for genome sequencing of two *Monsonia* species.

species	**sequencing platform**	**number of reads**	**average read length (bp)**	**insert size (bp)**
M. emarginata	Illumina HiSeq 2000	62,496,270	100	767±60
M. emarginata	PacBio	126,535	2,191	n/a
M. marlothii	Illumina HiSeq 2000	23,253,398	100	741±187
M. marlothii	PacBio	59,666	1,888	n/a

Appendix Table 4.2. Genes and genomic regions involved in indel estimation.

The names in the intron row indicate genes containing introns with indels. The gene pairs in the intergenic regions row indicate the genes flanking the intergenic regions analyzed.

Region[1]	Content
CDR	*rrn4.5, rrn5, rrn16, rrn23, atpA, atpB, atpE, atpF, atpH, atpI, ccsA, cemA, matK, petA, petB, petD, petG, petL, petN, psaA, psaB, psaC, psaI, psaJ, psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ, rbcL, rpl2, rpl14, rpl16, rpl20, rpl22, rpl32, rpl33, rpl36, rpoA, rpoB, rpoC1, rpoC2, rps2, rps3, rps4, rps7, rps11, rps14, rps15, rps16, rps19, ycf3, ycf4*
Intron	*atpF, ndhB, petB, petD, trnG-UCC, trnI-GAU, trnL-UAA, ycf3*
IG	*atpB - rbcL, atpH - atpF, cemA - petA, ndhJ - trnF-GAA, petA - psbJ, petB - petD, petG - petL, petG - trnW-CCA, petL - psbE, petN - psbM, psaC - ndhD, psbB - psbT, psbH - petB, psbK - psbI, psbL - psbJ, psbN - psbH, rpl14 - rps8, rpl16 - rpl14, rpl36 - rps11, rpoC2 - rps2, trnH-GUG - psbA, trnL-UAG - ccsA, trnP-UGG - trnW-CCA, trnY-GUA - trnD-GUC*

[1]Genome complexity abbreviations: CDR, number of indels in coding and ribosomal RNA regions; Intron, number of indels in intron regions; IG, number of indels in intergenic regions.

Appendix Table 4.3. Correlation among measures of genome rearrangement.

Inversion (IV) or Breakpoint (BP) distance was estimated based on either the order of local collinear blocks (LCBs) generated by pairwise genome alignments (LCB - p), or multiple genome alignments (LCB - g), or on the order of shared genes.

Measure	IV LCB - p	BP LCB - p	IV LCB - g	BP LCB - g	IV gene order
BP LCB - p	0.97				
IV LCB - g	0.95	0.95			
BP LCB - g	0.93	0.95	0.97		
IV gene order	0.87	0.87	0.86	0.81	
BP gene order	0.86	0.87	0.86	0.83	0.98

Appendix Table 4.4. Correlation among measures of genome complexity.

Measure	LCB - p[1]	Gene order[1]	SDR[1]	Tandem[1]	CDR[1]	Intron[1]	IG[1]	*dN* CP[1]	*dS* CP[1]
LCB - p[1]									
Gene order[1]	0.88								
SDR[1]	-0.12	-0.04							
Tandem[1]	-0.14	0.02	0.68						
CDR[1]	0.36	0.34	0.38	0.41					
Intron[1]	-0.22	-0.24	0.53	0.59	0.72				
IG[1]	0.39	0.47	0.33	0.4	0.49	0.21			
dN CP[1]	0.48	0.33	0.12	0.18	0.92	0.6	0.36		
dS CP[1]	0.56	0.41	0.13	0.16	0.77	0.42	0.67	0.83	

[1]Genome complexity abbreviations: LCBs - p, IV distance estimated from local collinear blocks; Gene order, IV distance estimated from gene order; SDR, small dispersal repeats; Tandem, repeats estimated from Tandem Repeat Finder (Benson, 1999); CDR, number of indels in coding and ribosomal RNA regions; Intron, number of indels in intron regions; IG, number of indels in intergenic regions; *dN* CP, nonsynonymous substitution rates of the plastid genome; *dS* CP, synonymous substitution rates of the plastid genome.

Appendix Table 4.5. Genes showing significant correlation of nonsynonymous substitution rate (*dN*) with genome complexity.

Measures of genome complexity lacking correlation to any genes were excluded. Detailed information for each gene is included in Supplemental Data File 4.1.

Gene	subcellular location[1]	CDR[3]	Intron[3]	IG[3]	*dN* CP[3]	*dS* CP[3]
pde194	P	+[2]	-	-	-	+
ccs1	P	-	-	+	-	+
rh22	P	+	+	-	+	-
RAP	P	+	-	-	-	-
ccp	P	-	-	+	-	-
tyfc	P	+	+	-	-	-
ppr	P	+	-	-	+	-
PTAC13	P	+	-	-	+	-
NOL1	P	+	-	+	-	-
TPRl	P	-	-	+	-	+
gyra	P	-	-	-	+	-
uvrB	P	+	+	-	+	-
why1	P	+	-	-	-	-
PSD1	M	-	-	+	-	-
MPPALPHA	P/M	-	-	+	-	-
maturase	P/M	-	-	-	+	-

[1]P, plastid targeted; M, mitochondrion targeted.

[2]Significant correlation is indicated with "+" and non-significant correlation is shown with "-".

[3]Genome complexity abbreviations: CDR, number of indels in coding and ribosomal RNA regions; Intron, number of indels in intron regions; IG, number of indels in intergenic regions; *dN* CP, nonsynonymous substitution rates of the plastid genome; *dS* CP, synonymous substitution rates of the plastid genome.

Appendix Table 4.6. Accession information for Brassicales data.

species	transcriptome	plastid genome
Arabidopsis lyrata	Phytozome[1]	JGI[2]
Arabidopsis thaliana	Phytozome	NC_000932
Capsella bursa-pastoris	SRR1198325	NC_009270
Arabis alpina	SRR1647718	NC_023367
Barbarea verna	SRR1198323	NC_009269
Brassica rapa	Phytozome	NC_015139
Brassica napus	ERX397793	NC_016734
Pachycladon cheesemanii	SRR364071	NC_021102
Raphanus sativus	SRR922465	NC_024469
Carica papaya	Phytozome	NC_010323

[1]Phytozome: http://phytozome.jgi.doe.gov/pz/portal.html
[2]JGI: http://genome.jgi-psf.org/Araly1/Araly1.download.ftp.html

Appendix Table 4.7. Genes used in rate comparisons between Geraniales and Brassicales.

