# Mutation Rate Variation and Organelle Genome Evolution in the Angiosperm Genus *Silene*

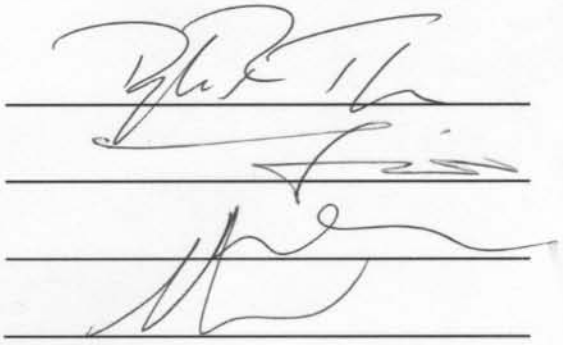
Daniel Benjamin Sloan
Kennebunk, Maine

B.A., Wesleyan University, 2003

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Biology

University of Virginia
May, 2011



## Abstract

The defining challenge to the field of molecular evolution in the genomic era is to identify the evolutionary forces that have shaped the striking diversity in genome size, structure, and organization across the tree of life. The mutational burden hypothesis offers a potentially unifying framework for explaining much of this diversity based on variation in the intensity and efficacy of selection associated with mutational processes. One key prediction from this hypothesis is that high mutation environments should select against large and complex genomes, because they are more susceptible to being altered or disrupted by mutation. For example, it has been argued that the abnormally large and complex mitochondrial genomes found in angiosperms reflect a history of relaxed selection resulting from the extremely low point mutation rates in these genomes. This dissertation establishes the angiosperm genus *Silene* (Caryophyllaceae) as an ideal system for testing this prediction. The genus *Silene* is shown to harbor dramatic variation in mitochondrial substitution rates both among genes and among evolutionary lineages. Several *Silene* species exhibit genome-wide accelerations of approximately 100-fold in mitochondrial substitution rates, which appear to reflect changes in underlying mutation rates. A comparative analysis of complete organelle genome sequences from four *Silene* species with highly divergent mitochondrial mutation rates shows that mutational acceleration has been associated with dramatic changes in mitochondrial genome architecture as well as some correlated differences in the rate of sequence and structural evolution in plastid genomes. However, many of the observed changes

run counter to the predictions of the mutational burden hypothesis. In particular, rather than showing evidence of streamlining, the fast-evolving mitochondrial genomes have experienced a massive proliferation in non-coding content, resulting in the largest mitochondrial genomes ever identified. The specific mechanisms underlying the observed differences in mutation rate and genome architecture remain unknown, but evidence for changes in recombinational processes within these genomes motivate the hypothesis that disruptions in the nuclear-encoded organelle recombination machinery may be responsible. Further unraveling the process of rapid extreme change in *Silene* organelles genomes should provide new insights into the evolutionary forces responsible for the tremendous variation in eukaryotic genome size and complexity.

Acknowledgements

This dissertation is the product of a highly collaborative effort, and I have many people to thank for guiding me through this process. I will begin with my advisor, Doug Taylor, whose support has been unwavering throughout this entire process. Regardless of the question or topic, Doug is always intellectually engaged, and he has been incredibly adept and dogged at dispensing with logistical hurdles, allowing me to focus squarely on the science. There were times when I seriously doubted that I would get to this point. With a different advisor, I might not have made it.

One of Doug's greatest skills as an advisor is in fostering a lab atmosphere that is both intellectual and social. When I first visited and met the grad students and postdocs in the Taylor lab (which blends rather seamlessly into the Antonovics lab), I knew that this was a place I wanted to be. The names and faces have almost all turned over, but the culture has remained the same. My Taylor/Antonovics labmates have been great friends, colleagues and collaborators, making this process both successful and enjoyable.

Perhaps the most stimulating aspect of the lab environment has been the extent to which students work on such diverse questions. Doug gives his students a great deal of freedom to develop their own projects. As a result, I often found myself in unchartered waters with respect to the expertise of the lab, and a large component of my training was accomplished with the help of outside collaborators and mentors. This begins with the members of my committee, Janis Antonovics, Stefan Bekiranov, Butch Brodie, Lei Li, and Martin Wu, who provided valuable input on planning and executing this research throughout. Collaborations also extended outside UVA. Dave McCauley, Matt Olson, Bengt Oxelman, and Helena Štorchová have all made important contributions to

individual projects, and I would especially like to thank Jeff Palmer and Andy Alverson. They have been incredibly generous with their time and energy and have essentially become co-advisors during this process. I could not have done it without them.

The execution of my dissertation research also depended on many other people here at UVA. There are too many to name them all, but I would like to specifically thank John Chuckalovcak, Wendy Crannage, and Tony Spano. John has put in many hours generating 454 sequencing data for this project, Wendy did an amazing job keeping thousands and thousands of plants alive and thriving in the greenhouse, allowing me to worry about other things, and Tony was a constant source of advice and help with equipment and lab techniques.

Any dissertation project involves a large amount of luck, both good and bad. I feel that I have enjoyed more than my share of good luck during my time at UVA. In fact, my good fortune began even before coming to Virginia. Unknowingly, I applied to UVA during the first year that the Jefferson Scholars Foundation was awarding graduate fellowships in Biology. I was fortunate enough to receive one of these fellowships, and the financial support it provided has been instrumental in performing my dissertation research. This project also benefited from generous financial support from the Faculty Senate, the Office of the VPR, the Graduate School of Arts and Sciences, the Department of Biology, the Society of Fellows, the Society for Molecular Biology and Evolution, and the NSF.

Finally, I would like to thank my family. They have made all of this possible and worthwhile. Six years ago, my wife Janet made the sacrifice of leaving the Pacific Northwest, one of the most beautiful places in the world, to move to the middle of

Virginia, more than one hundred miles from her beloved ocean. She has been a constant source of love, support and motivation, for which I am deeply grateful. I would like to end where it all began, with my parents. They have given me so much. Shortly after I arrived in Virginia, my father was diagnosed with ALS. Even with my father's declining health, my parents continued to go to enormous lengths to allow me to pursue my education. They have left me with a debt that I can never repay but will always remember.

## Table of Contents

# Introduction

The field of evolutionary biology is founded on the simple observation that biological organisms differ, often in spectacular ways. Most evolutionary studies have focused on explaining phenotypic variation, including differences in morphology, physiology, cell structure, etc. In this context, the study of DNA is largely restricted to its role in defining the genetic basis of phenotypic variation. However, the physical organization of DNA into a genome constitutes a phenotype in its own right. Elements of the genomic phenotype, including total size, structure, and chromosome number (Lewitsky 1931; Dobzhansky and Sturtevant 1938; Swift 1950a; Swift 1950b; Mirsky and Ris 1951), were an object of study even before the experiments that definitively identified DNA as the molecule of heredity (Hershey and Chase 1952) and elucidated its molecular structure (Watson and Crick 1953). With the rapidly expanding availability of DNA sequence data and advances in our understanding of the molecular biology of the genome, we have entered an era in which it is increasingly possible to study the evolution of the genome, *per se*.

The wealth of genomic data now available highlights an incredible diversity in genome size, structural organization, and function across the tree of life (Gregory 2005). Genome sizes in cellular organisms have been found to span six orders of magnitude (McCutcheon et al. 2009; Pellicer et al. 2010) and perhaps more (Cavalier-Smith 1985; Gregory 2005). Often, this variation does not correlate with obvious differences in organismal complexity, but instead reflects enormous divergence in gene density and the abundance of non-coding content (Gregory 2005). The organization and functional expression of individual genes also varies widely across organisms (Lynch 2006). For

example, the expression of eukaryotic genes can involve a number of complex modifications during RNA maturation that are largely absent in their prokaryotic counterparts, such as intron splicing and RNA editing (Gilbert 1978; Knoop 2011).

Attempts to explain the observed diversity in gene and genome architecture have generated numerous hypotheses, many of which fall out along the lines of the classic neutralist-selectionist debate that has pervaded the field of molecular evolution for decades (Doolittle and Sapienza 1980; Orgel and Crick 1980; Cavalier-Smith 1982; Petrov 2002; Wagner 2005; Lynch 2007; Gray et al. 2010). One of the most significant contributions to this debate in the 21$^{st}$ century is the work of Michael Lynch and the development of the mutational burden hypothesis (Lynch 2002; Lynch and Conery 2003; Lynch 2006; Lynch 2006; Lynch et al. 2006; Lynch 2007). The core of this hypothesis is based on the idea that many genomic features create an added liability for the organism that is directly related to the probability of mutational disruption. For example, the insertion of an intron into an essential gene generates the requirement that the intron be spliced out from RNA transcripts. Because proper splicing depends on conservation at multiple nucleotide positions, there is a non-zero probability that, in any given generation, a mutation will occur that disrupts splicing and reduces organismal fitness (Lynch 2002). Likewise, essential RNA editing events also depend on recognition sequences and are, therefore, sensitive to mutational disruption (Farre et al. 2001; Mulligan et al. 2007). The expansion of intergenic regions creates a related but somewhat distinct mutational vulnerability. Specifically, mutations in intergenic sequences can generate deleterious features, including improper transcription factor binding sites and translation initiation sites (Hahn et al. 2003; Lynch et al. 2005).

The origins of the mutational burden hypothesis are largely rooted in comparisons between eukaryotic and prokaryotic genome architectures. Whereas bacterial and archaeal genomes are small and highly streamlined, eukaryotic nuclear genomes are often characterized by enormous quantities of non-coding sequence and highly complex gene architectures. The mutational burden hypothesis suggests that these patterns can be explained by the disparity in effective population sizes ($N_e$) between these groups. In comparison to their eukaryotic counterparts, bacteria and archaea often maintain a very large $N_e$ (Lynch and Conery 2003). Therefore, selection on weakly deleterious or beneficial alleles is expected to be substantially more effective in prokaryotes. Based on this population genetic framework, the mutational burden hypothesis interprets many of the complexities found in eukaryotic nuclear genomes not as adaptations but as the consequence of failed selection. Under this interpretation, the mutational liability generated by a genomic feature such as an (expanded) intergenic region, intron, or RNA editing site can be associated with a selection coefficient. The magnitude of this selection coefficient should be proportional to the mutation rate per base pair per generation and, therefore, extremely small. The mutational burden hypothesis holds that selection in most eukaryotes, particularly large multicellular species, has been too inefficient to effectively act on these weakly deleterious alleles, whereas more efficacious selection in prokaryotes has been capable of preventing the proliferation of non-coding content.

Mitochondrial genome diversity across eukaryotes poses an empirical difficulty for this line of reasoning. The mutational burden hypothesis was developed based on the expected effects of different levels of $N_e$. Therefore, it would be predicted that large, multicellular plants and animals (both of which typically have very low values of $N_e$)

would exhibit similar patterns in mitochondrial genome evolution. However, the mitochondrial genomes of land plants and bilaterian animals have evolved to opposite extremes within the continuum of diversity found in eukaryotes (Gray et al. 1999; Lynch et al. 2006). Bilaterian animals have small and highly streamlined genomes that are reminiscent of prokaryotic genome architecture in many ways (Boore 1999). In contrast, land plant mitochondrial genomes have experienced a proliferation of non-coding content (Mower et al. In press).

To reconcile existing theory with these seemingly contradictory observations, Lynch et al. (2006) generalized the mutational burden hypothesis. They argued that, in addition to being dependent on $N_e$ and the *efficacy* of selection against mutational liabilities, the evolution of genome architecture is also determined by the *intensity* of that selection, which should be directly proportional to the mutation rate. Although plants and animals have grossly similar levels of $N_e$, they exhibit highly divergent mitochondrial mutation rates (Wolfe et al. 1987). Animal mitochondrial DNA evolves very rapidly relative to both nuclear DNA and to mitochondrial genomes in other eukaryotes. In contrast, plant mitochondrial genomes exhibit some of the slowest rates of nucleotide substitution ever documented. Based on these differences, Lynch et al. (2006) concluded that the low mutation rates in plant mitochondria have relaxed the intensity of selection associated with mutational vulnerabilities, creating a permissive environment for the evolution of large and complex genomes.

The mutational burden hypothesis offers a sweeping and ambitious explanation for the extreme diversity in genome architecture across living organisms. If correct in its major arguments, it would represent a truly revolutionary advance in our understanding

of genome evolution. However, empirical support for this hypothesis is largely limited to comparisons across very broad phylogenetic scales that confound countless biological variables (e.g., prokaryotes vs. eukaryotes or plants vs. animals). The primary goal of this dissertation is to establish and use angiosperm mitochondrial genomes, particularly within the genus *Silene* (Caryophyllaceae), as a model for testing the predictions of the mutational burden hypothesis. A handful of angiosperm lineages have recently been identified as having experienced dramatic accelerations in mitochondrial mutation rates, resulting in orders of magnitude more sequence divergence than typically observed in plant mitochondrial genomes (Cho et al. 2004; Parkinson et al. 2005; Mower et al. 2007). These lineages include several *Silene* species, but remarkably many other closely related species within this genus maintain their historically low rates of evolution (Mower et al. 2007; Sloan et al. 2008; Sloan et al. 2009). Therefore, this genus offers a unique comparative system to test the predicted effects of mutational processes on genome evolution.

Chapter 1 of this dissertation presents a literature review summarizing broader patterns of evolutionary rate variation in organelle genomes with a particular emphasis on the role of mutational processes. Chapter 2 presents an analysis of sequence evolution at synonymous sites in sequenced seed plant mitochondrial genomes. The primary objective of this chapter is to assess the extent to which synonymous sites are subject to selection in plant mitochondrial genomes. This question is relevant to the entire dissertation because the assumption of relative neutrality at synonymous sites is used throughout to infer mutational patterns from rates of synonymous substitution. This chapter identifies some evidence of selection on synonymous sites associated with biased codon usage and

conservation of recognition sequences around RNA editing sites, but the estimated magnitude of the effects are weak in all cases, supporting the use of synonymous positions as a relatively neutral set of sites within the genome.

Chapters 3 and 4 characterize patterns of mitochondrial substitution rate variation at multiple biological scales within the genus *Silene*. They identify substantial mitochondrial rate variation among genes, among species, and among lineages within species. Analysis of these patterns of rate variation suggests that they reflect dramatic variation in the underlying mutation rates. In contrast, comparable rate changes are not identified in a sample of nuclear and plastid genes.

Chapter 5 reports the first complete *Silene* mitochondrial genome sequence. This sequence (from the slowly evolving species *S. latifolia*) demonstrates many of the genomic complexities that typify angiosperm mitochondria as well as some unique patterns that distinguish *Silene* from other lineages. In particular, the *Silene* lineage has experienced rampant gene loss and sequence divergence in the mitochondrially-encoded components of its translation machinery.

Chapters 6-8 present comparative analyses of whole organelle genome data from multiple *Silene* species with highly divergent mitochondrial mutation rates. Chapter 6 shows that mutational acceleration in *Silene* mitochondrial genomes has been associated with a rapid reduction in the frequency of C-to-U RNA editing. This finding is consistent with the predictions of the mutational burden hypothesis. However, a detailed analysis of the pattern of editing site losses suggests that an alternative mechanism involving a C-to-T mutation bias at RNA editing sites resulting from retroprocessing may be a more likely explanation.

Chapter 7 shows that, contrary to the predictions of the mutational burden hypothesis, two fast-evolving *Silene* mitochondrial genomes (from *S. noctiflora* and *S. conica*) have experienced dramatic proliferations of non-coding content, making them the largest mitochondrial genomes ever identified. In contrast, two slowly evolving *Silene* mitochondrial genomes (from *S. latifolia* and *S. conica*) have maintained typical sizes for angiosperms. The fast-evolving *Silene* mitochondrial genomes also exhibit novel multi-circular structures and dramatically reduced frequencies of alternative genome conformations associated with intragenomic recombination.

Finally, chapter 8 presents a comparative analysis of complete plastid genomes from the same four species. Although there is no evidence for genome-wide mutation rate increases comparable to those observed in the mitochondria, the two species with fast evolving mitochondrial genomes also show elevated rates of evolution in the plastids, including higher substitution rates and multiple inversions, intron losses, large indels, and changes in the inverted repeat boundaries. In contrast to the mitochondrial genomes, elevated plastid substitution rates are restricted to a subset of genes and appear to be driven by altered selection pressures rather than increased mutation rates.

The results of this research show that *Silene* mitochondrial genomes have experienced rapid and extreme changes in genome architecture, providing a new model for investigating the forces that shape genome evolution. The observed genomic changes are associated with dramatic increases in the mitochondrial mutation rate. However, the specific patterns of change generally fail to support the predictions of the mutational burden hypothesis. In particular, the proliferation of intergenic regions and the overall expansion of mitochondrial genome size in *S. noctiflora* and *S. conica* runs directly

counter to *a priori* expectations. While these genomic analyses are based on an admittedly meager sample size, it is important to note that these four points span approximately 98% of the known range of organelle genome size across all eukaryotes. The inability of the mutational burden hypothesis to explain such a major component of variation in organelle genome size undermines its generality and raises questions about other evolutionary forces that may be at play, including directional mutational biases (Mira et al. 2001; Petrov 2002; Kuo and Ochman 2009; Sloan et al. 2010). These results also highlight the potentially complex interplay between mutational processes and genome evolution. The predictions from the mutational burden hypothesis treat mutation rates essentially as an independent variable that affects genome evolution. Reality is more complex, however, than any simple unidirectional effect. Mutation rate, selection, recombination and $N_e$ are all highly interrelated within a tangled web of cause-and-effect (Nordborg et al. 1996; Lercher and Hurst 2002; Lynch 2010; Sloan and Panjeti 2010). The results of this research indicate the further unraveling the evolutionary history of *Silene* organelle genomes will help to dissect some of these fundamental evolutionary forces.

## References

Boore J. L. 1999. Animal mitochondrial genomes. Nucleic Acids Res. 27:1767-1780.

Cavalier-Smith T. 1985. Eukaryote gene numbers, non-coding DNA and genome size. Pp. 69-103 *in* T. Cavalier-Smith, ed. The evolution of genome size. Wiley, Chichester.

Cavalier-Smith T. 1982. Skeletal DNA and the evolution of genome size. Annu. Rev. Biophys. Bioeng. 11:273-302.

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

Dobzhansky T., and A. H. Sturtevant. 1938. Inversions in the chromosomes of *Drosophila pseudoobscura*. Genetics 23:28-64.

Doolittle W. F., and C. Sapienza. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature 284:601-603.

Farre J. C., G. Leon, X. Jordana, and A. Araya. 2001. cis Recognition elements in plant mitochondrion RNA editing. Mol. Cell. Biol. 21:6731-6737.

Gilbert W. 1978. Why genes in pieces? Nature 271:501.

Gray M. W., G. Burger, and B. F. Lang. 1999. Mitochondrial evolution. Science 283:1476-1481.

Gray M. W., J. Lukeš, J. M. Archibald, P. J. Keeling, and W. F. Doolittle. 2010. Irremediable Complexity? Science 330:920-921.

Gregory T. R. 2005. The evolution of the genome. Elsevier, Amsterdam.

Hahn M. W., J. E. Stajich, and G. A. Wray. 2003. The effects of selection against spurious transcription factor binding sites. Mol. Biol. Evol. 20:901-906.

Hershey A. D., and M. Chase. 1952. Independent functions of viral protein and nucleic acid in growth of bacteriophage. J. Gen. Physiol. 36:39-56.

Knoop V. 2011. When you can't trust the DNA: RNA editing changes transcript sequences. Cell. Mol. Life Sci. 68:567-586.

Kuo C. H., and H. Ochman. 2009. Deletional bias across the three domains of life. Genome Biol. Evol. 1:145-152.

Lercher M. J., and L. D. Hurst. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. Trends in Genetics 18:337-340.

Lewitsky G. A. 1931. The karyotype in systematics. Bull. Appl. Bot. 27:220-240.

Lynch M. 2010. Evolution of the mutation rate. Trends Genet. 26:345-352.

--- 2007. The Origins of Genome Architecture. Sinauer Associates, Sunderland, MA.

--- 2006. Streamlining and simplification of microbial genome architecture. Annu. Rev. Microbiol. 60:327-349.

--- 2006. The origins of eukaryotic gene structure. Mol. Biol. Evol. 23:450-468.

--- 2002. Intron evolution as a population-genetic process. Proc. Natl. Acad. Sci. 99:6118.

Lynch M., and J. S. Conery. 2003. The origins of genome complexity. Science 302:1401-1404.

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

Lynch M., D. G. Scofield, and X. Hong. 2005. The evolution of transcription-initiation sites. Mol. Biol. Evol. 22:1137-1146.

McCutcheon J. P., B. R. McDonald, and N. A. Moran. 2009. Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. PLoS Genet. 5:e1000565.

Mira A., H. Ochman, and N. A. Moran. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17:589-596.

Mirsky A. E., and H. Ris. 1951. The desoxyribonucleic acid content of animal cells and its evolutionary significance. J. Gen. Physiol. 34:451-462.

Mower J. P., D. B. Sloan, and A. J. Alverson. In press. Plant mitochondrial diversity – the genomics revolution. *in* J. F. Wendel, ed. Plant Genome Diversity. Springer.

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

Mulligan R. M., K. L. C. Chang, and C. C. Chou. 2007. Computational analysis of RNA editing sites in plant mitochondrial genomes reveals similar information content and a sporadic distribution of editing sites. Mol. Biol. Evol. 24:1971-1981.

Nordborg M., B. Charlesworth, and D. Charlesworth. 1996. The effect of recombination on background selection. Genet. Res. 67:159-174.

Orgel L. E., and F. H. Crick. 1980. Selfish DNA: the ultimate parasite. Nature 284:604-607.

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

Pellicer J., M. F. Fay, and I. J. Leitch. 2010. The largest eukaryotic genome of them all? Bot. J. Linn. Soc. 164:10-15.

Petrov D. A. 2002. Mutational equilibrium model of genome size evolution. Theor. Popul. Biol. 61:531-544.