Group[1]	Genes
NUCP	*ccs1, rh22, pde194, EMB2219, RAP, HAD, ccp, HAD-like, tyfc, ppr, SUFS, amidase hydrose, PTAC13, NOL1, PCB2, UPM1, CGI-126L, amdmt, soulhb, TPR1, latglu, abhydrolase, Aspartase, SCY2, APX4, abHYD, PDE149*
NUMT	*OXase, pprf, rrc, rps5L, mpt, NAP10, PSD1, ISE1, pnat, ATP12r, Lojap, DECOY, Isu1, uboxase, tim, MPPALPHA, NDB1, NFU4, PPR336*
NUOT	*unf, eIF2A, rabGAP, m24, TF, Yabk, AtSec20, ATSFGH, polII, 26Sregulator, F-box, nprb7, AMP, amdmt, NagB, ARPC4, ELP6, CoG6L, amdmtf, unfp*
RRR	drt112, gyra, reca1, smr, mmr2
plastid	*atpA, atpB, atpE, atpF, atpH, atpI, ccsA, cemA, matK, petA, petB, petD, petG, petL, petN, psaA, psaB, psaC, psaI, psaJ, psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ, rbcL, rpl2, rpl14, rpl16, rpl20, rpl22, rpl32, rpl33, rpl36, rpoA, rpoB, rpoC1, rpoC2, rps2, rps3, rps4, rps7, rps11, rps14, rps15, rps16, rps19, ycf3, ycf4*

[1]Genome complexity abbreviation: NUCP, nuclear encoded plastid targeted control genes; NUMT, nuclear encoded mitochondrial targeted control genes; NUOT, other nuclear control genes; RRR, DNA replication, recombination and repair genes; plastid, plastid encoded genes.

References

CHAPTER 1

Zhang, J., Ruhlman, T.A., Mower, J.P., and Jansen, R.K. (2013). Comparative analyses of two Geraniaceae transcriptomes using next-generation sequencing. BMC Plant Biology 13: 228.

Zhang, J., Ruhlman, T.A., Sabir, J., Blazier, J.C., and Jansen, R.K. (2015). Coordinated rates of evolution between interacting plastid and nuclear genes in Geraniaceae. Plant Cell 27: 563–573.

CHAPTER 2

Adams, K.L., Qiu, Y.-L., Stoutemyer, M., and Palmer, J.D. (2002). Punctuated evolution of mitochondrial gene content: High and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. PNAS 99: 9905–9912.

Angeloni, F., Wagemaker, C. a. M., Jetten, M.S.M., Op den Camp, H.J.M., Janssen-Megens, E.M., Francoijs, K.-J., Stunnenberg, H.G., and Ouborg, N.J. (2011). *De novo* transcriptome characterization and development of genomic tools for *Scabiosa columbaria L.* using next-generation sequencing techniques. Mol Ecol Resour 11: 662–674.

Bakker, F.T., Breman, F., and Merckx, V. (2006). DNA sequence evolution in fast evolving mitochondrial DNA *nad1* exons in Geraniaceae and Plantaginaceae. Taxon 55: 887–896.

Birol, I. et al. (2009). *De novo* transcriptome assembly with ABySS. Bioinformatics 25: 2872–2877.

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

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31: 365–370.

Bräutigam, A., Mullick, T., Schliesky, S., and Weber, A.P.M. (2011). Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C3 and C4 species. J. Exp. Bot. 62: 3093–3102.

Cai, W., Ji, D., Peng, L., Guo, J., Ma, J., Zou, M., Lu, C., and Zhang, L. (2009). *LPA66* is required for editing *psbF* chloroplast transcripts in *Arabidopsis*. Plant Physiol. 150: 1260–1271.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

Cantacessi, C., Campbell, B.E., Young, N.D., Jex, A.R., Hall, R.S., Presidente, P.J.A., Zawadzki, J.L., Zhong, W., Aleman-Meza, B., Loukas, A., Sternberg, P.W., and Gasser, R.B. (2010). Differences in transcription between free-living and CO2-activated third-stage larvae of *Haemonchus contortus*. BMC Genomics 11: 266.

Chateigner-Boutin, A.-L., Ramos-Vega, M., Guevara-García, A., Andrés, C., de la Luz Gutiérrez-Nava, M., Cantero, A., Delannoy, E., Jiménez, L.F., Lurin, C., Small, I., and León, P. (2008). *CLB19*, a pentatricopeptide repeat protein required for editing of *rpoA* and *clpP* chloroplast transcripts. Plant J. 56: 590–602.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A.J., Müller, W.E.G., Wetter, T., and Suhai, S. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. 14: 1147–1159.

Cho, Y., Mower, J.P., Qiu, Y.-L., and Palmer, J.D. (2004). Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. Proc. Natl. Acad. Sci. U.S.A. 101: 17741–17746.

Chumley, T.W., Palmer, J.D., Mower, J.P., Fourcade, H.M., Calie, P.J., Boore, J.L., and Jansen, R.K. (2006). The complete chloroplast genome sequence of *Pelargonium* x *hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Molecular Biology and Evolution 23: 2175–2190.

Conesa, A. and Götz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics 2008: 619832.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676.

Cuenca, A., Petersen, G., Seberg, O., Davis, J.I., and Stevenson, D.W. (2010). Are substitution rates and RNA editing correlated? BMC Evolutionary Biology 10: 349.

Der, J.P., Barker, M.S., Wickett, N.J., dePamphilis, C.W., and Wolf, P.G. (2011). *De novo* characterization of the gametophyte transcriptome in bracken fern, *Pteridium aquilinum*. BMC Genomics 12: 99.

Downie, S.R., Katz-Downie, D.S., Wolfe, K.H., Calie, P.J., and Palmer, J.D. (1994). Structure and evolution of the largest chloroplast gene (*ORF2280*): internal plasticity and multiple gene loss during angiosperm evolution. Curr. Genet. 25: 367–378.

Duarte, J.M., Wall, P.K., Edger, P.P., Landherr, L.L., Ma, H., Pires, J.C., Leebens-Mack, J., and dePamphilis, C.W. (2010). Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. BMC Evolutionary Biology 10: 61.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics 14: 755–763.

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

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. 300: 1005–1016.

Feldmeyer, B., Wheat, C.W., Krezdorn, N., Rotter, B., and Pfenninger, M. (2011). Short read Illumina data for the *de novo* assembly of a non-model snail species transcriptome (*Radix balthica*, *Basommatophora*, *Pulmonata*), and a comparison of assembler performance. BMC Genomics 12: 317.

Fujii, S. and Small, I. (2011). The evolution of RNA editing and pentatricopeptide repeat genes. New Phytologist 191: 37–47.

Garg, R., Patel, R.K., Jhanwar, S., Priya, P., Bhattacharjee, A., Yadav, G., Bhatia, S., Chattopadhyay, D., Tyagi, A.K., and Jain, M. (2011). Gene discovery and tissue-specific transcriptome analysis in *chickpea* with massively parallel pyrosequencing and web resource development. Plant Physiol. 156: 1661–1678.

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

Gotz, S., Garcia-Gomez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talon, M., Dopazo, J., and Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Research 36: 3420–3435.

Grabherr, M.G. et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech 29: 644–652.

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.

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. U.S.A. 105: 18424–18429.

Hakimi, M.A., Privat, I., Valay, J.G., and Lerbs-Mache, S. (2000). Evolutionary conservation of C-terminal domains of primary sigma(70)-type transcription factors between plants and bacteria. J. Biol. Chem. 275: 9215–9221.

Hammani, K., Okuda, K., Tanz, S.K., Chateigner-Boutin, A.-L., Shikanai, T., and Small, I. (2009). A study of new *Arabidopsis* chloroplast RNA editing mutants reveals general features of editing factors and their target sites. Plant Cell 21: 3686–3699.

Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S.S., Topol, E.J., Levy, S., and Frazer, K.A. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biology 10: R32.

Hayes, M.L., Giang, K., and Mulligan, R.M. (2012). Molecular evolution of pentatricopeptide repeat genes reveals truncation in species lacking an editing target and structural domains under distinct selective pressures. BMC Evolutionary Biology 12: 66.

Helmann, J.D. and Chamberlin, M.J. (1988). Structure and function of bacterial sigma factors. Annu. Rev. Biochem. 57: 839–872.

Hou, R., Bao, Z., Wang, S., Su, H., Li, Y., Du, H., Hu, J., Wang, S., and Hu, X. (2011). Transcriptome sequencing and *de novo* analysis for Yesso Scallop (*Patinopecten yessoensis*) using 454 GS FLX. PLoS ONE 6: e21560.

Isono, K., Shimizu, M., Yoshimoto, K., Niwa, Y., Satoh, K., Yokota, A., and Kobayashi, H. (1997). Leaf-specifically expressed genes for polypeptides destined for chloroplasts with domains of sigma70 factors of bacterial RNA polymerases in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. U.S.A. 94: 14948–14953.

Kaur, S., Cogan, N.O., Pembleton, L.W., Shinozuka, M., Savin, K.W., Materne, M., and Forster, J.W. (2011). Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery. BMC Genomics 12: 265.

Kircher, M. and Kelso, J. (2010). High-throughput DNA sequencing--concepts and limitations. Bioessays 32: 524–536.

Kotera, E., Tasaka, M., and Shikanai, T. (2005). A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. Nature 433: 326–330.

Kozik, A., M, M., Kozik, I., Van, L.H., Van, D.A., and Michelmore, R. (2008). Eukaryotic ultra conserved orthologs and estimation of gene capture in EST librarie. Plant and Animal Genomes Conference 2008 Vol 16

Kumar, S. and Blaxter, M.L. (2010). Comparing *de novo* assemblers for 454 transcriptome data. BMC Genomics 11: 571.

Lamesch, P. et al. (2012). The *Arabidopsis* information resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 40: D1202–1210.

Langmead, B. and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Meth 9: 357–359.

Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat. Methods 7: 709–715.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009a). The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079.

Li, R. et al. (2009b). *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res.

Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008). SOAP: short oligonucleotide alignment program. Bioinformatics 24: 713–714.

Li, W. and Godzik, A. (2006). CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658–1659.

Logacheva, M.D., Kasianov, A.S., Vinogradov, D.V., Samigullin, T.H., Gelfand, M.S., Makeev, V.J., and Penin, A.A. (2011). *De novo* sequencing and characterization of floral transcriptome in two species of buckwheat (*Fagopyrum*). BMC Genomics 12: 30.

Lurin, C. et al. (2004). Genome-wide analysis of *Arabidopsis* pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. Plant Cell 16: 2089–2103.

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

Lysenko, E.A. (2007). Plant sigma factors and their role in plastid transcription. Plant Cell Rep. 26: 845–859.

Margam, V.M. et al. (2011). Transcriptome sequencing, and rapid development and application of SNP markers for the legume pod borer *Maruca vitrata* (Lepidoptera: Crambidae). PLoS ONE 6: e21388.

Margulies, M. et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.

Martin, J.A. and Wang, Z. (2011). Next-generation transcriptome assembly. Nat Rev Genet 12: 671–682.

Metzker, M.L. (2010). Sequencing technologies - the next generation. Nat. Rev. Genet. 11: 31–46.

Meyer, E., Aglyamova, G.V., Wang, S., Buchanan-Carter, J., Abrego, D., Colbourne, J.K., Willis, B.L., and Matz, M.V. (2009). Sequencing and *de novo* analysis of a coral larval transcriptome using 454 GSFlx. BMC Genomics 10: 219.