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

Sloan D. B., A. H. MacQueen, A. J. Alverson, J. D. Palmer, and D. R. Taylor. 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., C. M. Barr, M. S. Olson, S. R. Keller, and D. R. Taylor. 2008. Evolutionary rate variation at multiple levels of biological organization in plant mitochondrial DNA. Mol. Biol. Evol. 25:243-246.

Sloan D. B., and V. G. Panjeti. 2010. Evolutionary feedbacks between reproductive mode and mutation rate exacerbate the paradox of sex. Evolution 64:1129-1135.

Swift H. 1950a. The constancy of desoxyribose nucleic acid in plant nuclei. Proc. Natl. Acad. Sci. 36:643-654.

Swift H. H. 1950b. The desoxyribose nucleic acid content of animal nuclei. Physiol. Zool. 23:169-198.

Wagner A. 2005. Robustness and evolvability in living systems. Princeton University Press, Princeton, NJ.

Watson J. D., and F. H. C. Crick. 1953. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature 171:737-738.

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

# Chapter 1.

# Evolutionary rate variation in organelle genomes: the role of mutational processes[1]

[1]Formatted as a co-authored manuscript (Sloan DB and Taylor DR) in response to an invitation to contribute to the following volume:

*Organelle Genetics*. In Press. Bullerwell CE, ed. Springer-Verlag

**Abstract**

With the ever-expanding availability of DNA sequence data, it has become increasingly clear that genes and genomes can evolve at very different rates. Organelle genomes in particular, provide dramatic examples of variation in nucleotide substitution rates across species, which in many cases reflect variation in underlying mutational processes. In this chapter, we review the evidence for variation in organelle mutation rates and discuss the evolutionary causes and consequences of this variation. We suggest that the existence of mutation rate variation across diverse phylogenetic scales makes organelle genomes an ideal system for investigating the interplay between mutational processes and the evolution of genome architecture.

## 1. Introduction

The field of molecular evolution makes extensive use of the comparative analysis of evolutionary rates. Rates of nucleotide substitution are often used as tools to study the history of selection, with corrections being made to account for underlying differences in the mutation rate. But there is also broad interest in the mutation rate, *per se*, as an evolutionary force. For example, sex and recombination may be favored to enhance the clearance of deleterious mutations from the genome (Muller. 1964), complex genomes may be streamlined (Lynch. 2006, 2007) or genes may be moved to other compartments of the cell (i.e. the nucleus) (Brandvain and Wade. 2009) to reduce the occurrence of deleterious mutations, and mutational biases may influence many aspects of genome size and structure (Mira *et al*. 2001; Petrov. 2002). Organelle genomes are of particular

interest in this context because they exhibit some of the greatest natural variation in mutation rate.

Mitochondria and plastids represent the oldest examples of a much larger group of endosymbiotic relationships in which formerly free-living organisms have evolved to live exclusively inside the cells of other organisms. A common characteristic uniting this diverse group of organelles and endosymbionts is an accelerated rate of DNA sequence evolution (Brown *et al*. 1979; Andersson and Kurland. 1998; Moran *et al*. 2009). This acceleration can be partially explained by changes in the strength and efficacy of natural selection (and hence the probability of fixation of mutations) resulting from an intracellular lifestyle (Moran. 1996; Lynch and Blanchard. 1998). However, increases in the mutation rate (the probability of occurrence of mutations) are also responsible, as these genomes have lost large numbers of genes, including many involved in DNA replication and repair.

Mitochondrial and plastid genomes provide the most extreme examples of gene loss. In fact, most organelle genomes completely lack DNA replication and repair genes, as the control of organelle genome maintenance has been transferred to the nucleus (Timmis. 2004). It is clear that eukaryotic lineages differ significantly in their nuclear-encoded replication and repair machinery, which may provide an explanation for dramatic variation in organelle mutation rates. In addition, differences in physiology and life history may produce further variation in mutational input across lineages. In this chapter, we review the causes and consequences of substitution rate variation in organelle genomes, focusing predominantly on mutational processes. Given the ample evidence for

mitochondrial and plastid mutation rate variation, these systems represent a valuable model for investigating the role of mutational processes in shaping genome evolution.

## 2. Methods for Estimating Organelle Mutation Rates

To understand patterns of mutation rate variation in organelle genomes, we must first consider how mutation rates are estimated. The vast majority of rate estimates have been inferred from phylogenetic studies that quantify the extent of DNA sequence divergence among organisms. However, with the recent advent of high-throughput DNA sequencing technology, a complementary approach has also become feasible. Whole genome resequencing of mutation accumulation (MA) lines allows for genome-wide identification of genetic changes that have accumulated over a defined number of generations in the lab. In this section, we review these methods, highlighting their relative strengths and weaknesses.

### 2.1 Phylogenetic Methods

Molecular phylogenetic analyses are often used to generate an estimate of the number of substitutions that have occurred on each branch in an evolutionary tree. When combined with estimates of the age of that branch, this information can be used to calculate absolute substitution rates. Furthermore, restricting such analyses to neutrally evolving sequences can provide an estimate of the underlying mutation rate (see below). The advantage of these methods is that they are generally easy to implement and can often be conducted with publicly available data, making them amenable to comparisons of large numbers of

species. However, as discussed below, phylogenetic methods also suffer from a number of assumptions and technical challenges that potentially bias their mutation rate estimates (see also Lanfear et al. (2010) for a recent review of the use of phylogenetic methods to estimate evolutionary rates).

*2.1.1 Synonymous Substitution Rates and the Assumption of Neutrality*. One of the pillars of molecular evolution theory is that, if a sequence is not under selection, the nucleotide substitution rate is equal to its mutation rate (Kimura. 1983). Phylogenetic methods rely on this expectation to infer mutation rates from DNA sequence data. Synonymous substitutions are changes in DNA sequence that do not affect the corresponding amino acid sequence because of the redundancy of the genetic code. These so-called silent sites are therefore expected to be relatively free from selection pressures. Thus, the rate of synonymous substitutions in protein genes is commonly used as an estimate of the mutation rate (Table 1).

The assumption of complete neutrality at synonymous sites, however, is unrealistic. There is ample evidence for selective forces influencing rates of synonymous substitution. These include selection for biased codon usage, conservation of regulatory motifs, and stability of RNA secondary structure (Chamary *et al*. 2006). Therefore, there is likely to be some degree of purifying selection acting on synonymous sites in organelle genomes, resulting in a downward bias on mutation rate estimates based on synonymous substitution rates. Data from MA experiments and pedigree analyses suggest that this bias may be substantial—perhaps as much as one or two orders of magnitude (Denver *et al*. 2000; Howell *et al*. 2003; Haag-Liautard *et al*. 2008; Lynch *et al*. 2008).

An additional concern with the use of synonymous sites to estimate mutation rates is that they offer a potentially non-representative sample of the genome. For example, 4-fold synonymous sites in the *Drosophila* mitochondrial genome have an extremely skewed nucleotide composition (94% A+T), which reflects a strong mutational bias towards A:T base pairs (Haag-Liautard *et al*. 2008). Non-synonymous sites are also AT rich but not to the same extent (66% A+T), presumably because of functional constraint on amino acid sequence. MA experiments have shown that the rate of mutation in *Drosophila* mitochondrial DNA (mtDNA) is higher at non-synonymous than synonymous sites (Haag-Liautard *et al*. 2008). This result is not surprising given the differences in G+C content between synonymous and non-synonymous sites and the fact that most mutations in *Drosophila* mtDNA convert G/C to A/T. Therefore, in the case of *Drosophila*, the synonymous mutation rate is not representative of the genome-wide mutation rate. Similarly, Morton (Morton. 2003) found that, because of the constraints of the genetic code, synonymous sites in the plastid genomes of grasses are preferentially found adjacent to particular upstream and downstream nucleotides. Because the patterns of mutation at a given site depend on these flanking nucleotides, synonymous sites in plastid DNA can also be subject to different mutational pressures than the rest of the genome.

*2.1.2 Saturation*. One key to phylogenetic rate estimates is to accurately identify the number of nucleotide substitutions that have occurred along a particular branch. This can become challenging when there have been multiple substitutions at the same site. Models of sequence evolution are generally used to estimate the total number of substitutions that

occurred in a lineage when only a fraction of those changes are directly observable by sequence comparison (Sullivan and Joyce. 2005). In the extreme, however, when all sites have experienced one or more substitutions, such models cannot be effectively applied.

To circumvent this problem of "saturation", Nabholz et al. (2008, 2009) have employed a hierarchical strategy to analyze large datasets of mtDNA sequences in mammals and birds. Their method subdivides a dataset into smaller taxonomic groups, in which saturation is not a problem. The age of each group can be inferred based on more slowly evolving amino acid sequence data. Standard phylogenetic analyses are then performed separately on each subdivision. This approach represents a promising option that could be extended to other large datasets with levels of divergence that are too high for more standard phylogenetic analyses.

*2.1.3 Phylogenetic Artifacts.* Phylogenetic methods are subject to a number of biases that can potentially affect estimates of substitution rate. For example, although models of evolution attempt to account for the number of unobservable substitutions resulting from recurrent changes at the same site, there is often a bias towards underestimating the total number of substitutions in long branches. This is less of a problem in well-sampled clades with shorter internal branches. The discrepancy in rate estimates associated with sampling intensity is known as the node density effect (Venditti *et al*. 2006). Employing more phylogenetically balanced sampling strategies may minimize this effect. Alternatively, for imbalanced datasets, the magnitude of the node density effect can be assessed and possibly mitigated with statistical measures (Venditti *et al*. 2006; Venditti *et al*. 2008).

A second source of bias in phylogenetic analysis is intraspecific polymorphism. When only one or a small number of individuals are sampled from each species, estimates of between-species divergence can be inflated by the existence of within-species polymorphism (Peterson and Masel. 2009; Charlesworth. 2010). This bias is particularly important for divergence at young nodes in a tree and when relatively ancient polymorphisms are maintained across species boundaries. It leads to higher branch length estimates at the tips of a tree as compared to internal branches near the root. However, this effect does not explain asymmetry in synonymous branch lengths between sister taxa, which is the classic signature of substitution rate variation across lineages (Figure 1). One solution to this problem is to sample numerous individual per species to quantify the effect of polymorphism. For large-scale phylogenetic analyses, however, this solution may be impractical. A reasonable alternative may be to focus on deeper splits in the phylogenetic tree or on the relative rate differences between sister lineages.

A third issue with phylogenetic analyses is the basic assumption of a single tree-like model of evolution, which can be violated when organelle genomes undergo some forms of recombination. Organelle genomes are generally thought of as being uniparentally (usually maternally) inherited and, therefore, asexual. There, is, however, enormous variation in the modes of organelle inheritance across eukaryotes with countless exceptions to the rule of uniparental inheritance (Barr *et al*. 2005). When it occurs, biparental inheritance generates the opportunity for sexual recombination in organelle genomes, which can result in different sections of the genome having different genealogies. When sequence data from recombining genomes are analyzed under the assumption of a single genealogy, distortions in branch lengths and substitution rate

estimates can result. Cases of lateral gene transfer can have especially severe effects. For example, the "promiscuous" nature of plant mitochondrial genomes has resulted in transfer of genes or gene fragments among very distant related lineages (Ellis. 1982; Bergthorsson *et al*. 2003). In perhaps the most extreme examples, there is evidence of recent gene conversion between *anciently* homologous genes in the mitochondrial and plastid genomes of angiosperms (Hao and Palmer. 2009; Sloan *et al*. 2010). These cases of gene conversion affect relatively small stretches of DNA sequence and can superficially appear to be the result of local increases in substitution rate. Therefore, it is important to test phylogenetic datasets for evidence of recombination.

**2.2 Mutation Accumulation Lines**

Whereas phylogenetic estimates of mutation rates focus on portions of the genome that are believed to be relatively free from selection, MA lines represent an alternative approach in which experimental manipulation is used to reduce or eliminate the effect of selection across the entire genome. The logic behind MA experiments is that the efficacy of natural selection is proportional to the effective population size ($N_e$). Therefore, by repeated bottlenecking of a population through one or a small number of individuals, $N_e$ can be reduced to a point that only the most severe mutations will be removed by selection. Therefore, the genetic changes that accumulate over time in these experimental lines should very closely reflect the unfiltered mutational input. The ever-decreasing cost of DNA sequencing has made it feasible to re-sequence the entire genomes of individuals from MA lines. This approach has now been used to analyze mitochondrial mutation

patterns in a handful of classic laboratory model systems, including yeast, *Drosophila*, and *Caenorhabditis*.

*2.2.1 Findings from Mutation Accumulation Experiments.* The number of MA studies on organelle genomes is still very limited, but there is a clear trend suggesting that these more direct measures of mutation rates produce much higher estimates than those inferred from synonymous substitution rates. As opposed to phylogenetic analyses, which produce absolute (i.e., per year) rate estimates, the results of MA studies are measured on a per generation basis (Table 2). In the first major sequencing experiment based on MA lines, Denver et al. (Denver *et al*. 2000) reported a surprisingly high rate of point mutations in the mitochondrial genome of the nematode *Caenorhabditis elegans*. Assuming a 4-day generation time in *C. elegans*, this rate was roughly 2 orders of magnitude higher than previous estimates of mitochondrial mutation rates based on sequence divergence. A subsequent study of the related nematode *C. briggsae* found similarly high point mutation rates (although interestingly the mitochondrial genomes of these 2 closely-related species differed substantially in their rates of large deletion mutations)(Howe *et al*. 2010). MA experiments have also yielded unexpectedly high mitochondrial mutation rate estimates in yeast and *Drosophila* (Haag-Liautard *et al*. 2008; Lynch *et al*. 2008).

It appears that the disagreement between mutation rate estimates based on phylogenetic methods and MA experiments is not limited to mitochondrial genomes. Nuclear mutation rate estimates from MA experiments have also been elevated (e.g., Denver *et al*. 2004). However, at least in the cases of *Drosophila* and yeast, the

discrepancy between the two methods is more pronounced in the mitochondrial genome, resulting in higher estimates of the ratio of mitochondrial to nuclear mutation rates (Haag-Liautard *et al*. 2008; Lynch *et al*. 2008).

Although the discrepancy between mutation rate estimates derived from synonymous substitution rates and MA lines may reflect some of the shortcomings of phylogenetic analyses (Sect. 2.1), it is also important to consider that the substitution process for mitochondrial mutations is complex, reflecting the hierarchical organization of eukaryotic organisms. Below, we discuss these complexities in the context of MA experiments.

*2.2.2 Heteroplasmy and Mutation Accumulation Experiments.* Unlike the nuclear genome, which typically occurs as a single diploid copy in each cell, organelle genomes are highly polyploid often with many thousands of genome copies distributed across multiple organelles in each cell (Moraes. 2001; Day and Madesis. 2007). Therefore, to reach fixation, a novel mutation in an organelle genome must not only spread among individuals within a population but also among the many genome copies within a cell. The state of coexistence between different copies of an organelle genome within the same cell or individual is known as heteroplasmy.

The hierarchical organization of biological populations creates multiple levels at which selection can act. The bottlenecking approach employed in MA experiments is designed to reduce the efficacy of selection at the individual level. However, this approach does not necessarily reduce the effects of mechanisms such as replication advantage and mitophagy, which may act at lower levels of selection to bias the fate of

new organelle mutations (Taylor *et al*. 2002; Tolkovsky. 2009). For example, it has been found in experimental yeast populations that reducing the population size results in the dominance of within-cell selection pressures favoring the spread of mtDNA deletions that increase the rate of genome replication but eliminate the ability of the cell to respire (Taylor *et al*. 2002). In addition, mouse lines engineered to have higher mitochondrial mutation rates resulting from a proof-reading deficient mtDNA polymerase show evidence of substantial selection against non-synonymous changes in the mitochondrial genome within only a few generations (Stewart *et al*. 2008). The presumably small $N_e$ and rapid response to selection in these lines suggest that there may be mechanisms of selection occurring below the individual level.

Interestingly, the fraction of mutations found in the heteroplasmic state has differed substantially among MA experiments with different organisms (Denver *et al*. 2000; Haag-Liautard *et al*. 2008; Howe *et al*. 2010). This may reflect differences in the severity of the "mitochondrial bottleneck" among species, i.e., the number of mitochondria (or mitochondrial genome copies) transmitted to the offspring. To what extent do the dynamics of mitochondria within cells affect the accumulation and selective filtering of mutations? How does the mitochondrial bottleneck within individuals act to reduce selection among these variants? A better understanding of the replication and transmission dynamics of organelle genomes in a heteroplasmic state will be an essential step toward accurately interpreting both phylogenetic estimates of evolutionary rates and the results obtained from MA experiments.

**3. Phylogenetic Variation in Organelle Substitution Rates**

The branching structure of phylogenetic trees is inherently fractal, and comparing rates of sequence evolution across different evolutionary timescales suggests that rate variation has a fractal nature as well. From some of the deepest splits in the eukaryotic phylogeny all the way down to the intraspecific level, there is evidence for evolutionary rate variation in organelle genomes, much of which can be explained by differences in the underlying mutation rate. As discussed above (Sect. 2), most evidence for organelle mutation rate variation has been inferred from phylogenetic patterns of sequence divergence. In this section, we summarize patterns of substitution rate variation that have been indentified across different phylogenetic scales.

### 3.1 Early Evidence for Mitochondrial Substitution Rate Differences between Plants and Animals

In a classic study, Brown et al. (1979) showed that mtDNA from 4 primate species evolves approximately an order of magnitude faster than single-copy nuclear genes. As predicted by Brown et al., the rapid rate of mtDNA sequence evolution in most animals has made mtDNA a preferred tool for phylogenetic and population genetic studies. However, an elevated rate of sequence evolution in the mitochondrial genome (relative to the nucleus) is not a universal rule in eukaryotes. The opposite pattern generally occurs in plants. Wolfe et al. (1987) found that substitution rates in angiosperm mtDNA are approximately an order of magnitude slower than corresponding rates in the nucleus, while substitution rates in the plastid genome fall in between these two levels. By comparing rates of nucleotide substitution in absolute terms, Wolfe et al. (1987) established that, while rates of nuclear sequence evolution are comparable between plants

and animals, rates of mtDNA evolution in the two lineages have evolved to opposite extremes, differing by 100-fold or more.

### 3.2 Limited Data from Other Eukaryotic Lineages

In contrast to the wealth of data available for plants and animals, estimates of organelle substitution rates are sorely lacking in other eukaryotic lineages including fungi and protists. Nevertheless, there is some evidence to suggest that most eukaryotic lineages have experienced rates of mitochondrial evolution that fall in between the extremes observed in multicellular plants and animals. For example, global phylogenies of slowly-evolving rRNA genes exhibit intermediate branch lengths for protists and fungi (Yang *et al*. 1985; Gray *et al*. 1989). In addition, studies on a handful of protist and fungal lineages have found ratios of mitochondrial to nuclear divergence closer to 1:1, contrasting with biased ratios observed in plants and animals (Clark-Walker. 1991; Lynch and Blanchard. 1998; Lynch *et al*. 2006). However, there are at least some lineages, such as the mushroom order Boletales, that exhibit elevated ratios of mitochondrial to nuclear divergence (Bruns and Szaro. 1992). Furthermore, data from MA lines in yeast show a very high ratio of mitochondrial to nuclear mutation rates (~37:1) (Lynch *et al*. 2008), which is in conflict with estimates derived from synonymous substitutions (Clark-Walker. 1991; Lynch and Blanchard. 1998; Lynch *et al*. 2006).

Given the ever-increasing availability of DNA sequence data, systematic analyses of mitochondrial rate variation—similar to those recently conducted in diverse groups such as seed plants (Mower *et al*. 2007), mammals (Nabholz *et al*. 2008), and birds (Nabholz *et al*. 2009)—would be a valuable contribution to the field. Even in cases where

reliable divergence times cannot be estimated to calculate absolute substitution rates, comparisons of the relative rate of nuclear and mitochondrial substitution would be informative.

### 3.3 Mitochondrial Substitution Rate Variation within Major Taxonomic Groups

While the long-standing generalization that animal mtDNA evolves rapidly and plant mtDNA evolves slowly has remained largely intact, more recent research has also shown that there is substantial rate variation within each of these groups. Notably, it is not clear that the high rate observed in most animal mitochondrial genomes is the ancestral state for all animals, because many non-bilaterians (including corals and sponges) have markedly slower rates (Shearer *et al*. 2002; Huang *et al*. 2008). Instead, it has been proposed that there was a mitochondrial rate acceleration in the ancestor of all bilaterians (Hellberg. 2006; Huang *et al*. 2008).

There is also evidence for significant rate variation at lower taxonomic levels, even in vertebrate mtDNA, which for years was viewed as one the strongest cases for a molecular clock as evidenced by the famous "2% per million years" rule of thumb. For example, the rates of evolution in turtle and shark mitochondrial genomes are slow relative to other vertebrates (Avise *et al*. 1992; Martin *et al*. 1992), and recent in-depth studies of mitochondrial sequence divergence within both mammals and birds have revealed surprising levels of rate variation (Nabholz *et al*. 2008, 2009). Mammalian species in particular differ by 2 orders of magnitude in synonymous substitution rate, shattering the misconception of constant mutation rates even on relatively local phylogenetic scales (Galtier *et al*. 2009).

Research over the last decade on typically slow-evolving plant mtDNA has uncovered some of the most extreme examples of substitution rate variation ever identified. Flowering plants from multiple independent lineages (including the genera *Pelargonium*, *Plantago*, and *Silene*) exhibit massive accelerations in mitochondrial synonymous substitution rate, sometimes in excess of 1000-fold (Cho *et al*. 2004; Parkinson *et al*. 2005; Mower *et al*. 2007; Sloan *et al*. 2009). As a result, rates in these lineages often approach those of some of the fastest-evolving animal mitochondria. In contrast to the dramatic increases in mitochondrial substitution rates in these lineages, plastid and nuclear substitution rates appear generally unchanged, suggesting these species have experienced a mitochondrial-specific increase in mutation rate (Cho *et al*. 2004; Parkinson *et al*. 2005; Mower *et al*. 2007; Sloan *et al*. 2009; but see Erixon and Oxelman. 2008; Guisinger *et al*. 2008). In some cases, particularly within the genus *Silene*, these changes have occurred quite recently (<10 Mya), resulting in closely related species with highly divergent mitochondrial rates (Fig. 1)(Mower *et al*. 2007; Sloan *et al*. 2009). Further variation has been generated by apparent rate reversions in a subset of the accelerated lineages (Parkinson *et al*. 2005).