Mower, J.P., Touzet, P., Gummow, J.S., Delph, L.F., and Palmer, J.D. (2007). Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. BMC Evol Biol 7: 135.

MS Barker, H.V. Paleopolyploidy in the Brassicales: analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. Genome Biology and Evolution, v.2009, 391-399 (2009).

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344–1349.

Natarajan, P. and Parani, M. (2011). *De novo* assembly and transcriptome analysis of five major tissues of *Jatropha curcas L.* using GS FLX titanium platform of 454 pyrosequencing. BMC Genomics 12: 191.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. 10: 1–6.

Novaes, E., Drost, D.R., Farmerie, W.G., Pappas, G.J., Grattapaglia, D., Sederoff, R.R., and Kirst, M. (2008). High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics 9: 312.

Okuda, K., Chateigner-Boutin, A.-L., Nakamura, T., Delannoy, E., Sugita, M., Myouga, F., Motohashi, R., Shinozaki, K., Small, I., and Shikanai, T. (2009). Pentatricopeptide repeat proteins with the DYW motif have distinct molecular functions in RNA editing and RNA cleavage in *Arabidopsis* chloroplasts. Plant Cell 21: 146–156.

Okuda, K., Myouga, F., Motohashi, R., Shinozaki, K., and Shikanai, T. (2007). Conserved domain structure of pentatricopeptide repeat proteins involved in chloroplast RNA editing. Proc. Natl. Acad. Sci. U.S.A. 104: 8178–8183.

Palmer, J.D., Baldauf, S.L., Calie, P.J., and DePamphilis, C.W. (1990a). Chloroplast gene instability and transfer to the nucleus. 122: 97–106.

Palmer, J.D., Calie, P.J., dePamphilis, C.W., Jr, J.M.L., Katz-Downie, D.S., and Downie, S.R. (1990b). An evolutionary genetic approach to understanding plastid gene function: lessons from photosynthetic and nonphotosynthetic plants. In Current Research in Photosynthesis, M. Baltscheffsky, ed (Springer Netherlands), pp. 2381–2388.

Palmer, J.D., Nugent, J.M., and Herbon, L.A. (1987). Unusual structure of *Geranium* chloroplast DNA: A triple-sized inverted repeat, extensive gene duplications, multiple inversions, and two repeat families. Proc. Natl. Acad. Sci. U.S.A. 84: 769–773.

Parkinson, C.L., Mower, J.P., Qiu, Y.-L., Shirk, A.J., Song, K., Young, N.D., dePamphilis, C.W., and Palmer, J.D. (2005). Multiple major increases and decreases in mitochondrial substitution rates in the plant family Geraniaceae. BMC Evolutionary Biology 5: 73.

Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. Nat. Methods 6: S22–32.

Robbins, J.C., Heller, W.P., and Hanson, M.R. (2009). A comparative genomics approach identifies a PPR-DYW protein that is essential for C-to-U editing of the *Arabidopsis* chloroplast *accD* transcript. RNA 15: 1142–1153.

Roberts, S.B., Hauser, L., Seeb, L.W., and Seeb, J.E. (2012). Development of genomic resources for Pacific herring through targeted transcriptome pyrosequencing. PLoS ONE 7: e30908.

Rosenkranz, R., Borodina, T., Lehrach, H., and Himmelbauer, H. (2008). Characterizing the mouse ES cell transcriptome with Illumina sequencing. Genomics 92: 187–194.

Savory, E.A., Adhikari, B.N., Hamilton, J.P., Vaillancourt, B., Buell, C.R., and Day, B. (2012). mRNA-Seq analysis of the *Pseudoperonospora cubensis* transcriptome during cucumber (*Cucumis sativus* L.) Infection. PLoS ONE 7: e35796.

Schmitz-Linneweber, C. and Small, I. (2008). Pentatricopeptide repeat proteins: a socket set for organelle gene expression. Trends Plant Sci. 13: 663–670.

Shi, C.-Y., Yang, H., Wei, C.-L., Yu, O., Zhang, Z.-Z., Jiang, C.-J., Sun, J., Li, Y.-Y., Chen, Q., Xia, T., and Wan, X.-C. (2011). Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. BMC Genomics 12: 131.

Sloan, D.B., Barr, C.M., Olson, M.S., Keller, S.R., and Taylor, D.R. (2008). Evolutionary rate variation at multiple levels of biological organization in plant mitochondrial DNA. Mol. Biol. Evol. 25: 243–246.

Sloan, D.B., MacQueen, A.H., Alverson, A.J., Palmer, J.D., and Taylor, D.R. (2010). Extensive loss of RNA editing sites in rapidly evolving *Silene* mitochondrial genomes: Selection vs. Retroprocessing as the Driving Force. Genetics 185: 1369–1380.

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

Small, I.D. and Peeters, N. (2000). The PPR motif - a TPR-related motif prevalent in plant organellar proteins. Trends Biochem. Sci. 25: 46–47.

Sugiura, C., Kobayashi, Y., Aoki, S., Sugita, C., and 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 Res. 31: 5324–5331.

Tariq, M.A., Kim, H.J., Jejelowo, O., and Pourmand, N. (2011). Whole-transcriptome RNAseq analysis from minute amount of total RNA. Nucl. Acids Res.: gkr547.

Tseng, C.-C., Sung, T.-Y., Li, Y.-C., Hsu, S.-J., Lin, C.-L., and Hsieh, M.-H. (2010). Editing of *accD* and *ndhF* chloroplast transcripts is partially affected in the *Arabidopsis vanilla cream1* mutant. Plant Mol. Biol. 73: 309–323.

Vega-Arreguín, J.C., Ibarra-Laclette, E., Jiménez-Moraila, B., Martínez, O., Vielle-Calzada, J.P., Herrera-Estrella, L., and Herrera-Estrella, A. (2009). Deep sampling of the *Palomero* maize transcriptome by a high throughput strategy of pyrosequencing. BMC Genomics 10: 299.

Wall, P.K. et al. (2009). Comparison of next generation sequencing technologies for transcriptome characterization. BMC Genomics 10: 347.

Wang, X.-W., Luan, J.-B., Li, J.-M., Bao, Y.-Y., Zhang, C.-X., and Liu, S.-S. (2010). *De novo* characterization of a whitefly transcriptome and analysis of its gene expression during development. BMC Genomics 11: 400.

Ward, J.A., Ponnala, L., and Weber, C.A. (2012). Strategies for transcriptome analysis in nonmodel plants. Am. J. Bot. 99: 267–276.

Weber, A.P.M., Weber, K.L., Carr, K., Wilkerson, C., and Ohlrogge, J.B. (2007). Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. Plant Physiol. 144: 32–42.

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.

Wenping, H., Yuan, Z., Jie, S., Lijun, Z., and Zhezhi, W. (2011). *De novo* transcriptome sequencing in *Salvia miltiorrhiza* to identify genes involved in the biosynthesis of active ingredients. Genomics 98: 272–279.

Wheat, C.W. (2010). Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. Genetica 138: 433–451.

Wickett, N.J., Honaas, L.A., Wafula, E.K., Das, M., Huang, K., Wu, B., Landherr, L., Timko, M.P., Yoder, J., Westwood, J.H., and dePamphilis, C.W. (2011). Transcriptomes of the parasitic plant family Orobanchaceae reveal surprising conservation of chlorophyll synthesis. Curr. Biol. 21: 2098–2104.

Yu, Q.-B., Jiang, Y., Chong, K., and Yang, Z.-N. (2009). *AtECB2*, a pentatricopeptide repeat protein, is required for chloroplast transcript *accD* RNA editing and early chloroplast biogenesis in *Arabidopsis thaliana*. Plant J. 59: 1011–1023.

Zhou, W., Cheng, Y., Yap, A., Chateigner-Boutin, A.-L., Delannoy, E., Hammani, K., Small, I., and Huang, J. (2009). The *Arabidopsis* gene *YS1* encoding a DYW protein is required for editing of *rpoB* transcripts and the rapid development of chloroplasts during early growth. Plant J. 58: 82–96.

CHAPTER 3

Agrafioti, I., Swire, J., Abbott, J., Huntley, D., Butcher, S., and Stumpf, M.P.H. (2005). Comparative analysis of the *Saccharomyces cerevisiae* and *Caenorhabditis elegans* protein interaction networks. BMC Evol. Biol. 5: 23.

Barreto, F.S. and Burton, R.S. (2013). Evidence for compensatory evolution of ribosomal proteins in response to rapid divergence of mitochondrial rRNA. Mol. Biol. Evol. 30: 310–314.

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

Campo, D.S., Dimitrova, Z., Mitchell, R.J., Lara, J., and Khudyakov, Y. (2008). Coordinated evolution of the hepatitis C virus. Proc. Natl. Acad. Sci. USA 105: 9685–9690.

Chen, Y. and Dokholyan, N.V. (2006). The coordinated evolution of yeast proteins is constrained by functional modularity. Trends Genet. 22: 416–419.

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* x *hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Mol. Biol. Evol. 23: 2175–2190.

Clark, N.L., Alani, E., and Aquadro, C.F. (2012). Evolutionary rate covariation reveals shared functionality and coexpression of genes. Genome Res. 22: 714–720.

Clark, N.L. and Aquadro, C.F. (2010). A novel method to detect proteins evolving at correlated rates: identifying new functional relationships between coevolving proteins. Mol. Biol. Evol. 27: 1152–1161.

Davis, J.C. and Petrov, D.A. (2004). Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol. 2: e55.

Drouin, G., Daoud, H., and Xia, J. (2008). Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Mol. Phylogenet. Evol. 49: 827–831.

Duarte, J.M., Wall, P.K., Edger, P.P., Landherr, L.L., Ma, H., Pires, J.C., Leebens-Mack, J., and dePamphilis, C.W. (2010). Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. BMC Evol. Biol. 10: 61.

Durand, D., Halldórsson, B.V., and Vernot, B. (2006). A hybrid micro-macroevolutionary approach to gene tree reconstruction. J. Comput. Biol. 13: 320–335.

Dutheil, J. and Galtier, N. (2007). Detecting groups of coevolving positions in a molecule: a clustering approach. BMC Evol. Biol. 7: 242.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792–1797.

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

Fares, M.A. and Wolfe, K.H. (2003). Positive selection and subfunctionalization of duplicated CCT chaperonin subunits. Mol. Biol. Evol. 20: 1588–1597.

Favory, J.-J., Kobayshi, M., Tanaka, K., Peltier, G., Kreis, M., Valay, J.-G., and Lerbs-Mache, S. (2005). Specific function of a plastid sigma factor for *ndhF* gene transcription. Nucleic Acids Res. 33: 5991–5999.

Fraser, H.B., Hirsh, A.E., Wall, D.P., and Eisen, M.B. (2004). Coevolution of gene expression among interacting proteins. Proc. Natl. Acad. Sci. USA 101: 9033–9038.

Göbel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. Proteins 18: 309–317.

Greiner, S., Rauwolf, U., Meurer, J., and Herrmann, R.G. (2011). The role of plastids in plant speciation. Mol. Ecol. 20: 671–691.

Greiner, S., Wang, X., Herrmann, R.G., Rauwolf, U., Mayer, K., Haberer, G., and Meurer, J. (2008). The complete nucleotide sequences of the 5 genetically distinct plastid genomes of *Oenothera*, subsection *Oenothera*: II. A microevolutionary view using bioinformatics and formal genetic data. Mol. Biol. Evol. 25: 2019–2030.

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. USA 105: 18424–18429.

Hu, J. and Yan, C. (2009). A tool for calculating binding-site residues on proteins from PDB structures. BMC Struct. Biol. 9: 52.

De Juan, D., Pazos, F., and Valencia, A. (2013). Emerging methods in protein co-evolution. Nat. Rev. Genet. 14: 249–261.

Katoh, K. and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30: 772–780.

Lamesch, P. et al. (2012). The *Arabidopsis* information resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 40: D1202–1210.

Larkin, M.A. et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948.