It is not surprising that some of the most extreme examples of variation in mitochondrial substitution rates have been documented in mammals, birds and seed plants (Table 1). This almost certainly reflects the greater intensity of study in these groups and highlights the need for improved sampling in other eukaryotic lineages. Given the evidence that phylogenetic patterns of rate variation may extend all the way to the intraspecific level (Sloan *et al*. 2008) and that organelle mutation rate can vary among genomic regions (Wolfe *et al*. 1987) and even among individual genes (Sloan *et al*.

2009), it appears that our ability to detect rate variation in organelle genomes may be limited only by how close we are willing to look.

The existence of rate variation across these diverse biological scales is of fundamental importance for analyses of sequence data. In particular, many population genetic tests based on sequence diversity make the assumption of a constant underlying mutation rate, at least at some phylogenetic scale. Frequent violations of these assumptions in organelle genomes highlight the importance of using a local measure of the neutral substitution rate to correct estimates of diversity (Barr *et al*. 2007; Nabholz *et al*. 2009).

## 4. DNA Replication and Repair in Organelle Genomes

The substantial substitution rate variation among organelle genomes and among taxa at every phylogenetic scale begs for both mechanistic and evolutionary explanations. With respect to mechanistic explanations of mutation rate variation, most attention has been focused on systems of DNA replication and repair.

For the most part, organelle genomes completely lack the genes necessary for their own DNA replication and repair. Although there are occasional exceptions (e.g., the plastid genomes of many non-green algae contain a *dnaB*-like gene; Day and Madesis. 2007), the genetic control of organelle genome replication resides entirely in the nucleus. Even in *Reclinomonas americana*, a protist with the most gene-rich mitochondrial genome identified to date, there are no mitochondrially-encoded genes known to be

involved in genome replication and repair (although, unlike most eukaryotes, *Reclinomonas* does maintain a mitochondrially-encoded copy of a eubacterial-like RNA polymerase (Lang *et al*. 1997)). Furthermore, in angiosperms in which plastid ribosomes have been artificially eliminated, plastid genome replication still occurs, indicating that the replication process is not strictly dependent on plastid-encoded proteins (Zubko and Day. 2002).

In 1974, Clayton et al. showed that mammalian cells were incapable of repairing pyrimidine dimers in their mitochondrial DNA, indicating that mitochondria lacked the nucleotide excision repair pathways found in the nucleus. In some ways, this and other studies may have been over interpreted to mean that mitochondria lack DNA repair mechanisms altogether—a notion that has been clearly refuted with subsequent research identifying a host of different mechanisms involved in preventing and repairing mutations in mtDNA (Bogenhagen. 1999; Holt. 2009). Nevertheless, these early studies were important in establishing that the machineries involved in nuclear and organelle genome maintenance are not always the same. Instead, DNA replication and repair in organelle genomes depends on numerous genes that are specifically targeted to the mitochondria and/or plastids. Understanding the functional roles of these genes and how they vary across eukaryotic lineages is essential to understanding variation in organelle mutation rates.

### 4.1 Origins of Organelle DNA Replication and Repair Genes

*4.1.1 Endosymbiotic Gene Transfer.* The dominant pattern in the evolution of organelles genomes since their endosymbiotic origin is one of gene loss. The genomes of free-living

bacteria contain thousands of protein genes. In contrast, mitochondrial and plastid genomes contain fewer than 250, in most cases far fewer. Animal mitochondrial genomes contain a nearly universal complement of only 13 protein genes, and the mtDNA of *Plasmodium* species encodes only 3 proteins (Gray *et al*. 1999). Some of the reduction in organelle genome coding content can be explained by outright gene loss, but the history of eukaryotic evolution has also been characterized by a massive transfer of genes from the organelles to the nucleus—a process that remains active in many lineages (Timmis. 2004).

Evidence of endosymbiotic gene transfer (EGT) can be found in the genes responsible for organelle DNA replication and repair. Genes of both proteobacterial and cyanobacterial origin (presumably reflecting the progenitors of mitochondria and plastids, respectively) have been identified as components of organelle DNA replication and repair machinery (Van Dyck *et al*. 1992; Eisen and Hanawalt. 1999; Karlberg *et al*. 2000; Kimura *et al*. 2002; Wall *et al*. 2004; Lin *et al*. 2007; Shedge *et al*. 2007). These include some of the classic players in bacterial DNA repair such as *mutS*, *mutL* and *recA* (Eisen and Hanawalt. 1999; Lin *et al*. 2007; Shedge *et al*. 2007). In addition, some key processes in organelle DNA replication are apparently mediated by eubacterial-like proteins, including single stranded DNA binding and (in some eukaryotes) DNA polymerization (Van Dyck *et al*. 1992; Ono *et al*. 2007; Moriyama *et al*. 2008).

Despite the role of EGT in the evolution of organelle DNA replication and repair, it is also clear that many genes involved in these processes did not originate with the bacterial progenitors of mitochondria and plastids (Karlberg *et al*. 2000; Suzuki and Miyagishima. 2010). Instead, many components of organelle DNA replication and repair

machinery appear to have been co-opted or acquired from other sources, as we discuss below.

*4.1.2 Viral Origins*. Sequencing of the yeast mitochondrial RNA polymerase resulted in the surprising observation that it is not homologous to other known eukaryotic RNA polymerases or to the eubacterial RNA polymerase, as might be expected given the $\alpha$-proteobacterial origins of mitochondria (Masters *et al*. 1987). This finding has since been extended to diverse eukaryotic lineages (Cermakian *et al*. 1996). Rather than being derived from a eubacterial ancestor, it appears that mitochondrial RNA polymerases are related to those encoded by T3/T7 bacteriophages, indicating that in some cases the genes controlling organelle genome function in eukaryotes may have been acquired from viruses.

Subsequent studies have found that key components controlling organelle DNA replication may also be of bacteriophage origin, raising the possibility that these genes were simultaneously acquired from a viral ancestor early on in eukaryotic evolution, perhaps in association with the bacterial endosymbiont that gave rise to mitochondria (Filee and Forterre. 2005; Shutt and Gray. 2006a). For example, there is evidence in diverse eukaryotic lineages for another T7 bacteriophage homolog (known as Twinkle in humans) that is responsible for helicase activity in mitochondrial genome replication (Spelbrink *et al*. 2001; Shutt and Gray. 2006b). In addition, phylogenetic analysis suggests that DNA polymerase $\gamma$, the enzyme responsible for mitochondrial genome replication in animals and fungi, also has a T3/T7 bacteriophage homolog (Filee *et al*.

2002). Therefore, it is apparent that viral genes have played an important role in the evolution of organelle genome replication and expression.

*4.1.3 Dual Targeting to Mitochondria and Plastids.* The co-existence of 2 or more genomes within the eukaryotic cell creates the opportunity to share components of cellular machinery across genomic compartments. Nowhere is this more apparent than in the growing list of nuclear-encoded proteins that are targeted to both mitochondria and plastids. The set of known plant proteins that are dual targeted to the mitochondria and plastids is significantly enriched for genes involved in DNA synthesis and processing (Carrie *et al*. 2009a). Dual targeted proteins include DNA polymerases, helicases, topoisomerases and a RecA homolog (Wall *et al*. 2004; Christensen *et al*. 2005; Shedge *et al*. 2007; Carrie *et al*. 2009b;). These examples clearly demonstrate the ability of the evolutionary process to co-opt existing genetic machinery for function in other organelles. The importance of this process is further illustrated by numerous genes shown to be involved in the repair of both nuclear and mitochondrial DNA, including DNA glycosylases, an apurinic/apyrimidinic endonuclease, and DNA ligase III (Larsen *et al*. 2005; Holt. 2009; and references therein).

Even in cases where gene products are targeted to only the mitochondria or the plastids, there are often closely related paralogs that perform similar functions in the other organelle. For example, the plant-specific OSB gene family has been shown to be involved in the generation and maintenance of alternative conformations of the mitochondrial genome. This family includes paralogs that are targeted to the mitochondria, to the plastids, or possibly to both organelles (Zaegel *et al*. 2006). A

history of gene duplication and replacement has been clearly demonstrated in other organelle processes including translation. In particular, multiple angiosperm lineages appear to have experienced replacement of mitochondrial-encoded ribosomal protein genes by duplicated copies of (nuclear-encoded) plastid homologs (Adams *et al*. 2002; Mower and Bonen. 2009; Kubo and Arimura. 2010). Collectively, these phenomena illustrate a history of co-opting and modifying existing genes to function in organelles, and they have played an especially important role in the evolution of organelle DNA replication and repair genes.

Although our understanding of organelle genome replication and repair remains limited in many respects, the available data illustrate that these processes depend on a complex and evolutionary labile assemblage of viral, bacterial and eukaryotic genes. As we discuss in the next section, the flexibility of these systems has led to significant divergence in replication and repair across eukaryotic lineages.

### 4.2 Variation in Mitochondrial Genome Replication and Repair Machinery across Eukaryotes

The best-characterized mitochondrial systems are probably those from mammals and yeast. Comparisons between the processes of mtDNA replication and repair in these two lineages have revealed important differences that likely reflect enormous variation across the diversity of eukaryotes. In some cases, differences in repair machinery may explain observed variation in mitochondrial substitution rates.

Yeast nuclear genomes contain a homolog of the bacterial DNA repair gene *mutS* (*MSH1*), which encodes a protein that functions in mitochondrial mismatch repair

(Reenan and Kolodner. 1992). In contrast, mammalian mitochondria lack a *mutS*-based mismatch repair system, which may at least partially explain the higher rates of point mutations in mammalian mtDNA (Foury *et al*. 2004). Interestingly, the Msh1 protein has been shown to preferentially recognize mismatches that would result in transitions (Chi and Kolodner. 1994). Therefore the lack of *mutS*-based mismatch repair in mammals is consistent with the extreme bias observed in transition:transversion ratios in animal mtDNA, which can be well in excess of 10:1 (Tamura and Nei. 1993). In contrast, there is no significant excess of transitions observed in yeast mtDNA (Vanderstraeten *et al*. 1998; Lynch *et al*. 2008;).

There are also components of mammalian mtDNA replication and repair that have not been identified in yeast. For example, a homolog of the helicase Twinkle has not been found in yeast, and it is unclear what gene is responsible for helicase activity during yeast mtDNA replication. Yeast also lack an identifiable mitochondrial DNA polymerase $\gamma$ accessory factor, while such a factor is known to function in animal mitochondrial DNA synthesis and in the related T3/T7 bacteriophage system (Shutt and Gray. 2006a).

These few examples likely represent the tip of the iceberg when it comes to variation among eukaryotes in organelle DNA replication and repair. Although our understanding of the organelle genetic machinery remains limited (note that even in the most well characterized organelle systems there is ongoing uncertainty and controversy about the basic mechanisms of replication; Day and Madesis. 2007; Holt. 2009), there is evidence for distinct origins of the mitochondrial DNA polymerase in different eukaryotic lineages (Shutt and Gray. 2006a; Ono *et al*. 2007; Moriyama *et al*. 2008;). Such differences suggest that even the most central components of organelle DNA

replication and repair machinery may fundamentally differ from one species to the next. Furthermore, mutation screens and directed mutagenesis have been effective at identifying/generating variants with altered organelle genome replication and repair machinery and corresponding changes in mutation rates (Foury *et al*. 2004; Trifunovic *et al*. 2004). A valuable next step would be to identify the genetic basis of organelle mutation rate variation in natural populations. Cases of recent and extreme increases in organelle mutation rates would be a good place start (Mower *et al*. 2007; Sloan *et al*. 2009).

## 5. Evolutionary Explanations for Organelle Mutation Rate Variation

Up to this point, we have largely focused on mechanisms of organelle DNA repair as a potential cause of mutation rate variation. The observed mutation rate, however, depends not only on the efficacy of DNA repair, but also on the total amount of mutational input (Baer *et al*. 2007). Accordingly, extensive comparative work has been performed (particularly in vertebrates) to develop and test hypothesis about the causes of mitochondrial mutation rate variation, focusing on the role of physiology and life history. Historically, most hypotheses have treated organelle mutation rate variation as a byproduct of other biological differences (e.g. generation time and metabolic rate). In recent years, additional emphasis is being placed on the effects of mutation rate variation and the more direct role of natural selection in shaping the mutation rate.

### 5.1 Generation Time and Metabolic Rate Hypothesis

Comparative work exploring the causes of variation in mitochondrial substitution rates has focused predominantly on vertebrates and particularly mammals for which there are ample data on life history and physiology. In a now famous study, Martin and Palumbi (1993) noted the existence of a strong negative relationship between body size and the rate of DNA sequence evolution in mammals. Although it is unlikely that there is any direct effect of body size, this trait is strongly correlated with both metabolic rate and generation time, which are at the center of leading hypotheses about the sources of mutation rate variation in mitochondrial genomes.

*5.1.1 Metabolic Rate Hypothesis*. The basic operation of metabolic pathways in mitochondria is associated with the production of mutagenic byproducts including reactive oxygen species (Wallace. 2005). Therefore, one natural prediction is that species with higher metabolic rates will experience higher rates of mutational damage to their mitochondrial genomes. This prediction is consistent with the negative correlation between evolutionary rate and body size in mammals, because smaller mammals tend to have higher mass-specific metabolic rates. Nevertheless, studies that have attempted to decouple effects of metabolic rate from confounded variables have found limited support for this hypothesis, particularly outside of mammals (Lanfear *et al*. 2007; Nabholz *et al*. 2009).

*5.1.2 Generation Time Hypothesis*. An alternative hypothesis is based on the expectation that many or most mutations are the result of DNA replication errors and that species with shorter generation times will undergo more rounds of germ line DNA replication per

year. This hypothesis is also consistent with the negative relationship between body size and substitution rate, because smaller species tend to have shorter generation times. It is not entirely clear how this hypothesis should extend to other eukaryotic lineages, particularly those that do not have a sequestered germ line. Nevertheless, there is support for a generation time effect in invertebrates and plants, suggesting that it may have some generality outside of mammals (Smith and Donoghue. 2008; Thomas *et al*. 2010).

### 5.2 Variation in the Efficacy and Intensity of Selection on Mutation Rates

Although the idea that mutation rates can be shaped by the forces of natural selection is not new (Sturtevant. 1937), recent arguments have placed renewed emphasis on how variation in the selective environment may explain difference in organelle mutation rates. In general, selection is expected to favor reductions in the mutation rate because the vast majority of non-neutral mutations are deleterious. However, the strength of that selection and the ability of populations to respond to it may vary across species. For example, comparative analyses in vertebrates have found that mitochondrial synonymous substitution rates are correlated with lifespan, even when controlling for related life history traits including generation time. It has been proposed that these results reflect more intense selection in long-lived organisms for reduced mitochondrial mutation rates (Nabholz *et al*. 2008). This hypothesis represents an extension of the mitochondrial theory of aging, which posits that a positive feedback between the rate of mitochondrial mutations and the decline of mitochondrial function is responsible for the physiological signs of aging (Kujoth *et al*. 2007).

The outcome of selection on mutation rate may also depend on variation in the efficacy of selection. Based on the observation that species with low $N_e$ tend to have higher (per generation) point mutation rates, Lynch (2010) has argued that small populations cannot effectively select against weakly deleterious alleles that increase the mutation rate. This hypothesis, however, is not specific to organelle genomes, and Lynch notes that it cannot explain some of the major patterns in mitochondrial mutation rate variation across eukaryotes (e.g., the combination of extremely low mitochondrial rates and small $N_e$ in land plants).

The population genetic theory of mutation rate evolution in organelle genomes remains largely unexplored. In general, the magnitude of selection acting on mutation rate modifiers is dependent on genetic linkage between these modifiers and mutations throughout the genome (as well as any direct fitness effects of the modifier) (Sniegowski *et al*. 2000). Because organelle genomes generally experience little or no sexual recombination, a mutator located within the genome would remain tightly linked with resulting mutations and, therefore, be subject to strong selection. However, most of the genetic control of organelle mutation rates likely resides with nuclear-encoded DNA replication and repair machinery. Therefore, linkage between modifiers and mutation load should be dependent on the frequency of outcrossed sexual reproduction. A valuable area for theoretical and empirical population genetic research would be to investigate the evolutionary forces acting on nuclear-encoded modifiers of organelle mutation rates including the effects of mating system, $N_e$, and the relative frequency of deleterious and beneficial mutations.

## 6. Mutational Processes and the Evolution of Organelle Genome Architecture

Although mutational mechanisms are often viewed as a directionless player in evolution, generating "random" variation on which natural selection can act, they can also have clear directional effects. Based on a combination of empirical and theoretical arguments, it has been proposed that variation in mutational processes can explain some of the striking diversity of genome architecture found across living organisms, including differences in genome size, structure, and organization.

### 6.1 Biased Mutation as a Directional Force

Mutational patterns are often highly skewed, preferentially affecting certain portions of a genome and exhibiting a bias for certain types of nucleotide substitutions as well as disparities in the number and size of insertions vs. deletions (i.e., the indel spectrum). In the absence of counterbalancing selection, these biases represent a directional force in genome evolution. Directional mutation pressures have been linked to variation in nucleotide composition and genome size. In particular, differences in nuclear genome size in animals have been attributed to corresponding differences in the indel spectrum (Petrov. 2002). Likewise it has been proposed that high gene densities in bacteria result from a deletion bias (Mira *et al*. 2001; Kuo and Ochman. 2009).

Organelle genomes (particularly mitochondrial genomes) exhibit dramatic variation in genome size and gene density (Gray *et al*. 1999), but the extent to which directional mutation pressures can explain these differences in unclear. There is recent

evidence that rates of indels in mtDNA can vary even between very closely related species (Howe *et al*. 2010), but overall very little is known about variation in the mitochondrial indel spectrum. Comparative analyses determining the number and size of mitochondrial indels in diverse eukaryotic lineages would be a valuable contribution.

### 6.2 Mutation Pressure as a Selective Force

The mutational burden hypothesis presents another possible role for mutational pressures in shaping genome architecture (Lynch. 2006, 2007). The idea is that complex genomic features experience a small selective cost associated with the probability that they will be disrupted by mutation. For example a mutation altering the splice donor site of an intron can prevent proper splicing resulting in deleterious or lethal consequences depending on the functional importance of the gene. Lynch has argued that this form of selection acts as a general deterrent to the expansion of non-coding content. However, the selection coefficient on any single feature is expected to be quite small (proportional to the per nucleotide mutation rate). Therefore, the effects of this mechanism are predicted to vary across lineages, depending on both the efficacy and intensity of selection, which should be proportional to $N_e$ and the mutation rate, respectively (Lynch. 2007).

Consequently, the mutational burden hypothesis has been put forth as an explanation for some of the most dramatic differences in genome architecture observed among living organisms, e.g., prokaryotes (large $N_e$) vs. eukaryotes (small $N_e$) or animal mitochondria (high mutation rate) vs. plant mitochondria (low mutation rate). These comparisons, however, span enormous phylogenetic scales, which confound countless biological differences and raise alternative interpretations for observed variation in

genome architecture. Given the growing evidence for organelle mutation rate variation among much more closely related species, we suggest that mitochondrial and chloroplast genomes represent an ideal model for dissecting the genomic consequences of mutation rate variation.

## 7. Conclusion

The organelle genomes of eukaryotes exhibit remarkable variation in nucleotide substitution rates. Despite the challenges in estimating spontaneous mutation rates, differences in evolutionary rates at sites under relatively weak selection points to substantial mutation rate variation in organelle genomes. In most cases, the underlying molecular mechanisms remain elusive, though select examples of organelle mutation rate variation may be attributed to documented differences in DNA replication and repair machinery. Mutation rate variation also reflects more ultimate evolutionary causes. Recent studies have placed renewed focus on differences across species in the efficacy and intensity of selection on mutation rate modifiers. Finally, mutation itself may be a powerful evolutionary force. It has been proposed that biased mutation may drive many aspects of genome structure and that selection exerted by deleterious mutations may favor reduced genome complexity. Since mutation rate variation arises repeatedly over small phylogenetic scales, organelle genomes represent potentially powerful systems for testing these hypotheses.

**Acknowledgements**

Our research on mutation rates and the evolution of organelle genomes has been supported by the NSF (DEB-0808452 and MCB-1022128).

**References**

Adams KL, Daley DO, Whelan J, Palmer JD (2002) Genes for two mitochondrial ribosomal proteins in flowering plants are derived from their chloroplast or cytosolic counterparts. Plant Cell 14: 931-943.

Andersson SG, Kurland CG (1998) Reductive evolution of resident genomes. Trends Microbiol 6: 263-268.

Avise JC, Bowen BW, Lamb T, Meylan AB, Bermingham E (1992) Mitochondrial DNA evolution at a turtle's pace: Evidence for low genetic variability and reduced microevolutionary rate in the testudines. Mol Biol Evol 9: 457-473.

Baer CF, Miyamoto MM, Denver DR (2007) Mutation rate variation in multicellular eukaryotes: Causes and consequences. Nat Rev Genet 8: 619-631.

Barr CM, Neiman M, Taylor DR (2005) Inheritance and recombination of mitochondrial genomes in plants, fungi and animals. New Phytol 168: 39-50.

Barr CM, Keller SR, Ingvarsson PK, Sloan DB, Taylor DR (2007) Variation in mutation rate and polymorphism among mitochondrial genes in Silene vulgaris . Mol Biol Evol 24: 1783-1791.

Bergthorsson U, Adams KL, Thomason B (2003) Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature 424: 197-201.