Lerbs-Mache, S. (2011). Function of plastid sigma factors in higher plants: regulation of gene expression or just preservation of constitutive transcription? Plant Mol. Biol. 76: 235–249.

Li, L., Huang, Y., Xia, X., and Sun, Z. (2006). Preferential duplication in the sparse part of yeast protein interaction network. Mol. Biol. Evol 23: 2467–2473.

Lobo, F.P., Mota, B.E.F., Pena, S.D.J., Azevedo, V., Macedo, A.M., Tauch, A., Machado, C.R., and Franco, G.R. (2009). Virus-host coevolution: common patterns of nucleotide motif usage in *Flaviviridae* and their hosts. PLoS ONE 4: e6282.

Lovell, S.C. and Robertson, D.L. (2010). An integrated view of molecular coevolution in protein–protein interactions. Mol. Biol. Evol. 27: 2567–2575.

Marchler-Bauer, A. et al. (2013). CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Res. 41: D348–352.

Mintseris, J. and Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein–protein interactions. Proc. Natl. Acad. Sci. USA 102: 10930–10935.

Neher, E. (1994). How frequent are correlated changes in families of protein sequences? Proc. Natl. Acad. Sci. USA 91: 98–102.

Osada, N. and Akashi, H. (2012). Mitochondrial–nuclear interactions and accelerated compensatory evolution: evidence from the primate cytochrome c oxidase complex. Mol. Biol. Evol. 29: 337–346.

Pazos, F., Helmer-Citterich, M., Ausiello, G., and Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. J. Mol. Biol. 271: 511–523.

Pazos, F., Ranea, J.A.G., Juan, D., and Sternberg, M.J.E. (2005). Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J. Mol. Biol. 352: 1002–1015.

Pazos, F. and Valencia, A. (2008). Protein co-evolution, co-adaptation and interactions. EMBO J. 27: 2648–2655.

Pazos, F. and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng. 14: 609–614.

Pond, S.L.K., Frost, S.D.W., and Muse, S.V. (2005). HyPhy: hypothesis testing using phylogenies. Bioinformatics 21: 676–679.

Rao, V.S., Srinivas, K., Sujini, G.N., and Kumar, G.N.S. (2014). Protein-protein interaction detection: Methods and analysis. International Journal of Proteomics 2014: e147648.

Ruckle, M.E., DeMarco, S.M., and Larkin, R.M. (2007). Plastid signals remodel light signaling networks and are essential for efficient chloroplast biogenesis in Arabidopsis. Plant Cell 19: 3944–3960.

Sato, T., Yamanishi, Y., Kanehisa, M., and Toh, H. (2005). The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics 21: 3482–3489.

Schmitz-Linneweber, C., Kushnir, S., Babiychuk, E., Poltnigg, P., Herrmann, R.G., and Maier, R.M. (2005). Pigment deficiency in nightshade/tobacco cybrids is caused by the failure to edit the plastid ATPase α-subunit mRNA. Plant Cell 17: 1815–1828.

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

Sloan, D.B., Triant, D.A., Wu, M., and Taylor, D.R. (2014). Cytonuclear interactions and relaxed selection accelerate sequence evolution in organelle ribosomes. Mol. Biol. Evol. 31: 673–682.

Smith, L. (1915). Variegation in *Pelargonium*. Proceedings of the Royal Horticultural Society 41: 36.

Sobolev, V., Eyal, E., Gerzon, S., Potapov, V., Babor, M., Prilusky, J., and Edelman, M. (2005). SPACE: a suite of tools for protein structure prediction and analysis based on complementarity and environment. Nucleic Acids Res. 33: W39–W43.

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

Subramanian, S. and Kumar, S. (2004). Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics 168: 373–381.

Tan, S. and Troxler, R.F. (1999). Characterization of two chloroplast RNA polymerase sigma factors from *Zea mays*: photoregulation and differential expression. Proc. Natl. Acad. Sci. USA 96: 5316–5321.

Taylor, W.R. and Hatrick, K. (1994). Compensating changes in protein multiple sequence alignments. Protein Eng. 7: 341–348.

Touloumenidou, T., Bakker, F.T., and Albers, F. (2007). The phylogeny of *Monsonia* L. (Geraniaceae). Plant Syst. Evol. 264: 1–14.

Weihe, A., Apitz, J., Pohlheim, F., Salinas-Hartwig, A., and Börner, T. (2009). Biparental inheritance of plastidial and mitochondrial DNA and hybrid variegation in *Pelargonium*. Mol. Genet. Genomics 282: 587–593.

Weng, M.-L., Blazier, J.C., Govindu, M., and Jansen, R.K. (2014). Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. Mol. Biol. Evol. 31: 645–659.

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.

Widler-Kiefer, H. and Yeo, P.F. (1987). Fertility relationships of *Geranium* (Geraniaceae): *sect. Ruberta*, *Anemonifolia*, *Lucida* and *Unguiculata*. Pl. Syst. Evol. 155: 283–306.

Willett, C.S. and Burton, R.S. (2004). Evolution of interacting proteins in the mitochondrial electron transport system in a marine copepod. Mol. Biol. Evol. 21: 443–453.

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. USA 84: 9054–9058.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24: 1586–1591.

Yao, J.-L. and Cohen, D. (2000). Multiple gene control of plastome-genome incompatibility and plastid DNA inheritance in interspecific hybrids of *Zantedeschia*. Theor. Appl. Genet. 101: 400–406.

Yeang, C.-H. and Haussler, D. (2007). Detecting coevolution in and among protein domains. PLoS Comput. Biol. 3: e211.

Yu, S.-N. and Horn, W. a. H. (1988). Additional chromosome numbers in *Pelargonium* (Geraniaceae). Pl. Syst. Evol. 159: 165–171.

Zhang, F. and Broughton, R.E. (2013). Mitochondrial–nuclear interactions: compensatory evolution or variable functional constraint among vertebrate oxidative phosphorylation genes? Genome Biol. Evol. 5: 1781–1791.

Zhang, J., Ruhlman, T.A., Mower, J.P., and Jansen, R.K. (2013). Comparative analyses of two Geraniaceae transcriptomes using next-generation sequencing. BMC Plant Biol. 13: 228.

CHAPTER 4

Barreto, F.S. and Burton, R.S. (2013). Evidence for compensatory evolution of ribosomal proteins in response to rapid divergence of mitochondrial rRNA. Mol. Biol. Evol. 30: 310–314.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27: 573–580.

Bernt, M., Merkle, D., Ramsch, K., Fritzsch, G., Perseke, M., Bernhard, D., Schlegel, M., Stadler, P., and Middendorf, M. (2005). CREx: inferring genomic rearrangements based on common intervals.

Blazier, J., 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, Topics in Current Genetics. (Springer Berlin Heidelberg), pp. 29–63.

Boesch, P., Weber-Lotfi, F., Ibrahim, N., Tarasenko, V., Cosset, A., Paulus, F., Lightowlers, R.N., and Dietrich, A. (2011). DNA repair in organelles: pathways, organization, regulation, relevance in disease and aging. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research 1813: 186–200.

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

Carrie, C., Kühn, K., Murcha, M.W., Duncan, O., Small, I.D., O'Toole, N., and Whelan, J. (2009). Approaches to defining dual-targeted proteins in *Arabidopsis*. Plant J. 57: 1128–1139.

Cho, H.S., Lee, S.S., Kim, K.D., Hwang, I., Lim, J.-S., Park, Y.-I., and Pai, H.-S. (2004). DNA gyrase is involved in chloroplast nucleoid partitioning. Plant Cell 16: 2665–2682.

Chumley, T.W., Palmer, J.D., Mower, J.P., Fourcade, H.M., Calie, P.J., Boore, J.L., and Jansen, R.K. (2006). The complete chloroplast genome sequence of *Pelargonium* x *hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Molecular Biology and Evolution 23: 2175–2190.

Cosner, M.E., Raubeson, L.A., and Jansen, R.K. (2004). Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. BMC Evolutionary Biology 4: 27.

Cox, M.M. (2007). Regulation of bacterial *RecA* protein function. Crit. Rev. Biochem. Mol. Biol. 42: 41–63.

Darling, A.E., Mau, B., and Perna, N.T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS ONE 5: e11147.

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

Drouin, G., Daoud, H., and Xia, J. (2008). Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Mol. Phylogenet. Evol. 49: 827–831.

Duarte, J.M., Wall, P.K., Edger, P.P., Landherr, L.L., Ma, H., Pires, J.C., Leebens-Mack, J., and dePamphilis, C.W. (2010). Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. BMC Evolutionary Biology 10: 61.

Fajardo, D., Senalik, D., Ames, M., Zhu, H., Steffan, S.A., Harbut, R., Polashock, J., Vorsa, N., Gillespie, E., Kron, K., and Zalapa, J.E. (2012). Complete plastid genome sequence of *Vaccinium macrocarpon*: structure, gene content, and rearrangements revealed by next generation sequencing. Tree Genetics & Genomes 9: 489–498.

Gaut, B.S. (1998). Molecular clocks and nucleotide substitution rates in higher plants. In Evolutionary Biology, M.K. Hecht, R.J. Macintyre, and M.T. Clegg, eds, Evolutionary Biology. (Springer US), pp. 93–120.

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.

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. U.S.A. 105: 18424–18429.

Gutman, B.L. and Niyogi, K.K. (2009). Evidence for base excision repair of oxidative DNA damage in chloroplasts of *Arabidopsis thaliana*. J. Biol. Chem. 284: 17006–17012.

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.

Hsu, D.S., Kim, S.-T., Sun, Q., and Sancar, A. (1995). Structure and function of the *UvrB* protein. J. Biol. Chem. 270: 8319–8327.

Huang, C.-C., Hung, K.-H., Wang, W.-K., Ho, C.-W., Huang, C.-L., Hsu, T.-W., Osada, N., Hwang, C.-C., and Chiang, T.-Y. (2012). Evolutionary rates of commonly used nuclear and organelle markers of *Arabidopsis* relatives (Brassicaceae). Gene 499: 194–201.

Hübscher, U., Maga, G., and Podust, V.N. (1996). DNA replication accessory proteins. In DNA Replication in Eukaryotic Cells (De Pamphilis, M. L., ed) (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY), pp. 525–543.

Ishibashi, T., Kimura, S., and Sakaguchi, K. (2006). A higher plant has three different types of RPA heterotrimeric complex. J Biochem 139: 99–104.

Jansen, R.K. et al. (2007). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. PNAS 104: 19369–19374.

Jansen, R.K. and Ruhlman, T.A. (2012). Plastid genomes of seed plants. In Genomics of Chloroplasts and Mitochondria, R. Bock and V. Knoop, eds, Advances in Photosynthesis and Respiration. (Springer Netherlands), pp. 103–126.

Kagale, S., Robinson, S.J., Nixon, J., Xiao, R., Huebert, T., Condie, J., Kessler, D., Clarke, W.E., Edger, P.P., Links, M.G., Sharpe, A.G., and Parkin, I.A.P. (2014). Polyploid evolution of the Brassicaceae during the Cenozoic Era. Plant Cell: tpc.114.126391.

Katoh, K. and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30: 772–780.

Knox, E.B. (2014). The dynamic history of plastid genomes in the Campanulaceae sensu lato is unique among angiosperms. PNAS 111: 11097–11102.

Kornberg, A. and Baker, T.A. (2005). DNA Replication (University Science Books).

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

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

Martínez-Alberola, F., del Campo, E.M., Lázaro-Gimeno, D., Mezquita-Claramonte, S., Molins, A., Mateu-Andrés, I., Pedrola-Monfort, J., Casano, L.M., and Barreno, E. (2013). Balanced gene losses, duplications and intensive rearrangements led to an unusual regularly sized genome in *Arbutus unedo* chloroplasts. PLoS ONE 8: e79685.

Milligan, B.G., Hampton, J.N., and Palmer, J.D. (1989). Dispersed repeats and structural reorganization in subclover chloroplast DNA. Molecular Biology and Evolution.