Bogenhagen DF (1999) Repair of mtDNA in vertebrates. Am J Hum Genet 64: 1276-1281.

Brandvain Y, Wade MJ (2009) The functional transfer of genes from the mitochondria to the nucleus: The effects of selection, mutation, population size and rate of self-fertilization. Genetics 182: 1129-1139.

Brown WM, George M, Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. Proc Natl Acad Sci 76: 1967-1971.

Bruns TD, Szaro TM (1992) Rate and mode differences between nuclear and mitochondrial small-subunit rRNA genes in mushrooms. Mol Biol Evol 9: 836-855.

Carrie C, Giraud E, Whelan J (2009) Protein transport in organelles: Dual targeting of proteins to mitochondria and chloroplasts. FEBS J 276: 1187-1195.

Carrie C, et al (2009) Approaches to defining dual-targeted proteins in Arabidopsis. Plant J 57: 1128-1139.

Cermakian N, Ikeda TM, Cedergren R, Gray MW (1996) Sequences homologous to yeast mitochondrial and bacteriophage T3 and T7 RNA polymerases are widespread throughout the eukaryotic lineage. Nucleic Acids Res 24: 648-654.

Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: Non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7: 98-108.

Charlesworth D (2010) Apparent recent elevation of mutation rate: Don't forget the ancestral polymorphisms. Heredity [Epub ahead of print].

Chi NW, Kolodner RD (1994) Purification and characterization of MSH1, a yeast mitochondrial protein that binds to DNA mismatches. J Biol Chem 269: 29984-29992.

Cho Y, Mower JP, Qiu YL, Palmer JD (2004) Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. Proc Natl Acad Sci 101: 17741-17746.

Christensen AC, et al (2005) Dual-domain, dual-targeting organellar protein presequences in Arabidopsis can use non-AUG start codons. Plant Cell 17: 2805-2816.

Clark-Walker GD (1991) Contrasting mutation rates in mitochondrial and nuclear genes of yeasts versus mammals. Curr Genet 20: 195-198.

Clayton DA, Doda JN, Friedberg EC (1974) The absence of a pyrimidine dimer repair mechanism in mammalian mitochondria. Proc Natl Acad Sci U S A 71: 2777-2781.

Day A, Madesis P (2007) in Cell and Molecular Biology of Plastids, ed Bock R (Springer, pp 65-119).

Denver DR, Morris K, Lynch M, Vassilieva LL, Thomas WK (2000) High direct estimate of the mutation rate in the mitochondrial genome of Caenorhabditis elegans. Science 289: 2342-2344.

Denver DR, Morris K, Lynch M, Thomas WK (2004) High mutation rate and predominance of insertions in the Caenorhabditis elegans nuclear genome. Nature 430: 679-682.

Eisen JA, Hanawalt PC (1999) A phylogenomic study of DNA repair genes, proteins, and processes. Mutat Res 435: 171-213.

Ellis J (1982) Promiscuous DNA--chloroplast genes inside plant mitochondria. Nature 299: 678-679.

Erixon P, Oxelman B (2008) Whole-gene positive selection, elevated synonymous substitution rates, duplication, and indel evolution of the chloroplast clpP1 gene. PLoS ONE 3: e1386.

Filee J, Forterre P (2005) Viral proteins functioning in organelles: A cryptic origin?. Trends Microbiol 13: 510-513.

Filee J, Forterre P, Sen-Lin T, Laurent J (2002) Evolution of DNA polymerase families: Evidences for multiple gene exchange between cellular and viral proteins. J Mol Evol 54: 763-773.

Foury F, Hu J, Vanderstraeten S (2004) Mitochondrial DNA mutators. Cell Mol Life Sci 61: 2799-2811.

Galtier N, Nabholz B, Glemin S, Hurst GD (2009) Mitochondrial DNA as a marker of molecular diversity: A reappraisal. Mol Ecol 18: 4541-4550.

Gray MW, Cedergren R, Abel Y, Sankoff D (1989) On the evolutionary origin of the plant mitochondrion and its genome. Proc Natl Acad Sci 86: 2267-2271.

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

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

Haag-Liautard C, et al (2008) Direct estimation of the mitochondrial DNA mutation rate in Drosophila melanogaster . PLoS Biol 6: 1706-1714.

Hao W, Palmer JD (2009) Fine-scale mergers of chloroplast and mitochondrial genes create functional, transcompartmentally chimeric mitochondrial genes. Proc Natl Acad Sci 106: 16728-16733.

Hellberg ME (2006) No variation and low synonymous substitution rates in coral mtDNA despite high nuclear variation. BMC Evol Biol 6: 24.

Holt IJ (2009) Mitochondrial DNA replication and repair: All a flap. Trends Biochem Sci 34: 358-365.

Howe DK, Baer CF, Denver DR (2010) High rate of large deletions in Caenorhabditis briggsae mitochondrial genome mutation processes. Genome Biol Evol 2: 29-38.

Howell N, et al (2003) The pedigree rate of sequence divergence in the human mitochondrial genome: There is a difference between phylogenetic and pedigree rates. Am J Hum Genet 72: 659-670.

Huang D, Meier R, Todd PA, Chou LM (2008) Slow mitochondrial COI sequence evolution at the base of the metazoan tree and its implications for DNA barcoding. J Mol Evol 66: 167-174.

Karlberg O, Canback B, Kurland CG, Andersson SG (2000) The dual origin of the yeast mitochondrial proteome. Yeast 17: 170-187.

Kimura M (1983) The Neutral Theory of Molecular Evolution, (Cambridge University Press, Cambridge).

Kimura S, et al (2002) A novel DNA polymerase homologous to Escherichia coli DNA polymerase I from a higher plant, rice (oryza sativa L.). Nucleic Acids Res 30: 1585-1592.

Kubo N, Arimura S (2010) Discovery of the rpl10 gene in diverse plant mitochondrial genomes and its probable replacement by the nuclear gene for chloroplast RPL10 in two lineages of angiosperms. DNA Res 17: 1-9.

Kujoth GC, Bradshaw PC, Haroon S, Prolla TA (2007) The role of mitochondrial DNA mutations in mammalian aging. PLoS Genet 3: e24.

Kuo CH, Ochman H (2009) Deletional bias across the three domains of life. Genome Biol Evol 1: 145-152.

Lanfear R, Thomas JA, Welch JJ, Brey T, Bromham L (2007) Metabolic rate does not calibrate the molecular clock. Proc Natl Acad Sci 104: 15388-15393.

Lanfear R, Welch JJ, Bromham L (2010) Watching the clock: Studying variation in rates of molecular evolution between species. Trends Ecol Evol 25: 495-503.

Lang BF, et al (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature 387: 493-497.

Lin Z, Nei M, Ma H (2007) The origins and early evolution of DNA mismatch repair genes--multiple horizontal gene transfers and co-evolution. Nucleic Acids Res 35: 7591-7603.

Larsen NB, Rasmussen M, Rasmussen LJ (2005) Nuclear and mitochondrial DNA repair: Similar pathways? Mitochondrion 5: 89-108.

Lynch M (2006) Streamlining and simplification of microbial genome architecture. Annu Rev Microbiol 60: 327-349.

Lynch M (2007) The Origins of Genome Architecture, (Sinauer Associates, Sunderland, MA).

Lynch M (2010) Evolution of the mutation rate. Trends Genet 26: 345-352.

Lynch M, Blanchard JL (1998) Deleterious mutation accumulation in organelle genomes. Genetica 103: 29-39.

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

Lynch M, et al (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci U S A 105: 9272-9277.

Martin AP, Palumbi SR (1993) Body size, metabolic rate, generation time, and the molecular clock. Proc Natl Acad Sci 90: 4087-4091.

Martin AP, Naylor GJP, Palumbi SR (1992) Rates of mitochondrial DNA evolution in sharks are slow compared with mammals. Nature 357: 153-155.

Masters BS, Stohl LL, Clayton DA (1987) Yeast mitochondrial RNA polymerase is homologous to those encoded by bacteriophages T3 and T7. Cell 51: 89-99.

Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. Trends Genet 17: 589-596.

Moraes CT (2001) What regulates mitochondrial DNA copy number in animal cells?. Trends Genet 17: 199-205.

Moran NA (1996) Accelerated evolution and muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci U S A 93: 2873-2878.

Moran NA, McLaughlin HJ, Sorek R (2009) The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. Science 323: 379-382.

Moriyama T, Terasawa K, Fujiwara M, Sato N (2008) Purification and characterization of organellar DNA polymerases in the red alga Cyanidioschyzon merolae. FEBS J 275: 2899-2918.

Morton BR (2003) The role of context-dependent mutations in generating compositional and codon usage bias in grass chloroplast DNA. J Mol Evol 56: 616-629.

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

Mower JP, Touzet P, Gummow JS, Delph LF, Palmer JD (2007) Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. BMC Evol Biol 7: 135.

Muller HJ (1964) The relation of recombination to mutational advance. Mutat Res 106: 2-9.

Nabholz B, Glemin S, Galtier N (2008) Strong variations of mitochondrial mutation rate across mammals--the longevity hypothesis. Mol Biol Evol 25: 120-130.

Nabholz B, Glémin S, Galtier N (2009) The erratic mitochondrial clock: Variations of mutation rate, not population size, affect mtDNA diversity across birds and mammals. BMC Evol Biol 9: 54.

Ono Y, et al (2007) NtPolI-like1 and NtPolI-like2, bacterial DNA polymerase I homologs isolated from BY-2 cultured tobacco cells, encode DNA polymerases engaged in DNA replication in both plastids and mitochondria. Plant Cell Physiol 48: 1679-1692.

Parkinson CL, et al (2005) Multiple major increases and decreases in mitochondrial substitution rates in the plant family Geraniaceae. BMC Evol Biol 5: 73.

Peterson GI, Masel J (2009) Quantitative prediction of molecular clock and Ka/Ks at short timescales. Mol Biol Evol 26: 2595-2603.

Petrov DA (2002) Mutational equilibrium model of genome size evolution. Theor Popul Biol 61: 531-544.

Reenan RA, Kolodner RD (1992) Characterization of insertion mutations in the saccharomyces cerevisiae MSH1 and MSH2 genes: Evidence for separate mitochondrial and nuclear functions. Genetics 132: 975-985.

Shearer TL, Van Oppen MJH, Romano SL, Wörheide G (2002) Slow mitochondrial DNA sequence evolution in the anthozoa (cnidaria). Mol Ecol 11: 2475-2487.

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

Shutt TE, Gray MW (2006) Bacteriophage origins of mitochondrial replication and transcription proteins. Trends Genet 22: 90-95.

Shutt TE, Gray MW (2006) Twinkle, the mitochondrial replicative DNA helicase, is widespread in the eukaryotic radiation and may also be the mitochondrial DNA primase in most eukaryotes. J Mol Evol 62: 588-599.

Sloan DB, Barr CM, Olson MS, Keller SR, Taylor DR (2008) Evolutionary rate variation at multiple levels of biological organization in plant mitochondrial DNA. Mol Biol Evol 25: 243-246.

Sloan DB, Oxelman B, Rautenberg A, Taylor DR (2009) Phylogenetic analysis of mitochondrial substitution rate variation in the angiosperm tribe Dileneae (Caryophyllaceae). BMC Evol Biol 9: 260.

Sloan DB, Alverson AJ, Storchova H, Palmer JD, Taylor DR (2010) Extensive loss of translational genes in the structurally dynamic mitochondrial genome of the angiosperm Silene latifolia. BMC Evol Biol in press:

Smith SA, Donoghue MJ (2008) Rates of molecular evolution are linked to life history in flowering plants. Science 322: 86-89.

Sniegowski PD, Gerrish PJ, Johnson T, Shaver A (2000) The evolution of mutation rates: Separating causes from consequences. Bioessays 22: 1057-1066.

Spelbrink JN, et al (2001) Human mitochondrial DNA deletions associated with mutations in the gene encoding twinkle, a phage T7 gene 4-like protein localized in mitochondria. Nat Genet 28: 223-231.

Stewart JB, et al (2008) Strong purifying selection in transmission of mammalian mitochondrial DNA. PLoS Biol 6: e10.

Sturtevant AH (1937) Essays on evolution. I. on the effects of selection on mutation rate. Q Rev Biol 12: 464.

Sullivan J, Joyce P (2005) Model selection in phylogenetics. Annu Rev Ecol Evol Syst 36: 445-466.

Suzuki K, Miyagishima SY (2010) Eukaryotic and eubacterial contributions to the establishment of plastid proteome estimated by large-scale phylogenetic analyses. Mol Biol Evol 27: 581-590.

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10: 512-526.

Taylor DR, Zeyl C, Cooke E (2002) Conflicting levels of selection in the accumulation of mitochondrial defects in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 99: 3690-3694.

Thomas JA, Welch JJ, Lanfear R, Bromham L (2010) A generation time effect on the rate of molecular evolution in invertebrates. Mol Biol Evol 27: 1173-1180.

Timmis J (2004) Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. Nat Rev Genet 5: 123-U16.

Tolkovsky AM (2009) Mitophagy. Biochim Biophys Acta 1793: 1508-1515.

Trifunovic A, et al (2004) Premature ageing in mice expressing defective mitochondrial DNA polymerase. Nature 429: 417-423.

Van Dyck E, Foury F, Stillman B, Brill SJ (1992) A single-stranded DNA binding protein required for mitochondrial DNA replication in S. cerevisiae is homologous to E. coli SSB. EMBO J 11: 3421-3430.

Vanderstraeten S, Van den Brule S, Hu J, Foury F (1998) The role of 3'-5' exonucleolytic proofreading and mismatch repair in yeast mitochondrial DNA error avoidance. J Biol Chem 273: 23690-23697.

Venditti C, Meade A, Pagel M (2006) Detecting the node-density artifact in phylogeny reconstruction. Syst Biol 55: 637-643.

Venditti C, Meade A, Pagel M (2008) Phylogenetic mixture models can reduce node-density artifacts. Syst Biol 57: 286-293.

Wall MK, Mitchenall LA, Maxwell A (2004) Arabidopsis thaliana DNA gyrase is targeted to chloroplasts and mitochondria. Proc Natl Acad Sci 101: 7821-7826.

Wallace DC (2005) A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: A dawn for evolutionary medicine. Annu Rev Genet 39: 359-407.

Wolfe KH, Li WH, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci 84: 9054-9058.

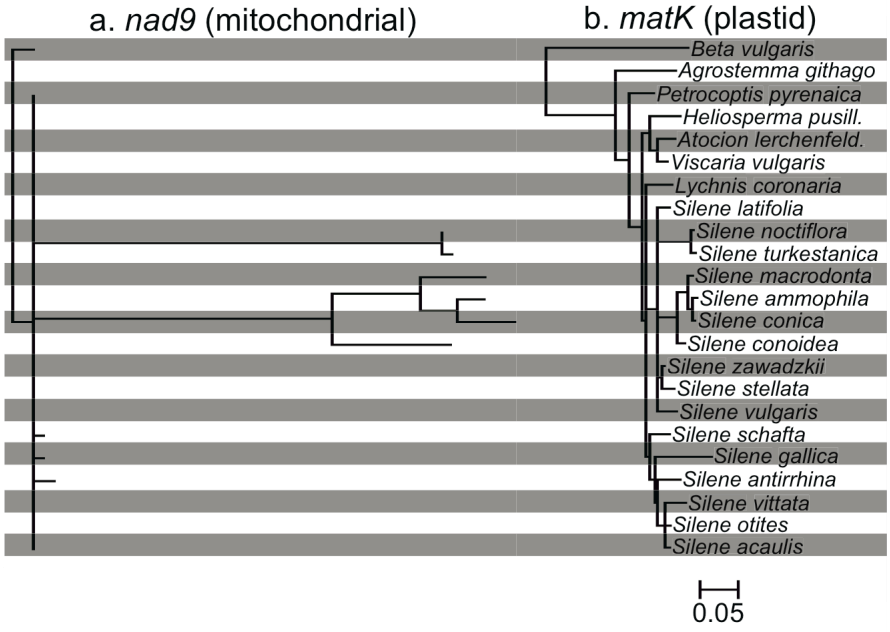
Yang D, Oyaizu Y, Oyaizu H, Olsen GJ, Woese CR (1985) Mitochondrial origins. Proc Natl Acad Sci 82: 4443-4447.

Zaegel V, et al (2006) The plant-specific ssDNA binding protein OSB1 is involved in the stoichiometric transmission of mitochondrial DNA in arabidopsis. Plant Cell 18: 3548-3563.

Zubko MK, Day A (2002) Differential regulation of genes transcribed by nucleus-encoded plastid RNA polymerase, and DNA amplification, within ribosome-deficient plastids in stable phenocopies of cereal albino mutants. Mol Genet Genomics 267: 27-37.

**Figures**

Figure 1. An example of recent and extreme acceleration in mitochondrial substitution rates within the angiosperm genus *Silene*. (a) A phylogenetic tree with branch lengths representing the number of substitutions per synonymous site in the mitochondrial gene *nad9*. (b) For comparison, a tree based on a gene from the plastid genome (which does not exhibit comparable increases in substitution rate). Both trees are based on previously published data (Sloan *et al*. 2009).



**Tables**

Table 1. Phylogenetic estimates of mitochondrial mutation rates

| Taxon | Mitochondrial Synonymous Substitution Rate ($x10^{-9}$ per site per *year*) | Source |
|---|---|---|
| Mammals | 18.2 - 54.5 | Wolfe et al. 1987[a] |
| | 7.0 - 643.4 | Nabholz et al. 2008[b] |
| Birds | 3.0 - 90.0 | Nabholz et al. 2009[b] |
| Amphibians | 13.8 - 21.6 | Lynch et al. 2006 |
| Insects | 16.6 - 34.0 | Lynch et al. 2006 |
| Seed Plants | 0.2 - 1.1 | Wolfe et al. 1987[a] |
| | 0.02 - 90.1 | Mower et al. 2007 |

[a]The early study of Wolfe et al. was limited to a small number of species comparisons within both mammals and seed plants. The wider ranges of rate estimates in subsequent studies reflect much more extensive sampling in these groups.

[b]Based on the original authors' interpretation, the indicated ranges for each of the Nabholz et al. studies exclude the most extreme 5% of rate estimates as potentially unreliable outliers.

Table 2. Estimates of mitochondrial mutation based on resequencing of laboratory mutation accumulation lines

| Species | Mitochondrial Mutation Rate ($x10^{-9}$ per site per *generation* +/- SEM) | Source |
|---|---|---|
| *Caenorhabditis briggsae* (HK104) | 110 (+/- 38) | Howe et al. 2010 |
| *Caenorhabditis briggsae* (PB800) | 72 (+/- 34) | Howe et al. 2010 |
| *Caenorhabditis elegans* | 97 (+/- 27) | Denver et al. 2000 |
| *Drosophila melanogaster* (Florida) | 43 (+/- 19) | Haag-Liautard et al. 2008[a] |
| *Drosophila melanogaster* (Madrid) | 81 (+/- 24) | Haag-Liautard et al. 2008[a] |
| *Saccharomyces cerevisiae* | 12.2 (+/- 3.6) | Lynch et al. 2008 |

[a]The standard errors indicated for the *Drosophila* lines were approximated as ¼ of the 95% confidence intervals reported in the original study.

# Chapter 2.

## Testing for selection on synonymous sites in plant mitochondrial DNA: the role of codon bias and RNA editing[1]

[1]Formatted as a co-authored manuscript and published as:

Sloan DB, Taylor DR. 2010. *J. Mol. Evol.* 70(5):479-491

Referenced supplementary material is available online at:

http://www.springerlink.com/content/mv56225h42032841/

ABSTRACT

Because plant mitochondrial genomes exhibit some of the slowest known synonymous substitution rates, it is generally believed that they experience exceptionally low mutation rates. However, the use of synonymous substitution rates to infer mutation rates depends on the implicit assumption that synonymous sites are evolving neutrally (or nearly so). To assess the validity of this assumption in plant mitochondrial genomes, we examined coding sequence for footprints of selection acting at synonymous sites. We found that synonymous sites exhibit an AT rich and pyrimidine skewed nucleotide composition compared to both non-synonymous sites and non-coding regions. We also found some evidence for selection associated with both biased codon usage and conservation of regulatory sequences involved in mRNA processing, although some of these findings are subject to alternative non-adaptive interpretations. Regardless, the inferred strength of selection appears too weak to account for the variation in substitution rates between the mitochondrial genomes of plants and other multicellular eukaryotes. Therefore, these results are consistent with the interpretation that plant mitochondrial genomes experience a substantially lower mutation rate rather than increased functional constraints acting on synonymous sites. Nevertheless, there are important nucleotide composition patterns (particularly the differences between synonymous sites and non-coding DNA) that remain largely unexplained.

## INTRODUCTION

Disentangling the effects of selection and mutation is one of the most vexing challenges in the field of molecular evolution. Patterns in DNA sequence data can simultaneously reflect either mutation bias or preferential fixation of certain mutations. For example, differences in nucleotide composition between complementary DNA strands may result from asymmetrical mutation pressures on the leading and lagging strand in DNA replication or from differential selection pressures arising from differences in gene density between the two strands (Morton and Morton 2007). The classic approach to separating these confounded effects relies on identifying a set of sites that are evolving neutrally to use as a basis for comparison (Li et al. 1981; Hughes and Nei 1989; McDonald and Kreitman 1991; Petrov et al. 1996). In the absence of selection, the substitution rate should provide an unbiased estimate of the mutation rate (Kimura 1983). Selection can then be inferred by observing departures from the neutral rate. The difficulty with these methods lies in finding a suitable set of sites that is (1) effectively free of selection, (2) large enough to provide statistically precise estimates of mutation parameters, and (3) representative of the mutational environment experienced by the region of interest.

Synonymous sites (i.e. those that do not change the amino acid sequence because of the redundancy of the genetic code) are often assumed to be relatively neutral, and hence a useful basis of comparison for detecting selection or estimating mutation rates. Synonymous sites are abundant in coding regions, and because synonymous and non-synonymous sites are interspersed along the length of each gene, both classes presumably experience similar mutation pressures. However, a blanket assumption of neutrality at

synonymous sites is unrealistic. Although these sites should be free from selection acting at the protein sequence level, there are a number of selective forces that are not dependent on amino acid sequence, including preferential use of codons for translational efficiency, maintenance of regulatory sequences, and conservation of mRNA secondary structure and stability (Chamary et al. 2006). Such effects can be distributed across the entirety of a gene (Ikemura 1985) or localized to specific sites (Kimchi-Sarfaty et al. 2007; Kudla et al. 2009).

The mitochondrial genomes of land plants exhibit intriguing patterns of evolution at synonymous sites. Generally, their rates of synonymous substitution are low relative to nuclear and chloroplast genomes as well as to the mitochondrial genomes of other eukaryotes (Wolfe et al. 1987; Palmer and Herbon 1988; Drouin et al. 2008). In the last decade, however, a growing number of plant species have been identified as exceptions to this rule, with rates of synonymous substitution that are orders of magnitude higher than other plants (Cho et al. 2004; Parkinson et al. 2005; Bakker et al. 2006; Mower et al. 2007; Sloan et al. 2008; Sloan et al. 2009, Ran et al. 2010). Examples of rate variation among genes and among lineages within species have also been documented (Barr et al. 2007; Sloan et al. 2008; Sloan et al. 2009). The general interpretation of these findings has been that the variation in synonymous substitution rate reflects the underlying mutation rate. Therefore, low mutation rates are believed to be the norm for plant mtDNA with occasional cases of dramatic mutational acceleration.

It is conceivable, however, that variation in synonymous substitution rates among mitochondrial genomes is a product of variance in selective constraint rather than mutation rate. For example, selective pressures that act on synonymous sites in other

genomes, e.g. codon usage bias, could be more intense in plant mtDNA, or the unique requirements for the expression of plant mtDNA could be responsible for novel sources of selection. In particular, C-to-U RNA editing is a widespread process in plant mitochondrial genomes in which cytidines are systematically converted to uridines by deamination at the mRNA level (Yu and Schuster 1995). This phenomenon affects hundreds of (mostly non-synonymous) sites, restoring mRNA codons so that the resulting peptides retain phylogenetically conserved amino acids (Giege and Brennicke 1999). Although the machinery responsible for RNA editing in plant organelle genomes is not yet fully characterized, it is clear that the process depends on the recognition of the sequence surrounding the editing site, most likely by members of the pentatricopeptide repeat (PPR) gene family (Kotera et al. 2005; Rüdinger et al. 2008; Zehrmann et al. 2009; Hammani et al. 2009). Comparisons across editing sites within a genome have shown that particular nucleotides are preferentially found at positions immediately adjacent to editing sites (Farre et al. 2001; Mulligan et al. 2007). Although this pattern of conservation across editing sites appears to be fairly weak and restricted to a narrow window, it is also possible that each editing site maintains a unique recognition sequence corresponding to its own recognition elements. Indeed, targeted deletion experiments have shown that sequence information up to 40 bp away from an editing site can be important for efficient processing (Choury et al. 2004; Takenaka et al. 2004; Hayes et al. 2006). Therefore, the recognition of RNA editing sites may represent a novel selective constraint that restricts the rate of mitochondrial sequence divergence among plant species.

To assess the potential importance of selection on synonymous sites in plant mtDNA, we analyzed the coding sequence from all published land plant mitochondrial genome sequences. Although we found some evidence for preferential codon usage and conservation of recognition sequences around RNA editing sites, neither of these selective forces are large enough to explain the variation in substitution rates between plants and other eukaryotes. We discuss the implications of these findings for interpreting patterns of molecular evolution in plant mitochondrial genomes as well as the general limitations for teasing apart the effects of selection and mutation bias at synonymous sites.

## METHODS

**Plant mtDNA Sequence Data**. The full length genomic sequences for all protein coding genes and their associated introns from 18 complete land plant mitochondrial genomes were downloaded from GenBank with custom scripts utilizing BioPerl modules (Table 1) (Stajich et al. 2002). The resulting dataset was manually curated to correct assembly errors associated with trans-spliced genes and to remove a large number of unknown open reading frames (ORFs) that are likely to be non-functional. Trans-spliced introns were excluded from the dataset, because their physical bounds are typically undefined. Homologous genes were aligned and trimmed to remove large 5' or 3' extensions that may have resulted from ambiguous gene models. To examine the effects of RNA editing, cDNA sequences were obtained from the REDIdb database for 3 species that have been the subject of whole-genome analysis: *Arabidopsis thaliana*, *Beta vulgaris*, and *Oryza*

*sativa* (Picardi et al. 2007). To serve as a basis for comparison, coding sequences were also downloaded for a diverse sample of animal mitochondrial genomes.

**Analysis of Codon Usage.** The program CodonW v1.4.4 (Peden 2000) was used to calculate GC content at synonymous third codon positions ($GC_{3S}$) and the effective number of codons ($N_c$) for a given mitochondrial sequence. $N_c$ is a standard measure of codon bias (Wright 1990). At one extreme, $N_c = 20$ if only a single codon is used for each of the 20 amino acids. At the other extreme, if all synonymous codons are used with equal frequency, then $N_c = 61$, reflecting the total number of sense codons (under the standard genetic code). Reductions in $N_c$ can result from selection for preferred codon usage as well as neutral forces such as mutation bias. The expected value for $N_c$ based purely on nucleotide composition can be approximated as a function of $GC_{3S}$ (Wright 1990):

$$E(N_c) \approx 2 + GC_{3S} + \frac{29}{GC_{3S}^{\ 2} + (1 - GC_{3S})^2} \qquad \text{(Eq. 1)}$$

$GC_{3S}$ and $N_c$ were calculated based on each gene individually as well as based on a concatenation of all mitochondrial genes for each species. These analyses were also repeated on the RNA editing dataset.

CodonW was also used to perform a correspondence analysis on the mitochondrial cDNA sequences of *Arabidopsis thaliana*. Correspondence analysis is an ordination technique that reduces the multi-dimensional nature of codon usage data. For

each gene, relative synonymous codon usage (RSCU; Sharp et al. 1986) values are calculated for all sense codons other than ATG (Met) and TGG (Trp), which are the only codons for their respective amino acids. Therefore, each gene is associated with 59 different RSCU values, which can be reduced to a set of orthogonal axes. RSCU values provide a measure of whether a codon is over-represented (RSCU>1) or under-represented (RSCU<1) given the frequency of an amino acid and the number of different codons for that amino acid.

The principal axis from a correspondence analysis is often assumed to reflect different intensities of selection for preferred codon usage among genes (e.g. those with high vs. low expression levels; Grantham et al. 1981; Liu et al. 2004; Zhang et al. 2007; Zhou and Li 2008). However, it is important to test this assumption with independent measures of expression, because other mechanisms such as local differences in mutation patterns can generate differences in codon usage patterns across genes (Peden 2000; Duret 2002). *Arabidopsis* was chosen for this analysis because of the availability of microarray based expression data for mitochondrial genes. We used normalized (GCRMA $\log_2$) expression values for the first two true leaves from the AtGenExpress development series dataset (Schmid et al. 2005). Affymetrix probe set identifiers for each gene are provided as supplementary material (Table S1). Pearson correlation coefficients were calculated to test for a relationship between either the expression level or $GC_{3S}$ content of a gene and the position of that gene on the principal axis from the correspondence analysis.

To test for a genome-wide pattern of selection on synonymous codon usage, a pair of maximum likelihood models was implemented in PAML v4.1 (Yang 2007). Yang and

Nielsen (2008) recently developed a likelihood-based approach for separating the confounded effects of mutation bias and selection on synonymous codon usage. In one model (FMutSel), a unique fitness coefficient (scaled by the effective population size, $N_e$) is applied to all 61 sense codons. In a second more constrained model (FMutSel0), all codons that code for the same amino acid share the same fitness coefficient. In each model, all fitness coefficients are expressed relative to a single codon whose value is arbitrarily set to 0. Therefore, there are 60 free selection parameters in FMutSel model, but only 19 in FMutSel0. Both models implement an HKY model of nucleotide substitution to account for mutation bias and a genome-wide $d_N/d_S$ ratio to account for selection on non-synonymous mutations. A likelihood ratio test (LRT) with 41 (=60-19) degrees of freedom can be used to compare the two models. If FMutSel provides a significantly better fit to the data, it suggests that there is selection acting on synonymous codon usage. Furthermore, the magnitude of selection on any synonymous mutation can be calculated based on the absolute value of the difference in scaled fitness coefficients between the two synonymous codons. Averaging these values over all possible pairs of synonymous codons differing by a single nucleotide provides an estimate of the mean intensity of selection on synonymous mutations (ΔF). These models were used to analyze the pairwise divergence between *Arabidopsis* and *Oryza* for a concatenated dataset of all protein-coding mitochondrial cDNAs. *Arabidopsis* and *Oryza* were selected, because they represented one of the most divergent species pairs for which complete RNA editing data were available, while still maintaining similar nucleotide compositions (Table 1) to better conform to the assumptions of the reversible and stationary phylogenetic model.

**Conservation of Recognition Sequences around RNA Editing Sites.** To assess the potential effect of RNA editing on the synonymous substitution rate, mitochondrial cDNA sequences from *Arabidopsis thaliana*, *Beta vulgaris*, and *Oryza sativa* were aligned. If the maintenance of recognition sequences for RNA editing acts as a selective constraint, we would predict that substitution rates would be reduced in the immediate vicinity of editing sites that are conserved among the species. Therefore, the extent of synonymous divergence was determined based on the proximity of each nucleotide to a conserved RNA editing site. Divergence at four-fold synonymous sites was calculated for windows of both -40/+20 bp and -2/+2 bp around conserved editing sites and compared to the level of synonymous divergence at sites >50bp from the nearest editing site. These windows were selected *a priori* based on empirical studies identifying the flanking sequences that are important for efficient editing (Farre et al. 2001; Choury et al. 2004; Takenaka et al. 2004; Hayes et al. 2006; Mulligan et al. 2007). The larger (-40/+20 bp) window encompasses the maximum range of flanking sequence that has been shown to affect editing site recognition, and its asymmetry reflects the greater importance of upstream sequence. The smaller (-2/+2) window spans the immediate flanking nucleotides, which exhibit enough conservation to define a general recognition motif for editing sites (Mulligan et al. 2007). Statistical significance for differences in divergence was assessed, using one-tailed Fisher's exact tests. For a more fine scale assessment of patterns of sequence conservation, divergence was calculated for a 6 bp sliding window ranging from 50 bp before to 50 bp after conserved editing sites. For each of these analyses, divergence was calculated based on a comparison of 4-fold synonymous sites, excluding codons for which the amino acid was not conserved between species. Each

analysis was performed separately on the pairwise alignment between *Arabidopsis* and *Oryza* as well as between *Arabidopsis* and *Beta*.

## RESULTS

**Codon Usage in Plant mtDNA.** As is the case in almost all genomes, plant mitochondrial genes do not utilize synonymous codons with equal frequency. A comparison of nucleotide composition at 4-fold synonymous sites revealed that, on average, there is a much greater abundance of codons ending in T (37.4%) or A (29.8%) than in C (18.1%) or G (14.6%), reflecting both a low GC content and bias towards pyrimidines on the coding strand. Notably, the AT richness at synonymous sites was consistently greater than in the rest of the genome (both coding and non-coding; Table 1). Intron sequences, in particular, showed much higher GC content as well as a contrasting skew towards purines rather than pyrimidines. For the most part, the nucleotide composition at synonymous sites appeared to be consistent across amino acids, genes and phylogenetic lineages (Figure 1; Figures S1-S2). There were, however, some exceptions. For example, *matR* (the sole intron encoded gene) was a consistent outlier relative to other genes with $GC_{3S}$ values in excess of 55% (Table S2; Unseld et al. 1997). In addition, whole genome comparisons of $GC_{3S}$ values revealed two distinct species clusters. While all of the vascular plants had $GC_{3S}$ values between 34 and 38%, three independent bryophyte lineages had $GC_{3S}$ values between 24 and 28%. These two groups also exhibited a corresponding divergence in codon usage bias with the bryophytes maintaining more skewed usage ($N_c$ = 47 to 51) than the vascular clade ($N_c$ = 53 to 56). In each case, the $N_c$ values fall very close to but slightly below the levels predicted based

on GC content alone (Figure 2). In comparison, animal mitochondrial genomes exhibited a broader range of $GC_{3S}$ values—often with more substantial departures from the corresponding $N_c$ predictions (Figure 2).

**The Effects of RNA Editing on Synonymous Codon Usage.** The preceding results were based on an analysis of genomic sequence data. Therefore, they ignored the potential effects of RNA editing, which is pervasive in the mitochondrial genomes of land plants. A comparison of genomic and cDNA sequences in three angiosperms showed a small but consistent effect of RNA editing on codon usage (Figure 3). Although both C-to-U and U-to-C editing occur in land plants, the latter is rare or absent altogether in angiosperms. As a result, RNA editing decreases $GC_{3S}$ values. However, this decrease is small (a genome-wide average of 0.63% in the three species), because editing at the third codon position is infrequent.

Editing also results in slightly more biased codon usage, illustrated by a mean $N_c$ reduction of 0.93. Table 2 summarizes the effects of editing on each codon in *Arabidopsis*. Codons that appear preferentially in genomic DNA (RSCU > 1) exhibit a slight increase in relative usage after editing (mean $\Delta_{RSCU} = 0.021$), while underrepresented codons (RSCU < 1) experience a slight decrease (mean $\Delta_{RSCU} = -0.018$). The difference between these two classes was marginally significant ($t_{df=57} = 1.99$; $p = 0.05$). Accounting for RNA editing also explained some of the apparent variation in codon usage across species. For example, the frequency of the CGG (Arg) codon in genomic sequence was highly variable across species (Figure 2). CGG was overrepresented (33.0% of CGN codons) in *Isoetes*, whereas it was exceedingly rare

(<4.0%) in *Marchantia and Physcomitrella*. The codon was found at intermediate frequencies in other species. This codon, however, appears to be a frequent target of conversion to UGG (Trp) by RNA editing, making it rare in the mRNA of all species. Therefore, variation among species in the frequency of genomic CGG codons largely reflects divergence in the extent of RNA editing rather actual differences in synonymous codon usage (Giege and Brennicke 1999; Chaw et al. 2008; Rüdinger et al. 2008; Grewe et al. 2009; Rüdinger et al. 2009).

**Selection on Codon Usage.** An analysis of mitochondrial sequence divergence between *Arabidopsis* and *Oryza* found significant evidence for a genome-wide pattern of selection on synonymous sites. A likelihood ratio test strongly favored the more complex FMutSel model, which applies a unique fitness coefficient to every codon, over a null model that constrains all synonymous codons to share the same coefficient ($\chi^2_{df=41} = 206.9$; $p << 0.001$). The magnitude of selection inferred from the model, however, tended to be quite small. The average magnitude of pairwise difference in scaled fitness coefficients ($|N_e s|$) among synonymous codons was $\Delta F = 0.81$. Yang and Nielsen (2008) performed the same analysis on mitochondrial divergence between humans and chimpanzees, and showed highly significant selection on synonymous codons ($\chi^2_{df=41} = 281.4$; $p << 0.001$). We used their data to calculate a $\Delta F$ value of 1.35. Because $|N_e s| = 1$ is often viewed as an approximate threshold below which neutral forces begin to dominate over selection, the $\Delta F$ values from both studies suggest a generally weak pattern of selection on synonymous sites.

Correspondence analysis was used to summarize patterns of codon usage across genes within the *Arabidopsis* mitochondrial genome. The principal axis explained 23.2% of the variance in codon usage (based on RSCU values), while the second and third axes explained 10.2% and 7.8%, respectively. It did not appear that the major trend in the data could be explained by differences in selection pressure associated with high vs. low expression. There was essentially no relationship between the expression level of a gene and its placement on the principal axis from the correspondence analysis (Figure 4A). In contrast, there was a strong relationship between the first axis and GC content (Figure 4B), suggesting that local differences in nucleotide composition within the genome may be the driving force behind differences in codon usage across genes. Meanwhile, the second axis of the correspondence analysis was not significantly correlated with either gene expression or GC content (data not shown).

**Conservation of Recognition Sequences around RNA Editing Sites.** Synonymous substitution rates were lower in sequences flanking RNA editing sites than in other parts of the mitochondrial genome. Divergence at four-fold synonymous sites was significantly lower for nucleotides that fell within a 60 bp window (-40/+20 bp) around a conserved editing site than for nucleotides that were >50 bp away from an editing site (Figure 5). This was true for comparisons of *Arabidopsis* with both *Beta* ($p = 0.005$) and *Oryza* ($p = 0.05$). In both comparisons, divergence was even lower in a narrower 4 bp window (-2/+2 bp), but this rate was only significantly different from background in the *Arabidopsis-Beta* analysis ($p = 0.009$), perhaps reflecting a lack of power given the small the number of synonymous sites in the 4 bp window. A sliding window analysis showed that

synonymous substitution rates were generally lowest in the 20bp region immediately upstream of each editing site (Figure 6). The rates were quite variable, however, and their 95% confidence intervals generally overlapped with the baseline substitution rate observed in other parts of the genome.

## DISCUSSION

The AT richness and pyrimidine skew at synonymous sites in plant mtDNA (Table 1) are clearly non-random patterns, raising questions about the role that both adaptive and non-adaptive processes play in determining nucleotide composition and the rate of molecular evolution. The focus of this study was to test for potential sources of selection on synonymous sites, and we identified some evidence for such selection, which we discuss below.

**Genome-wide Patterns of Codon Usage.** Molecular biologists have long known that, convenient as it may be, the assumption of effective neutrality at synonymous sites is often unrealistic. The existence of purifying selection on synonymous sites can result in a downward bias on mutation rate estimates based on synonymous divergence (Ikemura 1985; Sharp and Li 1987; Denver et al. 2004). Likewise, variation in the intensity of purifying selection may falsely suggest differences in the underlying mutation rate. Therefore, identifying sources of selection on synonymous sites is necessary to interpret patterns of molecular evolution in coding sequence data.

Preferential codon usage is one of the most well known forms of selection on synonymous sites. While it is thought to be pervasive in prokaryotes with large $N_e$, its

role in species with smaller $N_e$ (e.g. multicellular eukaryotes) may be more limited. This study of plant mitochondrial genomes revealed evidence for significant—albeit weak—selection on synonymous codon usage based on a hierarchical model comparison in PAML. As is the case with any test of selection, it is important to consider alternative interpretations of this analysis. In particular, it should be stressed that a significant LRT indicates that observed codon frequencies cannot be fully explained by the assumed model for mutation and selection at the amino acid level. While such a failure may be explained by the action of selection on synonymous codons, it may also reflect an incomplete and improper model of the other substitutional processes. As noted by Yang and Nielsen (2008), violations of phylogenetic assumptions such as the independence of substitutions among sites could falsely support the existence of selection on synonymous sites. Furthermore, their study found that the inferred magnitude of selection on synonymous mutations could depend greatly on the assumed mutational models. Given the high sensitivity of the FmutSel/FmutSel0 test (which detected significant evidence for selection on synonymous sites in more than 90% of mammalian genes analyzed; Yang and Nielsen 2008), it is important to find corroborating evidence for selection.

A close correspondence between $N_c$ values and the level predicted based on $GC_{3S}$ alone is typically interpreted as evidence for neutral causes of biased codon usage (Wright 1990). Therefore, the relationship between GC content and codon usage in plant mtDNA (Figure 2) offers little support for a role of selection in shaping codon usage. However, this interpretation assumes that selection does not exhibit consistent nucleotide preferences across codon groups. In other words, if selection favored the same

nucleotide(s) at all synonymous sites, it could simultaneously affect both $N_c$ and $GC_{3S}$ values in a way that superficially appears non-adaptive.

Much of the existing evidence for selection on synonymous codon usage involves optimizing the efficiency of translation given the relative abundance of different tRNAs (Ikemura 1985). The possibility of selection for translational efficiency in plant mitochondria is intriguing considering the complex mixture of tRNAs involved in the process. Gene expression in plant mitochondria relies on the coordinated action of tRNAs with three different evolutionary histories (Glover et al. 2001). First, plant mitochondrial genomes contain a set of "native" tRNAs that are descended from the genome of the α-proteobacterial ancestor of mitochondria. Second, transfer of DNA between organelle genomes has resulted in the functional acquisition of "chloroplast-like" tRNAs that are descended from the genome of the cyanobacterial ancestor of chloroplasts. Finally, plant mitochondria also import a number of (nuclear encoded) tRNAs from the cytosol. This tRNA mélange appears to be evolutionarily labile, as plant species differ widely in the number of tRNA genes found within their mitochondrial genomes (Kubo and Newton 2008; Grewe et al. 2009).

Given what is known about mitochondrial tRNA populations, what evidence is there that selection has acted to optimize synonymous codon usage in plant mitochondrial genomes? Unfortunately, to the best of our knowledge, a comprehensive quantification of expressed tRNAs is not available for plant mitochondria, so it is not possible to correlate codon usage with corresponding tRNA abundances. Nevertheless, it is possible to make some general assessments based on the presence or absence of different tRNA species. Plant mitochondria do not contain a full set of tRNAs with all 61 anticodons. Rather, they

utilize a reduced set of approximately 32-35 tRNAs (Glover et al. 2001), which rely on third position "wobble" pairing to successfully recognize all possible codons (Crick 1966). Notably, in the *Triticum* mitochondrial tRNA population (which is the most thoroughly characterized of any of the species in this study), all synonymous codon pairs ending in C/T are recognized by the same tRNA (Glover et al. 2001). Therefore, the much higher frequency of T than C at synonymous sites (Figure 1; Figure S1) cannot be explained by differences in tRNA abundance. Moreover, the standard rules for preferential wobble pairing (Ikemura 1985) would predict that C (not T) would be favored at synonymous positions for amino acids such as Asn, Asp, Cys, Gly, His, Phe, Ser, and Tyr, because of the G at the first anticodon position of the corresponding tRNA. No preference is expected for synonymous C/T codon pairs for other amino acids, because the corresponding tRNA uses the modified nucleoside inosine at the first anticodon position, which is not thought to discriminate between C vs. T in wobble pairing (Ikemura 1985; Glover et al. 2001). Therefore, given our current understanding of plant mitochondrial tRNA populations, it is difficult to interpret the major trends in synonymous codon usage within plant mitochondrial genomes as evidence for selection on translational efficiency.