Mori, Y., Kimura, S., Saotome, A., Kasai, N., Sakaguchi, N., Uchiyama, Y., Ishibashi, T., Yamamoto, T., Chiku, H., and Sakaguchi, K. (2005). Plastid DNA polymerases from higher plants, *Arabidopsis thaliana*. Biochem. Biophys. Res. Commun. 334: 43–50.

Odahara, M., Masuda, Y., Sato, M., Wakazaki, M., Harada, C., Toyooka, K., and Sekine, Y. (2015). *RECG* maintains plastid and mitochondrial genome stability by suppressing extensive recombination between short dispersed repeats. PLoS Genet 11: e1005080.

Pang, Q., Hays, J.B., and Rajagopal, I. (1993). Two cDNAs from the plant *Arabidopsis thaliana* that partially restore recombination proficiency and DNA-damage resistance to *E. coli* mutants lacking recombination-intermediate-resolution activities. Nucleic Acids Res 21: 1647–1653.

Pazos, F., Helmer-Citterich, M., Ausiello, G., and Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. J. Mol. Biol. 271: 511–523.

Pazos, F., Ranea, J.A.G., Juan, D., and Sternberg, M.J.E. (2005). Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J. Mol. Biol. 352: 1002–1015.

Pazos, F. and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng. 14: 609–614.

Perry, A.S., Brennan, S., Murphy, D.J., Kavanagh, T.A., and Wolfe, K.H. (2002). Evolutionary re-organisation of a large operon in Adzuki bean chloroplast DNA caused by inverted repeat movement. DNA Res 9: 157–162.

Ruhlman, T.A., Chang, W.-J., Chen, J.J., Huang, Y.-T., Chan, M.-T., Zhang, J., Liao, D.-C., Blazier, J.C., Jin, X., Shih, M.-C., Jansen, R.K., and Lin, C.-S. (2015). *NDH* expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. BMC Plant Biology 15: 100.

Ruhlman, T.A. and Jansen, R.K. (2014). The plastid genomes of flowering plants. Methods Mol. Biol. 1132: 3–38.

Sabir, J., Schwarz, E., Ellison, N., Zhang, J., Baeshen, N.A., Mutwakil, M., Jansen, R., and Ruhlman, T. (2014). Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. Plant Biotechnol J 12: 743–754.

Samach, A., Melamed-Bessudo, C., Avivi-Ragolski, N., Pietrokovski, S., and Levy, A.A. (2011). Identification of plant *RAD52* homologs and characterization of the *Arabidopsis thaliana RAD52-Like* Genes. Plant Cell 23: 4266–4279.

Saotome, A., Kimura, S., Mori, Y., Uchiyama, Y., Morohashi, K., and Sakaguchi, K. (2006). Characterization of four *RecQ* homologues from rice (*Oryza sativa* L. cv. Nipponbare). Biochemical and Biophysical Research Communications 345: 1283–1291.

Shedge, V., Arrieta-Montiel, M., Christensen, A.C., and Mackenzie, S.A. (2007). Plant mitochondrial recombination surveillance requires unusual *RecA* and *MutS* homologs. Plant Cell 19: 1251–1264.

Sloan, D.B., Triant, D.A., Wu, M., and Taylor, D.R. (2014). Cytonuclear interactions and relaxed selection accelerate sequence evolution in organelle ribosomes. Mol Biol Evol 31: 673–682.

Sorhannus, U. and Fox, M. (1999). Synonymous and nonsynonymous substitution rates in diatoms: a comparison between chloroplast and nuclear genes. J Mol Evol 48: 209–212.

Tesler, G. (2002a). Efficient algorithms for multichromosomal genome rearrangements. Journal of Computer and System Sciences 65: 587–609.

Tesler, G. (2002b). GRIMM: genome rearrangements web server. Bioinformatics 18: 492–493.

Wall, M.K., Mitchenall, L.A., and Maxwell, A. (2004). *Arabidopsis thaliana* DNA gyrase is targeted to chloroplasts and mitochondria. Proc. Natl. Acad. Sci. U.S.A. 101: 7821–7826.

Weng, M.-L., Blazier, J.C., Govindu, M., and Jansen, R.K. (2014). Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. Mol Biol Evol 31: 645–659.

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. U.S.A. 84: 9054–9058.

Wolf, P.G. (2012). Plastid genome diversity. In Plant Genome Diversity Volume 1, J.F. Wendel, J. Greilhuber, J. Dolezel, and I.J. Leitch, eds (Springer Vienna), pp. 145–154.

Xu, Y.-Z., Arrieta-Montiel, M.P., Virdi, K.S., de Paula, W.B.M., Widhalm, J.R., Basset, G.J., Davila, J.I., Elthon, T.E., Elowsky, C.G., Sato, S.J., Clemente, T.E., and Mackenzie, S.A. (2011). *MutS HOMOLOG1* is a nucleoid protein that alters mitochondrial and plastid properties and plant response to high light. Plant Cell 23: 3428–3441.

Ye, J. and Sayre, R.T. (1990). Reduction of chloroplast DNA content in *Solanum nigrum* suspension cells by treatment with chloroplast DNA synthesis inhibitors 1. Plant Physiol 94: 1477–1483.

Zaegel, V., Guermann, B., Le Ret, M., Andrés, C., Meyer, D., Erhardt, M., Canaday, J., Gualberto, J.M., and Imbault, P. (2006). The plant-specific ssDNA binding protein *OSB1* is involved in the stoichiometric transmission of mitochondrial DNA in *Arabidopsis*. Plant Cell 18: 3548–3563.

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

Zhang, J., Ruhlman, T.A., Mower, J.P., and Jansen, R.K. (2013). Comparative analyses of two Geraniaceae transcriptomes using next-generation sequencing. BMC Plant Biology 13: 228.

Zhang, J., Ruhlman, T.A., Sabir, J., Blazier, J.C., and Jansen, R.K. (2015). Coordinated rates of evolution between interacting plastid and nuclear genes in Geraniaceae. Plant Cell 27: 563–573.