Regardless of the discrepancy between the PAML analysis and other measures of selection on synonymous sites, all analyses fail to implicate preferential codon usage and the associated functional constraint as an explanation for the genome-wide reduction in plant mitochondrial substitution rates relative to other eukaryotes. Codon bias appears to be similar or even more extreme in animal mitochondrial genomes, which have

synonymous substitution rates that are generally orders of magnitude higher than in plants.

**Variation among Genes in Codon Usage.** Previous studies of plant mitochondrial genomes have concluded that variation in selection associated with translational demands is a major contributor to the differences in codon usage among genes (Liu et al. 2004; Zhang et al. 2007; Zhou and Li 2008). None of these studies, however, have confirmed that patterns of codon usage are correlated with any direct measure of gene expression. The apparent absence of such a correlation in *Arabidopsis thaliana* (Figure 4A) raises questions about the importance of translational selection inferred by previous studies. However, there are some limitations to the use of microarray-based measurements of gene expression, which should be considered. First, they assay expression at the transcription level, whereas selection on codon usage would be more directly related to the amount or rate of translation. Second, because polyadenylation is associated with transcript degradation in plant mitochondria (Holec et al. 2006; Adamo et al. 2008), the use of polyA selected cDNA libraries may yield "expression" levels that are more related to transcript turnover than total transcript number. Finally, the existence of paralogous gene copies in the nucleus (Stupar et al. 2001) may result in an inaccurate measure of actual mitochondrial genome expression. Therefore, more direct measures of the rates of transcription (or better yet translation) in plant mitochondrial genomes would be valuable for testing whether variation in codon usage among genes is related to their expression.

Contradictions among earlier studies cast further doubt on the evidence for differential selection on codon usage across mitochondrial genes. For example, Zhang et

al. (2007) and Zhou and Li (2008) both analyzed the *Triticum aestivum* mitochondrial genome to identify a set of "preferred codons". Each study identified about ten preferred codons, but the two sets only have a single codon in common—no more than would be expected at random. Therefore, the available data appear insufficient to reject the null hypothesis that codon usage differences among genes are the result of local differences in nucleotide composition and stochastic variation. Notably, the gene with the most divergent pattern of codon usage (*matR*) is also the only intron-encoded gene, found within domain IV of a group II intron within *nad1*. Given the high GC content of plant mitochondrial introns (Table 1), it is intriguing that *matR* is the one gene with $GC_{3S}$ values consistently above 50% (Table S2).

It is also important to note that the total amount of codon usage variation among genes is relatively low, especially in comparison with the amount of variation in plant nuclear genomes (Zhang et al. 2007). Therefore, even if selection is responsible for a significant portion of the variance in codon usage among mitochondrial genes, it may still have a relatively small effect.

**Selection around RNA Editing Sites.** In contrast to codon bias, which must be stronger in plants than other eukaryotes to explain the differences in mitochondrial substitution rate, widespread RNA editing does not exist in most eukaryotes and, therefore, represents a potential source of novel selection in plants. This study did find a reduced synonymous substitution rate around editing sites, which is consistent with a role of purifying selection maintaining recognition sequences for editing machinery.

First and foremost, it should be stressed that this result is only a correlation and one that is subject to alternative explanations. In particular, it is possible that genomic regions with higher mutation rates are relatively devoid of RNA editing sites. A simple mechanism to lose RNA editing sites is for genomic C-to-T mutations to "overwrite" the editing that would have been done at the mRNA level. Chloroplast RNA editing sites have been shown to preferentially occur in contexts that experience low mutation rates (Tillich et al. 2006), suggesting that the rate of mutational loss may be an important determinant of editing site distribution. In addition, the rate of editing site loss may be higher in regions with frequent DNA damage and double stranded breaks if repair occasionally involves recombination/gene conversion with reverse transcribed cDNA (Parkinson et al. 2005; Mulligan et al. 2007). An adaptive argument has also been made, suggesting that a high mutation rate itself selects for the loss (or against the proliferation) of editing sites by increasing the probability that essential recognition sequences will be disrupted by mutation (Lynch et al. 2006). These alternative interpretations highlight the difficulty of distinguishing the roles of selection and mutation in molecular evolution.

Even ignoring the possibility that mutation variation rather than purifying selection explains the lower rates of synonymous substitution around editing sites, it does not appear that the effect is particularly strong in the context of the genome-wide substitution rate. The synonymous substitution rate in nucleotides within a 60 (-40 to +20) bp window around conserved editing sites appears to be reduced by less than 30%. In our angiosperm comparisons, about one third of mitochondrial coding sequence fell within such a window, so the net reduction on the overall synonymous substitution rate could be as much as 10%. Such an effect is fairly trivial relative to the difference in

mitochondrial substitution rates between plants and other eukaryotes, which can span multiple orders of magnitude.

The two base pairs on either side of RNA editing site represent the one area where there is enough conservation across editing sites to define a consensus recognition motif—albeit a degenerate one: HY**C**GK (Mulligan et al. 2007). Therefore, it is not surprising that synonymous substitution rates were particularly low in this range (Figure 5). Nevertheless, even in this most critical window, conservation was not absolute. This finding suggests a high degree of flexibility in the recognition machinery involved in RNA editing, a conclusion that is supported by the recent identification of *trans*-acting factors that are required for the editing of multiple target sites with very limited sequence similarity (Chateigner-Boutin et al. 2008; Kobayashi et al. 2008; Okuda et al. 2009; Zehrmann et al. 2009; Hammani et al. 2009).

**Nucleotide Composition at Synonymous Sites**

The striking differences in nucleotide composition between synonymous sites and intronic/intergenic DNA in plant mitochondrial genomes beg for an explanation (Table 1). The contrast in AT richness and pyrimidine skew is reminiscent of the pattern observed in chloroplast genomes. Morton (2003) showed that low GC content at synonymous sites in cpDNA is likely the result of context-dependent mutational bias, while the higher GC content in non-coding DNA is not in mutational equilibrium. It has been proposed that recent insertions in non-coding DNA could explain the deviation because they would not yet have had the time to reach equilibrium (Morton 2003). This hypothesis has some appeal in the context of plant mitochondrial genomes because of the

high turnover rate and mysterious origins of intergenic mtDNA (Unseld et al. 1997; Kubo and Newton 2008). It is difficult, however, for this mechanism to explain the high GC content of introns, many of which have been maintained over long evolutionary timescales (Qiu et al. 1998). Moreover, given the low average GC content in plant nuclear and chloroplast genomes (Shimada and Sugiura 1991; Arabidopsis Genome Initiative 2000), potential sources of high GC content insertions are not immediately obvious.

Therefore, it is also important to consider the role of selection in generating the differences between synonymous sites and non-coding DNA. For example, the high GC content in introns might reflect the constraints of maintaining a stable secondary structure in these self-splicing ribozymes (Michel et al. 1989; Löhne and Borsch 2005). Likewise, it has been suggested that intergenic sequences can be subject to substantial selective pressures (Andolfatto 2005). Unfortunately, the high rate of structural evolution in plant mitochondrial genomes makes it difficult to analyze substitution patterns in intergenic regions.

**Detecting Selection on Synonymous Sites.** Not surprisingly, this analysis suggests that synonymous sites do not behave in a completely neutral fashion in plant mitochondrial genomes. The evidence for selection, however, points to only weak effects, which are far from sufficient to explain the greatly reduced substitution rates in plant mtDNA. Therefore, these data are consistent with the common interpretation that plant mitochondrial genomes generally experience low rates of point mutations. In reaching this conclusion, it is wise to consider the detection limits for identifying selection on

synonymous sites. Statistical methods generally require averaging across large numbers of sites (e.g. across genes, species, or codon groups). As a result, selection pressures that are specific to individual sites may go undetected, particularly if such pressures act in opposite directions at different sites. Recent experimental approaches have illustrated the potential to determine the functional consequences of individual synonymous mutations (Kudla et al. 2009). Coupling such approaches with broader bioinformatic analyses should expand our capabilities to detect patterns of selection acting on synonymous mutations.

ACKNOWLEDGEMENTS

We would like to thank Janis Antonovics, Stefan Bekiranov, Lei Li and Martin Wu for helpful discussion of our results. This work was supported by a grant from the NSF (DEB-0808452).

REFERENCES

Adamo A, Pinney JW, Kunova A, Westhead DR, Meyer P (2008) Heat stress enhances the accumulation of polyadenylated mitochondrial transcripts in *Arabidopsis thaliana*. PLoS One 3:e2889.

Andolfatto P (2005) Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437:1149-1152.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796-815.

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

Barr CM, Keller SR, Ingvarsson PK, Sloan DB, Taylor DR (2007) Variation in mutation rate and polymorphism among mitochondrial genes in *Silene vulgaris*. Mol. Biol. Evol. 24:1783-1791.

Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat. Rev. Genet. 7:98-108.

Chateigner-Boutin AL, Ramos-Vega M, Guevara-Garcia A, Andres C, de la Luz Gutierrez-Nava M, Cantero A, Delannoy E, Jimenez LF, Lurin C, Small I, Leon P (2008) CLB19, a pentatricopeptide repeat protein required for editing of rpoA and clpP chloroplast transcripts. Plant J. 56:590-602.

Chaw SM, Shih AC, Wang D, Wu YW, Liu SM, Chou TY (2008) The mitochondrial genome of the gymnosperm *Cycas taitungensis* contains a novel family of short interspersed elements, Bpu sequences, and abundant RNA editing sites. Mol. Biol. Evol. 25:603-615.

Cho Y, Mower JP, Qiu YL, Palmer JD (2004) Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. Proc. Natl. Acad. Sci. 101:17741-17746.

Choury D, Farre JC, Jordana X, Araya A (2004) Different patterns in the recognition of editing sites in plant mitochondria. Nucleic Acids Res. 32:6397-6406.

Crick FH (1966). Codon--anticodon pairing: the wobble hypothesis. J. Mol. Biol. 19:548-555.

Denver DR, Morris K, Lynch M, Thomas WK (2004) High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. Nature 430:679-682.

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

Duret L (2002) Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. 12:640-649.

Farre JC, Leon G, Jordana X, Araya A (2001) cis Recognition elements in plant mitochondrion RNA editing. Mol. Cell. Biol. 21:6731-6737.

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

Glover KE, Spencer DF, Gray MW (2001) Identification and structural characterization of nucleus-encoded transfer RNAs imported into wheat mitochondria. J. Biol. Chem. 276:639-648.

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9:r43-74.

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

Hammani K, Okuda K, Tanz SK, Chateigner-Boutin AL, Shikanai T, 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

Hayes ML, Reed ML, Hegeman CE, Hanson MR (2006) Sequence elements critical for efficient RNA editing of a tobacco chloroplast transcript in vivo and in vitro. Nucleic Acids Res. 34:3742-3754.

Holec S, Lange H, Kuhn K, Alioua M, Borner T, Gagliardi D (2006) Relaxed transcription in *Arabidopsis* mitochondria is counterbalanced by RNA stability control mediated by polyadenylation and polynucleotide phosphorylase. Mol. Cell. Biol. 26:2869-2876.

Hughes AL, and Nei M (1989) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc. Natl. Acad. Sci. 86:958-962.

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2:13-34.

Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science 315:525-528.

Kimura M (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Kobayashi Y, Matsuo M, Sakamoto K, Wakasugi T, Yamada K, Obokata J (2008) Two RNA editing sites with cis-acting elements of moderate sequence identity are recognized by an identical site-recognition protein in tobacco chloroplasts. Nucleic Acids Res. 36:311-318.

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

Kubo T, Newton KJ (2008) Angiosperm mitochondrial genomes and mutations. Mitochondrion 8:5-14.

Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. Science 324:255-258.

Li WH, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292:237-239.

Liu Q, Feng Y, Xue Q (2004) Analysis of factors shaping codon usage in the mitochondrion genome of *Oryza sativa*. Mitochondrion 4:313-320.

Löhne C, Borsch T (2005) Molecular evolution and phylogenetic utility of the petD group II intron: a case study in basal angiosperms. Mol. Biol. Evol. 22:317-332.

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

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351:652-654.

Michel F, Umesono K, Ozeki H (1989) Comparative and functional anatomy of group II catalytic introns--a review. Gene 82:5-30.

Morton BR (2003) The role of context-dependent mutations in generating compositional and codon usage bias in grass chloroplast DNA. J. Mol. Evol. 56:616-629.

Morton RA, Morton BR (2007) Separating the effects of mutation and selection in producing DNA skew in bacterial chromosomes. BMC Genomics 8:369.

Mower JP, Touzet P, Gummow JS, Delph LF, Palmer JD (2007) Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. BMC Evol. Biol. 7:135.

Mulligan RM, Chang KLC, Chou CC (2007) Computational analysis of RNA editing sites in plant mitochondrial genomes reveals similar information content and a sporadic distribution of editing sites. Mol. Biol. Evol. 24:1971-1981.

Okuda K, Chateigner-Boutin AL, Nakamura T, Delannoy E, Sugita M, Myouga F, Motohashi R, Shinozaki K, Small I, 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.

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

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

Peden JF (2000) Analysis of codon usage. PhD Thesis. University of Nottingham.

Petrov DA, Lozovskaya ER, Hartl DL (1996) High intrinsic rate of DNA loss in *Drosophila*. Nature 384:346-349.

Picardi E, Regina TM, Brennicke A, Quagliariello C (2007) REDIdb: the RNA editing database. Nucleic Acids Res. 35:D173-177.

Qiu YL, Cho Y, Cox JC, Palmer JD (1998) The gain of three mitochondrial introns identifies liverworts as the earliest land plants. Nature 394:671-674.

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

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

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

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU (2005) A gene expression map of *Arabidopsis thaliana* development. Nat. Genet. 37:501-506.

Sharp PM, Li WH (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. 4:222-230.

Sharp PM, Tuohy TM, Mosurski KR (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14:5125-5143.

Shimada H, and Sugiura M (1991) Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. Nucleic Acids Res. 19:983-995.

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

Sloan DB, Barr CM, Olson MS, Keller SR, Taylor DR (2008) Evolutionary rate variation at multiple levels of biological organization in plant mitochondrial DNA. Mol. Biol. Evol. 25:243-246.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 12:1611-1618.

Stupar RM, Lilly JW, Town CD, Cheng Z, Kaul S, Buell CR, Jiang J (2001) Complex mtDNA constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2: implication of potential sequencing errors caused by large-unit repeats. Proc. Natl. Acad. Sci. 98:5099-5103.

Takenaka M, Neuwirt J, Brennicke A (2004) Complex cis-elements determine an RNA editing site in pea mitochondria. Nucleic Acids Res. 32:4137-4144.

Tillich M, Lehwark P, Morton BR, Maier UG (2006) The evolution of chloroplast RNA editing. Mol. Biol. Evol. 23:1912-1921.

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

Wolfe KH, Li WH, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. 84:9054-9058.

Wright F (1990) The 'effective number of codons' used in a gene. Gene 87:23-29.

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

Yang Z, Nielsen R (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol. Biol. Evol. 25:568-579.

Yu W, Schuster W (1995) Evidence for a site-specific cytidine deamination reaction involved in C to U RNA editing of plant mitochondria. J. Biol. Chem. 270:18227-18233.

Zehrmann A, Verbitskiy D, van der Merwe JA, Brennicke A, Takenaka M (2009) A DYW domain-containing pentatricopeptide repeat protein is required for RNA editing at multiple sites in mitochondria of *Arabidopsis thaliana*. Plant Cell 21:558-567.

Zhang WJ, Zhou J, Li ZF, Wang L, Gu X, Zhong Y (2007) Comparative analysis of codon usage patterns among mitochondrion, chloroplast and nuclear genes in *Triticum aestivum* L. J. Integrative Plant Biol. 49:246-254.

Zhou M, Li X (2009) Analysis of synonymous codon usage patterns in different plant mitochondrial genomes. Mol. Biol. Rep. 36: 2039-2046.

TABLE 1. Nucleotide composition in plant mitochondrial genomes

| | | | | | 4-Fold Synonymous Sites[a] | | Introns[b] | |
|---|---|---|---|---|---|---|---|---|
| **Bryophytes** | **GenBank Accession** | **Coding Genes** | **GC (%)** | **GC-Coding (%)** | **GC (%)** | **Purines (%)** | **GC (%)** | **Purines (%)** |
| *Marchantia polymorpha* | NC_001660 | 41 | 42.4 | 36.9 | 28.7 | 43.0 | 47.7 | 54.2 |
| *Megaceros aenigmaticus* | NC_012651 | 20 | 46.0 | 38.9 | 27.4 | 41.8 | 49.2 | 54.4 |
| *Physcomitrella patens* | NC_007945 | 20 | 40.6 | 35.8 | 23.8 | 42.3 | 46.3 | 54.2 |
| **Lycophytes** | | | | | | | | |
| *Isoetes engelmannii* | FJ010859; FJ176330; FJ390841; FJ536259; FJ628360 | 24 | 49.0 | 44.9 | 37.2 | 45.1 | 55.2 | 54.6 |
| **Gymnosperms** | | | | | | | | |
| *Cycas taitungensis* | NC_010303 | 39 | 46.9 | 45.4 | 35.4 | 46.3 | 50.8 | 52.6 |
| **Angiosperms** | | | | | | | | |
| *Arabidopsis thaliana* | NC_001284 | 31 | 44.8 | 42.6 | 32.5 | 44.5 | 50.7 | 52.5 |
| *Beta vulgaris* | NC_002511 | 30 | 43.9 | 41.8 | 32.2 | 44.2 | 50.2 | 52.8 |
| *Brassica napus* | NC_008285 | 32 | 45.2 | 42.7 | 32.6 | 44.4 | 51.1 | 52.4 |
| *Carica papaya* | NC_012116 | 38 | 45.1 | 43.1 | 33.3 | 44.7 | 51.8 | 53.1 |
| *Nicotiana tabacum* | NC_006581 | 37 | 45.0 | 42.7 | 32.4 | 44.6 | 51.1 | 52.5 |
| *Oryza sativa* | NC_011033 | 35 | 43.9 | 43.1 | 34.0 | 45.0 | 52.3 | 52.8 |
| *Sorghum bicolor* | NC_008360 | 32 | 43.7 | 42.8 | 33.8 | 44.8 | 52.4 | 52.9 |
| *Tripsacum dactyloides* | NC_008362 | 32 | 43.9 | 42.9 | 34.0 | 45.0 | 52.3 | 53.0 |
| *Triticum aestivum* | NC_007579 | 33 | 44.4 | 42.7 | 34.2 | 44.5 | 52.8 | 52.8 |
| *Vitis vinifera* | NC_012119 | 37 | 44.1 | 43.4 | 33.8 | 44.5 | 51.4 | 52.9 |
| *Zea luxurians* | NC_008333 | 32 | 43.9 | 43.0 | 34.6 | 45.0 | 52.4 | 52.9 |
| *Zea mays* | NC_007982 | 32 | 43.9 | 43.0 | 34.4 | 45.0 | 52.5 | 53.1 |
| *Zea perennis* | NC_008331 | 32 | 43.9 | 43.0 | 34.5 | 45.0 | 52.4 | 53.0 |

[a]Four-fold synonymous site data are based on a set of 16 genes shared by all species.

[b]Excludes *trans*-spliced introns

TABLE 2. Codon usage and the effect of RNA editing

| | | Genomic DNA | | Change after Editing | |
|---|---|---|---|---|---|
| | **Codon** | **Count** | **RSCU** | **Count** | **RSCU** |
| **Ala** | GCA | 133 | 0.94 | -1 | 0.00 |
| | GCC | 133 | 0.94 | -2 | 0.00 |
| | GCG | 68 | 0.48 | -2 | -0.01 |
| | GCU | 231 | 1.64 | 0 | 0.01 |
| **Arg** | AGA | 92 | 1.22 | 0 | 0.16 |
| | AGG | 55 | 0.73 | 0 | 0.10 |
| | CGA | 104 | 1.38 | 0 | 0.18 |
| | CGC | 45 | 0.60 | -9 | -0.06 |
| | CGG | 57 | 0.76 | -22 | -0.24 |
| | CGU | 98 | 1.30 | -20 | -0.13 |
| **Asn** | AAC | 78 | 0.61 | 0 | 0.00 |
| | AAU | 178 | 1.39 | 0 | 0.00 |
| **Asp** | GAC | 79 | 0.62 | -2 | -0.02 |
| | GAU | 177 | 1.38 | 2 | 0.02 |
| **Cys** | UGC | 40 | 0.68 | 8 | -0.02 |
| | UGU | 77 | 1.32 | 21 | 0.02 |
| **Gln** | CAA | 178 | 1.54 | 0 | 0.00 |
| | CAG | 53 | 0.46 | 0 | 0.00 |
| **Glu** | GAA | 214 | 1.35 | 0 | 0.00 |
| | GAG | 102 | 0.65 | 0 | 0.00 |
| **Gly** | GGA | 213 | 1.47 | 0 | 0.00 |
| | GGC | 70 | 0.48 | -1 | 0.00 |
| | GGG | 99 | 0.68 | 0 | 0.00 |
| | GGU | 198 | 1.37 | 1 | 0.00 |
| **His** | CAC | 40 | 0.40 | -5 | -0.02 |
| | CAU | 162 | 1.60 | -12 | 0.02 |
| **Ile** | AUA | 170 | 0.79 | 2 | 0.00 |
| | AUC | 179 | 0.83 | -7 | -0.04 |
| | AUU | 297 | 1.38 | 11 | 0.04 |
| **Leu** | CUA | 147 | 0.95 | 17 | -0.02 |
| | CUC | 90 | 0.58 | -3 | -0.09 |
| | CUG | 83 | 0.54 | 11 | -0.01 |
| | CUU | 194 | 1.26 | 13 | -0.08 |
| | UUA | 242 | 1.57 | 54 | 0.11 |
| | UUG | 168 | 1.09 | 40 | 0.09 |
| **Lys** | AAA | 179 | 1.24 | 0 | 0.00 |
| | AAG | 109 | 0.76 | 0 | 0.00 |
| **Met** | AUG | 234 | -- | 3 | -- |
| **Phe** | UUC | 243 | 0.82 | 9 | -0.05 |
| | UUU | 348 | 1.18 | 52 | 0.05 |
| **Pro** | CCA | 117 | 1.10 | -28 | -0.03 |
| | CCC | 80 | 0.75 | -10 | 0.09 |
| | CCG | 64 | 0.60 | -19 | -0.06 |
| | CCU | 164 | 1.54 | -35 | 0.01 |

| | | | | | |
|---|---|---|---|---|---|
| **Ser** | AGC | 81 | 0.68 | -1 | 0.10 |
| | AGU | 129 | 1.08 | 1 | 0.19 |
| | UCA | 135 | 1.13 | -43 | -0.23 |
| | UCC | 102 | 0.86 | -15 | -0.01 |
| | UCG | 98 | 0.82 | -32 | -0.17 |
| | UCU | 169 | 1.42 | -11 | 0.13 |
| **Thr** | ACA | 95 | 0.93 | -2 | 0.00 |
| | ACC | 113 | 1.11 | -4 | -0.02 |
| | ACG | 51 | 0.50 | -3 | -0.02 |
| | ACU | 149 | 1.46 | 0 | 0.03 |
| **Trp** | UGG | 135 | -- | 22 | -- |
| **Tyr** | UAC | 68 | 0.51 | 4 | 0.00 |
| | UAU | 197 | 1.49 | 13 | 0.00 |
| **Val** | GUA | 163 | 1.22 | 1 | -0.01 |
| | GUC | 104 | 0.78 | -3 | -0.03 |
| | GUG | 113 | 0.84 | 2 | 0.01 |
| | GUU | 156 | 1.16 | 5 | 0.03 |

FIGURES

**Figure 1.** Nucleotide frequencies at 4-fold synonymous sites in plant mtDNA. Each bar represents the average 3rd position nucleotide frequencies for a set of codons specifying the same amino acid (from left to right: Ala, Arg, Gly, Leu, Pro, Ser, Thr, Val). Error bars describe the full range of means across 18 different land plants. The data presented are based on the subset of 16 genes that were found in all 18 species. Nucleotide frequencies at 2-fold and 3-fold synonymous sites are available as supplementary data (Figure S1).

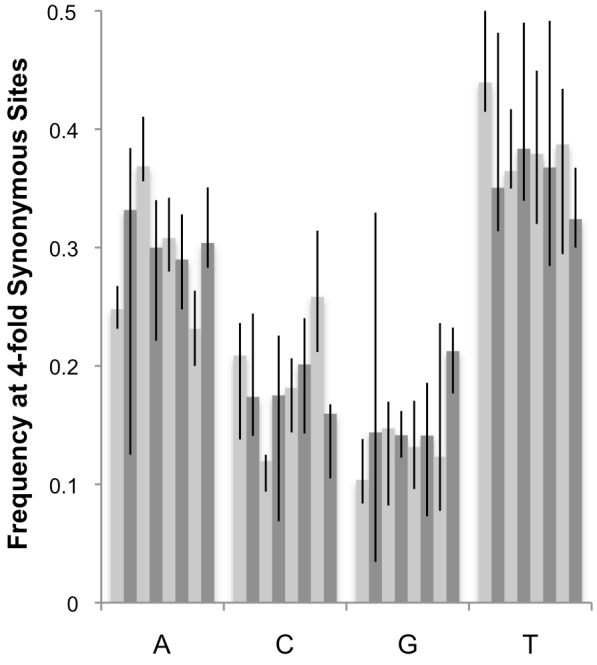


**Figure 2.** $N_c$ and $GC_{3S}$ data based on the mitochondrial genomes of vascular plants (black squares), bryophytes (black triangles), and metazoans (grey circles). The black line shows the expected $N_c$ as a function of $GC_{3S}$ given random codon usage (Eq. 1). Animal data points are labeled with initials corresponding to the following species: *Axinella corrugata, Briareum asbestinum, Caenorhabditis elegans, Colpopphyllia natans, Danio rerio, Drosophila melanogaster, Gallus gallus, Homo sapiens, Hydra oligactis, Mus musculus, Mytilus edulis, Tribolium castaneum, Trichoplax adhaerens*. Codon usage data for each plant species broken down by gene and by codon are provided as supplementary material (Tables S2 and S3).

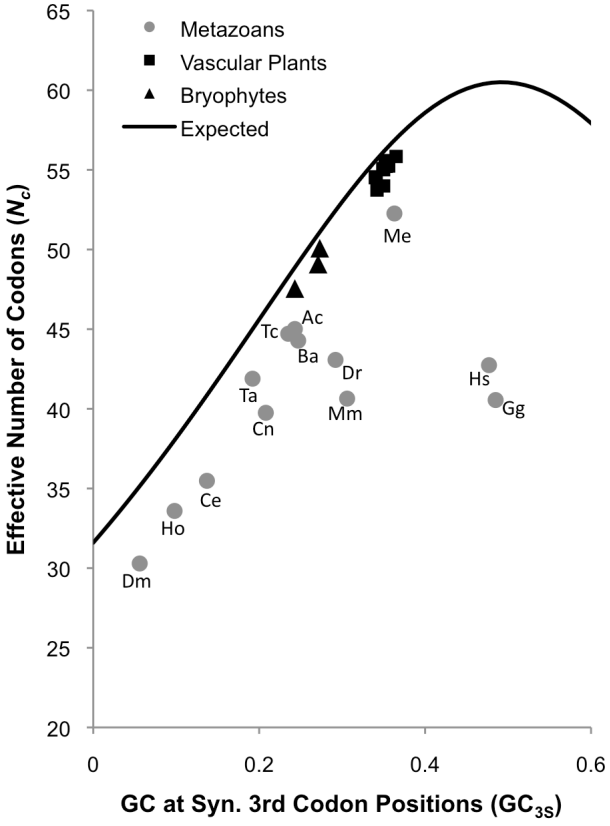
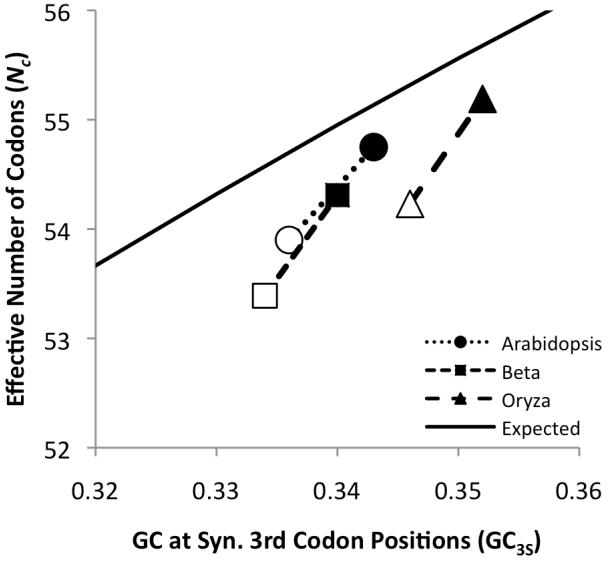


**Figure 3.** The effect of RNA editing on $N_c$ and $GC_{3S}$ based on whole mitochondrial genome datasets for three angiosperms. Data points connected by broken lines represent values before (filled symbols) and after (open symbols) RNA editing for each species. The solid black line shows the expected $N_c$ as a function of $GC_{3S}$ given random codon usage (Eq. 1).



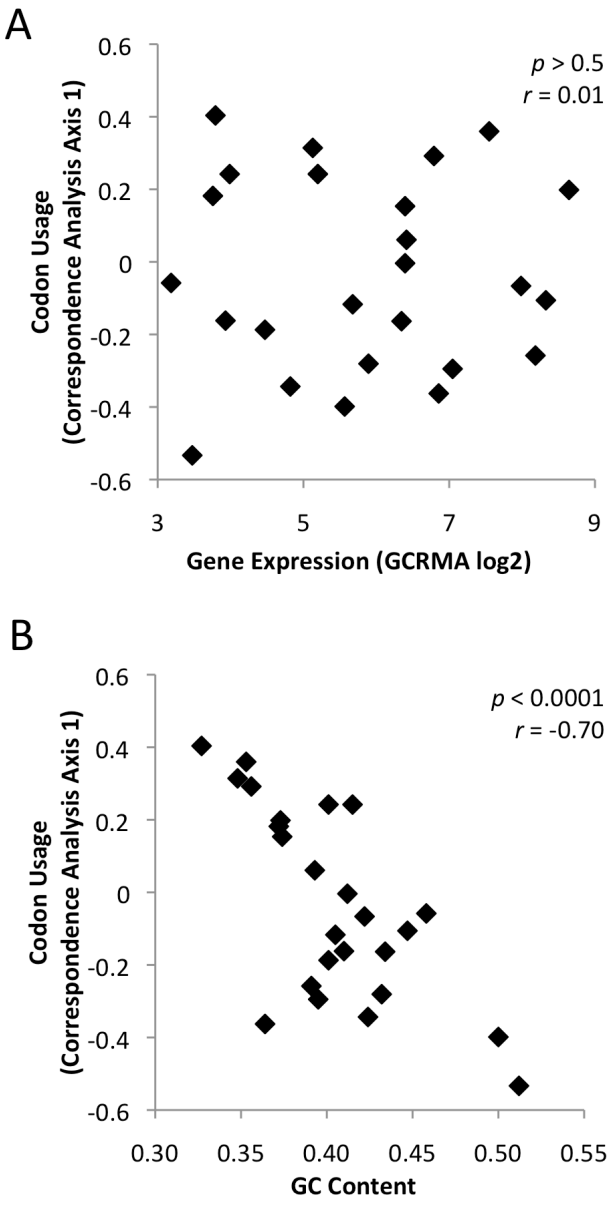
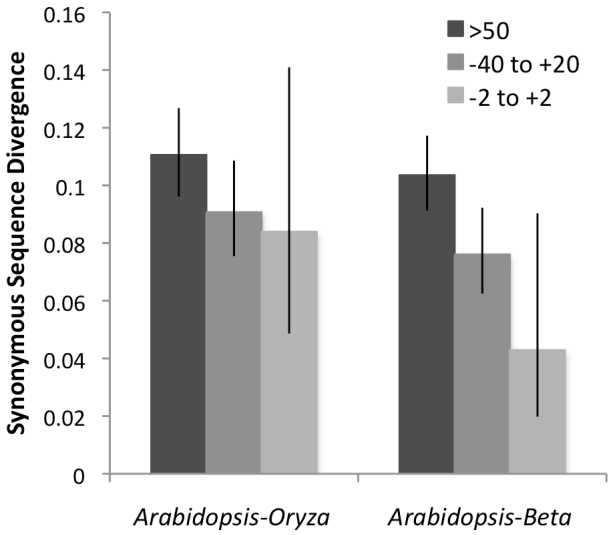



**Figure 4.** The relationship between codon usage and either gene expression (A) or GC content (B) in *Arabidopsis thaliana*. Each data point represents a gene, and codon usage is defined as the placement of that gene on the first axis of a correspondence analysis (see Methods). Gene expression values for the first pair of true leaves were taken from the AtGenExpress developmental dataset (see Methods). Expression values from other tissue types were also compared with codon usage and yielded similar results (data not shown). The second axis of the correspondence analysis was not significantly correlated with either gene expression or GC content (data not shown). Pearson correlation coefficients ($r$) are reported along with corresponding $p$-values.

**Figure 5.** Pairwise synonymous sequence divergence based on proximity to conserved RNA editing sites. The proportion of divergent 4-fold synonymous sites is reported for separate comparisons of *Arabidopsis* with both *Oryza* and *Beta* (see Methods for details). Values are reported separately for sites that are >50 bp away from the nearest editing site (dark bars) or within a -40 to +20 bp window (medium bars) or -2 to +2 bp window (light bars) from the nearest editing site. Errors bars represent 95% confidence intervals as calculated from the binomial distribution.



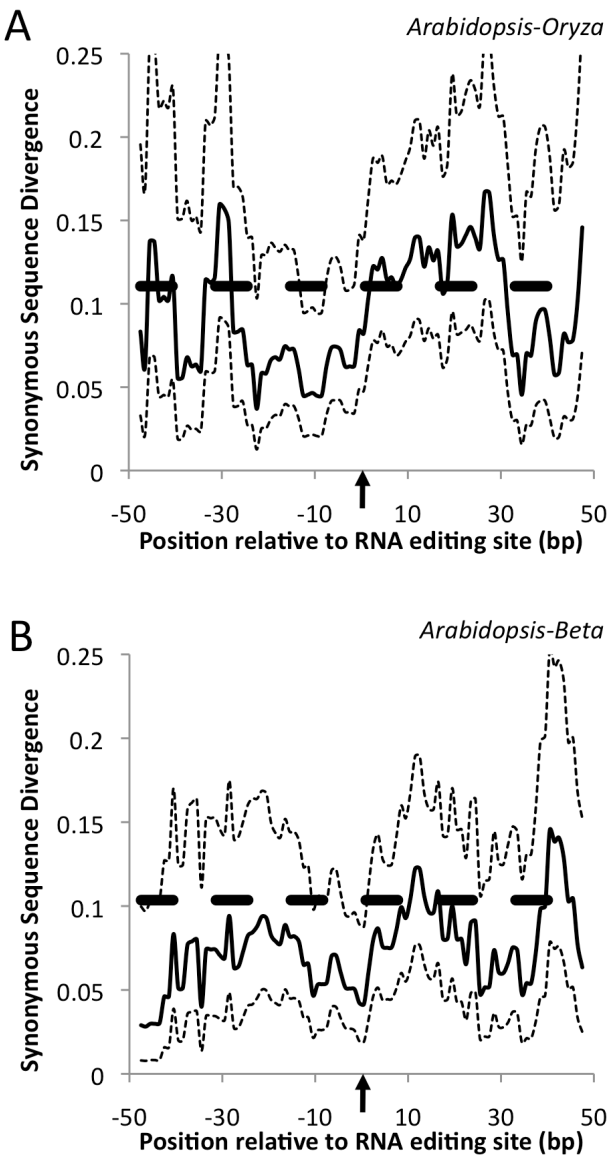



**Figure 6.** Sliding window analysis of synonymous divergence based on proximity to RNA editing sites. The solid line traces the proportion of divergent 4-fold synonymous sites based on a 6 bp window centered at that point on the x-axis, which represents the distance to the nearest conserved editing site conserved between *Arabidopsis* and either (A) *Oryza* or (B) *Beta*. The dotted lines show the bounds of the 95% confidence interval calculated from the binomial distribution. The dashed line shows the average substitution rate for 4-fold synonymous sites that are >50 bp from the nearest conserved editing site.

# Chapter 3.

# Evolutionary rate variation at multiple levels of biological organization in plant mitochondrial DNA[1]

[1]Formatted as a co-authored manuscript and published as:

Sloan DB, Barr CM, Olson MS, Keller SR, Taylor DR. 2008. *Mol. Biol. Evol.* 25(2): 243-246.

Referenced supplementary material is available online at:

http://mbe.oxfordjournals.org/content/25/2/243.full

ABSTRACT

We examined patterns of mitochondrial polymorphism and divergence in the angiosperm genus *Silene* and found substantial variation in evolutionary rates among species and among lineages within species. Moreover, we found corresponding differences in the amount of polymorphism within species. We argue that, along with our earlier findings of rate variation among genes, these patterns of rate heterogeneity at multiple phylogenetic scales are most likely explained by differences in underlying mutation rates. In contrast, no rate variation was detected in nuclear or chloroplast loci. We conclude that mutation rate heterogeneity is a characteristic of plant mitochondrial sequence evolution at multiple biological scales and may be a crucial determinant of how much polymorphism is maintained within species. These dramatic patterns of variation raise intriguing questions about the mechanisms driving and maintaining mutation rate heterogeneity in plant mitochondrial genomes. Additionally, they should alter our interpretation of many common phylogenetic and population genetic analyses.

MAIN TEXT

Classic molecular evolutionary studies have established a general pattern of low substitution rates in plant mitochondrial DNA (Wolfe, Li, Sharp 1987; Palmer and Herbon 1988). In recent years, however, exceptions to this pattern of slow plant mitochondrial sequence evolution have been found, highlighted by major rate accelerations in *Plantago* and *Pelargonium* (Cho et al. 2004; Parkinson et al. 2005). Given that the mechanistic basis remains elusive in these cases of elevated substitution rate, the full implications and extent of rate variation in plant mitochondrial genomes is unclear. For example, at what biological levels does rate variation occur? Does it only occur among species or is it found within species as well? Does it affect the levels of genetic variation (polymorphism) that are maintained within species? Here, we address these questions with an empirical analysis of DNA sequence polymorphism and divergence in *Silene*, expanding on earlier single-species studies in this genus (Städler and Delph 2002; Houliston and Olson 2006; Barr et al. 2007).

In comparing patterns of mitochondrial divergence, we found substantial rate variation among closely related species (fig. 1). Most notably, mitochondrial genes in *S. noctiflora* showed extreme divergence relative to the rest of the genus—a result that was recently discovered in an independent study of rate variation among species (Mower et al. 2007). Even after excluding *S. noctiflora*, rate variation among species was still evident; a likelihood ratio test found that a model of evolution allowing for rate heterogeneity among species provided a significantly better fit to the mitochondrial data than one that enforced a molecular clock ($\chi^2_{df=5} = 54.6$; $p < 0.0001$). Among the remaining species, substitution rates differed by more than 8-fold between the fastest (*S. paradoxa*) and

slowest (*S. latifolia*) evolving taxa. In contrast, substitution rates of two nuclear (*X4* and *ITS*) and two chloroplast loci (*trnL* and *rps16* introns) showed no evidence for variation among species ($\chi^2_{df=5} = 4.6$; $p = 0.47$).

Analysis of multiple individuals within *S. vulgaris* and *S. latifolia* found that patterns of rate heterogeneity extended across multiple phylogenetic scales. The concatenation of seven mitochondrial loci revealed extensive rate variation among the different lineages of *S. vulgaris* (fig. 2), and a molecular clock test strongly rejected homogeneous rates in this species ($\chi^2_{df=39} = 111.7$; $p < 0.0001$). The consequences of intraspecific rate variation were startling. For example, slowly evolving lineages in both species have not experienced a single substitution in the highly polymorphic *atp1* locus so that both species retain the same common ancestral haplotype (as identified by haplotype reconstruction in PAML). At the other extreme, rapidly evolving lineages within *S. vulgaris* have accumulated as many as 16 substitutions in this gene. In contrast to *S. vulgaris*, no rate variation was detected within the largely invariant *S. latifolia* sample ($\chi^2_{df=27} = 30.3$; $p = 0.30$). Findings of rate heterogeneity both within and among species were consistently supported with pairwise relative rate tests (data not shown; Tajima 1993), which are not sensitive to recombination or phylogenetic uncertainty (Posada 2001).

Challenges to the concept of a molecular clock in which the accumulation of substitutions occurs at a constant rate over time have arisen since its inception (Zuckerkandl and Pauling 1965), and violations of the clock are now well established including examples in plant mitochondrial DNA (Britten 1986; Martin and Palumbi 1993, Eyre-Walker and Gaut 1997; Whittle and Johnston 2002; Mower et al. 2007). The present

study and previous work showing rate variation among genes in *S. vulgaris* (Barr et al. 2007) demonstrate not only that rate variation occurs, but that it extends broadly across multiple biological scales.

As seen in other examples of rate acceleration in plant mitochondria (Cho et al. 2004; Parkinson et al. 2005), rate heterogeneity within *Silene* predominates at synonymous sites (data not shown), suggesting that differences in mutation rate are the underlying cause (Kimura 1983). The existence of rate heterogeneity both within and among species raises important and as yet unanswered questions about the relationship of mutation rate variability at different scales. Is the variation in substitution rate among *Silene* species caused by the same mechanisms that produce rate variation among genes or among lineages within *S. vulgaris*? Is the extreme divergence of *S. noctiflora* simply a case of the more modest interspecific variation writ large, or is it the result of an entirely different mechanism? How does selection act on mutation rate variation? Identifying the mechanistic basis of mutation rate variation at these different scales is an important goal to resolve these questions and thereby better understand the processes that drive sequence evolution in the plant mitochondrial genome.

Given mutation rate variation among species, neutral theory would predict differences in intraspecific polymorphism. Comparison of 7 mitochondrial loci found substantially greater polymorphism in *S. vulgaris* than *S. latifolia* (table 1). A maximum likelihood model of the neutral coalescent process produced an estimate of the scaled mutation rate ($\theta$) based on synonymous segregating sites that was more than 5-fold higher for *S. vulgaris* than for *S. latifolia*. For each species-specific estimate of $\theta$, the approximate 95% confidence intervals do not overlap the corresponding maximum

likelihood estimate of the other species (fig. 3); therefore, we consider these differences in polymorphism significant. In contrast, analysis of a single locus in both the nuclear and chloroplast genomes found that polymorphism in *S. latifolia* was equal to or greater than in *S. vulgaris*, consistent with previous studies of other loci in these genomes (Ingvarsson and Taylor 2002; Taylor and Keller 2007; but see Ingvarsson 2004).

At least a portion of the elevated mitochondrial polymorphism in *S. vulgaris* appears to result from accelerated mutation in a subset of lineages within the species. A local clock comparison in PAML (Yoder and Yang 2002) found the substitution rate in *S. vulgaris* to be 2-fold higher than in *S. latifolia*, but this difference was not significant. A model allowing for separate rate estimates for each species offered only a marginal improvement over one that enforced a single rate ($\chi^2_{df=1} = 3.1$; $p = 0.08$). This species-wide analysis, however, obscures the rate variation among *S. vulgaris* lineages. Relative rate tests found that the most rapidly evolving *S. vulgaris* lineage has a significantly elevated rate compared to *S. latifolia* even after considering the multiple comparisons based on 40 *S. vulgaris* samples ($\chi^2_{df=1} = 11.7$; $p = 0.0006$). At the other extreme, the most slowly-evolving *S. vulgaris* lineages were indistinguishable in rate from *S. latifolia* ($\chi^2_{df=1} = 0.3$; $p = 0.57$).

Previous studies in *Silene* have interpreted patterns of polymorphism as possible evidence for balancing selection or selective sweeps acting on mitochondrial genomes (Ingvarsson and Taylor 2002; Städler and Delph 2002; Houliston and Olson 2006). While historical patterns of selection in these species may contribute to the difference in mitochondrial polymorphism, these interpretations are generally made under an assumption of a constant mutation rate. Our findings of extensive mitochondrial rate

variation suggest that the role of mutation should be taken into account even when comparing closely related species. The existence of such striking variation at multiple biological scales has far-reaching impacts for phylogenetic and population genetic analyses which commonly disregard mutation rate variation at least at some levels. Meanwhile, these patterns of rate variation suggest that the plant mitochondrial genome and the genus *Silene* in particular may be fertile ground for examining the evolutionary causes and consequences of mutation rates.

## METHODS

We extracted DNA from a single individual from each of 40 and 28 geographically dispersed populations of *S. vulgaris* and *S. latifolia*, respectively, as well as from a single individual from 6 related species (supplementary material online). As described previously (Barr et al. 2007), we PCR amplified and sequenced 9 loci representing all 3 genomes (table 1; supplementary material online). For the nuclear *X4* gene, heterozygosities were scored manually in Sequencher v4.5, and multiple clones from each PCR product were sequenced to exclude paralogous Y-linked copies in *S. latifolia* males (Invitrogen TOPO TA Cloning Kit).

We calculated the number of segregating sites (S) and nucleotide diversity ($\pi$) for each gene in *S. latifolia* and *S. vulgaris* with DnaSP v4.0 (Rozas et al. 2003). To assess the statistical significance of the difference in mitochondrial polymorphism between the two species, we used neutral coalescent estimates of the scaled mutation rate ($\theta$) as described previously (Hudson 1991; Barr et al. 2007).

We performed molecular clock tests and branch length analysis with BASEML within PAML v3.15 (Yang 1997). Intraspecific tree topologies were determined by neighbor-joining in MEGA v3.1 (Kumar, Tamura, Nei 2004), while interspecific topologies were constrained by a supertree analysis (supplementary material online). Models of substitution were chosen with ModelTest v3.7 (Posada and Crandall 1998).

## SUPPLEMENTARY MATERIAL

Supplementary files containing geographic information, PCR primer sequences, GenBank accession numbers, sequence alignments and phylogenetic trees are available at *Molecular Biology and Evolution* online at http://www.mbe.oxfordjournals.org/.

## ACKNOWLEDGEMENTS

We greatly appreciate the assistance and comments of Rachel Carr, Whit Farnum, Michael Hood, Ellen McRae, Janet Miller, Keiko Miyake, Maurine Neiman, Dexter Sowell, three anonymous reviewers and all those who have contributed to the *Silene* collection. We also would like to thank Jeffrey Palmer for sharing his manuscript prior to publication. This work was supported by NSF grants to DRT (DEB-0349558) and MSO (DEB-0317115).

## LITERATURE CITED

Barr CM, Keller SR, Ingvarsson PK, Sloan DB, Taylor DR. 2007. Variation in mutation rate and polymorphism among mitochondrial genes in *Silene vulgaris*. Mol. Biol. Evol. 24: 1783-1791.

Britten RJ. 1986. Rates of DNA sequence evolution differ between taxonomic groups. Science 231:1393-1398.

Cho Y, Mower JP, Qiu YL, Palmer JD. 2004. Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. Proc Natl Acad Sci 101:17741-17746.

Eyre-Walker A, Gaut BS. 1997. Correlated rates of synonymous site evolution across plant genomes. Mol. Biol. Evol. 14:455-460.

Houliston GJ, Olson MS. 2006. Nonneutral evolution of organelle genes in *Silene vulgaris*. Genetics 174:1983-1994.

Hudson RR. 1991. Gene genealogies and the coalescent process. In: Futuyma D, Antonovics J, editors. Oxford Surveys of Evolutionary Biology, vol. 7. Oxford: Oxford University Press. p. 1-44.

Ingvarsson PK. 2004. Population subdivision and the hudson-kreitman-aguade test: Testing for deviations from the neutral model in organelle genomes. Genet. Res. 83:31-39.

Ingvarsson PK, Taylor DR. 2002. Genealogical evidence for epidemics of selfish genes. Proc Natl Acad Sci 99:11265-11269.

Kimura M. 1983. *The neutral theory of molecular evolution*. Cambridge: Cambridge University Press.

Kumar S, Tamura K, Nei M. 2004. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Briefings in Bioinformatics 5:150-163.

Martin AP, Palumbi SR. 1993. Body size, metabolic rate, generation time, and the molecular clock. Proc Natl Acad Sci 90:4087-4091.

Mower JP, Touzet P, Gummow JS, Delph JS, Palmer JD. 2007. Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. BMC Evol. Biol. 7:135.

Palmer JD, Herbon LA. 1988. Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J. Mol. Evol. 28:87-97.

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

Posada D. 2001. Unveiling the molecular clock in the presence of recombination. Mol. Biol. Evol. 18:1976-1978.

Posada D and Crandall KA. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14:817-818.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496-2497.

Städler T, Delph LF. 2002. Ancient mitochondrial haplotypes and evidence for intragenic recombination in a gynodioecious plant. Proc. Natl. Acad. Sci. U. S. A. 99:11730-11735.

Tajima F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135:599-607.

Taylor DR, Keller SR. 2007. Historical range expansion determines the phylogenetic diversity introduced during contemporary species invasion. Evolution 61:334-345.

Whittle CA, Johnston MO. 2002. Male-driven evolution of mitochondrial and chloroplastidial DNA sequences in plants. Mol. Biol. Evol. 19:938-949.

Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear dnas. Proc. Natl. Acad. Sci. 84:9054-9058.

Yang Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. Bioinformatics 13:555.

Yoder AD, Yang Z. 2000. Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. 17:1081-1090.

Zuckerkandl E, Pauling L. 1965. Molecules as documents of evolutionary history. J. Theor. Biol. 8:357-366.

**Table 1. Polymorphism statistics for *S. vulgaris* and *S. latifolia* by gene**

S = Segregating sites. π = nucleotide diversity. Values in parentheses reflect polymorphism at synonymous sites.

| | *S. vulgaris* (n = 40) | | *S. latifolia* ( n= 28) | |
|---|---|---|---|---|
| **Mitochondrial** | S | π (%) | S | π (%) |
| *atp1* (970 bp) | 24 (19) | 0.65 (2.32) | 4 (2) | 0.08 (.27) |
| *atp4* (291 bp) | 0 (0) | 0 (0) | 1 (0) | 0.18 (0) |
| *atp6* (696 bp) | 14 (6) | 0.21 (0.46) | 1 (1) | 0.02 (0.08) |
| *cob* (919 bp) | 8 (4) | 0.20 (0.55) | 1 (0) | 0.01 (0) |
| *cox3* (602 bp) | 6 (0) | 0.22 (0) | 4 (1) | 0.05 (0.05) |
| *nad9* (393 bp) | 1 (0) | 0.09 (0) | 4 (1) | 0.20 (0.59) |
| *nad4L-atp4* (134 bp) | 4 | 0.32 | 1 | 0.05 |
| **Total (4005 bp)** | 57 (33) | 0.29 (0.75) | 16 (6) | 0.07 (0.15) |
| **Chloroplast** | | | | |
| *trnL* (504 bp) | 6 | 0.06 | 7 | 0.12 |
| **Nuclear** | | | | |
| *X4* (578 bp) | 57 (36) | 1.16 (3.71) | 51 (37) | 2.65 (8.09) |

FIGURES

Figure 1. (a) Mitochondrial and (b) nuclear/cholorplast trees with branch length estimates (substitutions/site) performed in PAML (BASEML).

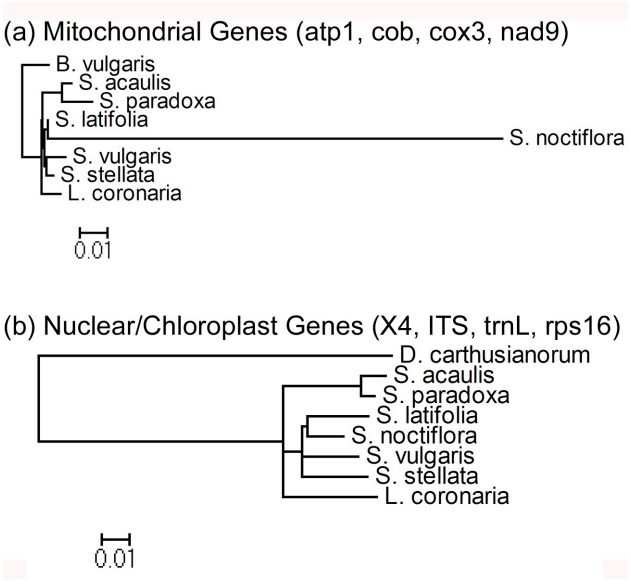


Figure 2. Mitochondrial tree for all *S. vulgaris* and *S. latifolia* individuals based on 7 concatenated loci with branch length estimates (substitutions/site) performed in PAML (BASEML).

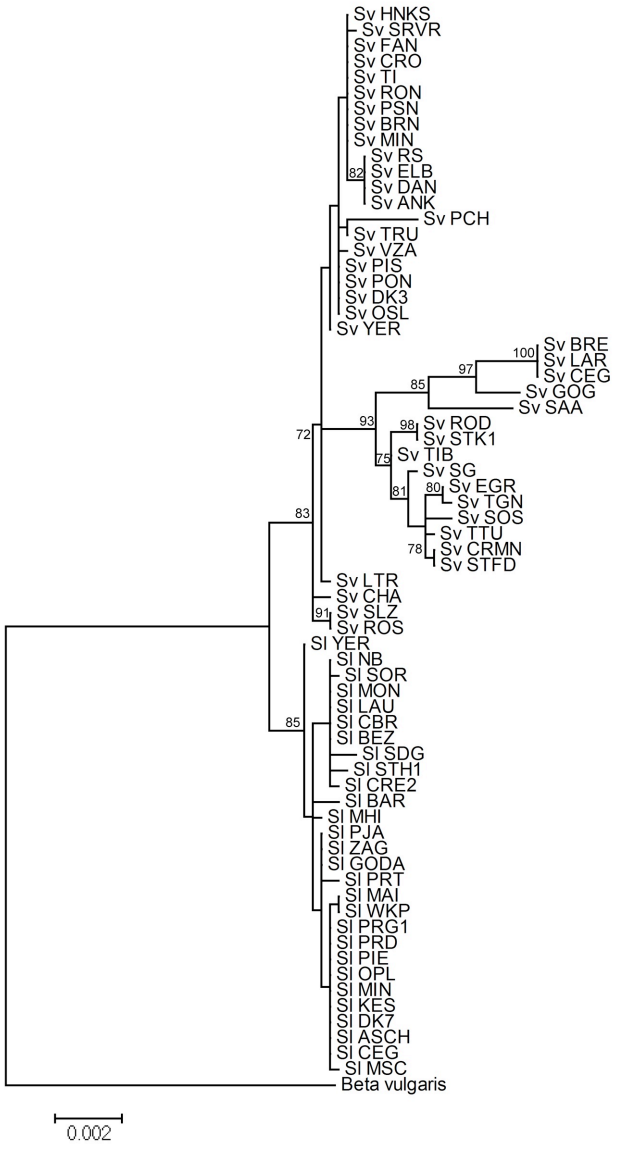
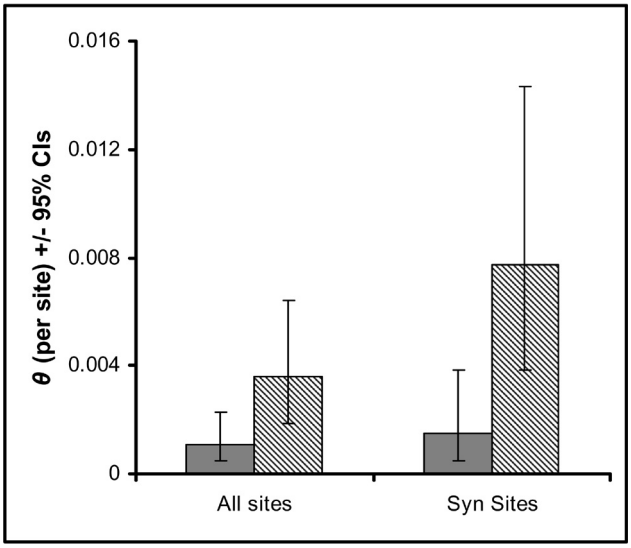


Figure 3. Estimates of the scaled mutation rate ($\theta$) based on the number of segregating sites in the mitochondrial genes of *S. latifolia* (solid) and *S. vulgaris* (striped).



# Chapter 4.

# A phylogenetic analysis of mitochondrial substitution rate variation in the angiosperm tribe *Sileneae* (Caryophyllaceae)[1]

[1]Formatted as a co-authored manuscript and published as:

Sloan DB, Oxelman B, Rautenberg A, Taylor DR. 2009. *BMC Evol. Biol.* 9:260.

Referenced supplementary material is available online at:

http://www.biomedcentral.com/1471-2148/9/260

# Abstract


## Background

Recent phylogenetic studies have revealed that the mitochondrial genome of the angiosperm *Silene noctiflora* (Caryophyllaceae) has experienced a massive mutation-driven acceleration in substitution rate, placing it among the fastest evolving eukaryotic genomes ever identified. To date, it appears that other species within *Silene* have maintained more typical substitution rates, suggesting that the acceleration in *S. noctiflora* is a recent and isolated evolutionary event. This assessment, however, is based on a very limited sampling of taxa within this diverse genus.

## Results

We analyzed the substitution rates in 4 mitochondrial genes (*atp1*, *atp9*, *cox3* and *nad9*) across a broad sample of 74 species within *Silene* and related genera in the tribe *Sileneae*. We found that *S. noctiflora* shares its history of elevated mitochondrial substitution rate with the closely related species *S. turkestanica*. Another section of the genus (*Conoimorpha*) has experienced an acceleration of comparable magnitude. The phylogenetic data remain ambiguous as to whether the accelerations in these two clades represent independent evolutionary events or a single ancestral change. Rate variation among genes was equally dramatic. Most of the genus exhibited elevated rates for *atp9* such that the average tree-wide substitution rate for this gene approached the values for the fastest evolving branches in the other three genes. In addition, some species exhibited major accelerations in *atp1* and/or *cox3* with no correlated change in other genes. Rates of non-synonymous substitution did not

increase proportionally with synonymous rates but instead remained low and relatively invariant.

**Conclusions**

The patterns of phylogenetic divergence within *Sileneae* suggest enormous variability in plant mitochondrial mutation rates and reveal a complex interaction of gene and species effects. The variation in rates across genomic and phylogenetic scales raises questions about the mechanisms responsible for the evolution of mutation rates in plant mitochondrial genomes.

## Background

Substitution rates in plant mitochondrial genomes are generally low relative to their nuclear and chloroplast counterparts, as well as relative to the mitochondrial genomes of other organisms [1-3]. In fact, absolute rates of sequence evolution in seed plant mitochondrial DNA (mtDNA) are among the slowest ever estimated (Figure 1; [4]). A series of recent studies, however, has revealed notable exceptions to this generalization [4-7]. There are angiosperm species that not only deviate from the slow substitution rates typical of plant mtDNA but also exhibit some of the highest eukaryotic substitution rates ever documented (Figure 1). With such a substantial fraction of known rate variation captured in a relatively small twig within the tree of life, plant mitochondrial genomes represent an intriguing system for investigating the evolutionary forces that shape substitution rates [8-14].

Studies of rate accelerations in plant mitochondrial genomes have consistently shown that these effects are most pronounced at so-called synonymous sites, which do not affect the corresponding amino acid sequence (*e.g.* [5]). One of the pillars of the neutral theory of molecular evolution is that the rate of neutral substitutions (*i.e.* those with no fitness effect) is expected to equal the mutation rate [15]. Synonymous substitutions are not completely neutral, however. They are subject to a variety of selection pressures including translational efficiency, mRNA stability and the conservation of regulatory motifs (reviewed in [16]), and direct measurements of mutation rates can be more than an order of magnitude higher than those estimated from synonymous substitution rates [17]. Nevertheless, synonymous sites still offer one of our best approximations of the underlying mutation rate. Therefore,

considering the absence of well-supported alternative hypotheses, the extreme synonymous substitution rates observed in certain plant mitochondrial genomes are most likely a result of mutational acceleration.

*Silene noctiflora* (Caryophyllaceae) is a recent addition to a growing list of angiosperms exhibiting major accelerations in mitochondrial synonymous substitution rate [4, 7]. In other well-documented examples (*e.g. Plantago* and *Pelargonium*), rate accelerations appear relatively old (*ca.* 30-80 million years) having preceded the divergence of large clades or even an entire genus [6]. In contrast, the extreme mitochondrial substitution rates of *S. noctiflora* appear unique relative to other *Silene* species, suggesting a very recent acceleration. Estimates of mitochondrial substitution rate, however, are available for only a few *Silene* species, representing a tiny fraction of this large and diverse genus. The sparse sampling severely limits the phylogenetic resolution to detect historical changes in substitution rate.

The scarcity of mitochondrial sequence data within *Silene* reflects a broader under-representation of plant mtDNA in studies of molecular evolution. Whereas chloroplast DNA (cpDNA) and animal mtDNA are utilized extensively in phylogenetic studies, the low baseline substitution rates and growing evidence for rate heterogeneity in plant mtDNA often limit its utility in this context—particularly at local phylogenetic scales [18]. Understanding the causes and consequences of mutation rate variation is a fundamental problem in evolutionary biology [12, 19-21], but the lack of plant mtDNA sequence data is a hindrance to investigating this question. To characterize the pattern of mitochondrial substitution rate variation throughout *Silene* and related genera, we sequenced four mitochondrial loci in a sample of 74 species that were selected to capture the phylogenetic diversity of this

genus and its closest relatives (Table 1). To our knowledge, this effort represents the most extensive species-level sampling to date of mitochondrial sequence divergence in a plant genus.

To compare absolute substitution rates in a gene across lineages requires an estimate of the genealogy with dated nodes (*i.e.* divergence times). In cases of extreme rate variation, generating such a tree directly from the gene in question is problematic. With rate variation, slowly-evolving taxa can be difficult to resolve, and long branch attraction can favor incorrect topologies [22]. Even with an accurate topology, rate variation can bias the estimate of divergence times with molecular clock based methods. For this reason, previous studies of substitution rate variation in plant mitochondrial genomes have constrained their analyses based on phylogenies and divergence times inferred from nuclear and chloroplasts sequences.

Because both mitochondrial and chloroplast genomes are predominantly maternally inherited in *Silene*, they are expected to share a common genealogy [23, 24] (although breakdowns in uniparental inheritance may potentially disrupt this relationship [25-27]). Therefore, we chose the chloroplast gene *matK* to estimate phylogenetic relationships and divergence times. This gene has proven to be highly informative in phylogenetic reconstruction, partly because of its high rates of substitution [28, 29]. It has also been used in two recent analyses of divergence times within *Silene* and the Caryophyllaceae [4, 30].

We identified substantial rate accelerations in multiple lineages within the *Silene* phylogeny as well as major rate differences among mitochondrial genes. Here, we discuss the complex patterns of mitochondrial rate variation in the genus *Silene*

and the implications they have for the evolution of mitochondrial mutation rates and the patterns of selection on mtDNA at the sequence level.

# Results

### Chloroplast DNA phylogeny

Likelihood, parsimony and Bayesian phylogenetic methods were in general agreement for the *matK* dataset. The 70% parsimony bootstrap consensus tree (Figure 2) did not conflict with any of the nodes from either the ML or Bayesian analysis. The results were also generally consistent with previous cpDNA studies of the tribe *Sileneae* [31, 32]. The analysis recovered the two previously identified subgenera (*e.g.* [32])—*Silene* and *Behenantha* (Otth.) Endl. (=subgenus *Behen* (Dumort.) Rohrb.)—along with the relationships among the major clades in subgenus *Silene*. There was, however, incomplete resolution in some parts of the tree—particularly among the major lineages within subgenus *Behenantha*, which appear as a large radiation. Four *Silene* species were not grouped with either of the two major subgenera. As found in the analysis of other chloroplast loci, *S. sordida* was placed in a clade with *Lychnis* [32]. *Silene odontopetala* was also assigned to this clade with strong support. The relationships between subgenus *Behenantha*, subgenus *Silene*, the *Lychnis*/*S. odontopetala*/*S. sordida* clade, and a fourth lineage consisting solely of *S. cordifolia* could not be confidently resolved. Finally, there was unexpected support for a sister relationship between *S. delicatula* and the rest of our *Silene*/*Lychnis* sample.

**Divergence times**

We used three different dating methods, which produced roughly similar estimates of divergence times, but there was a consistent pattern distinguishing them [see Additional files 1 and 2]. Specifically, the Langley-Fitch method produced the youngest estimates of divergence times within *Sileneae*, while a penalized likelihood method produced the oldest. For example, the estimated age of the root node for the entire *Silene*/*Lychnis* clade differed by 50% between the two methods (21.0 vs. 14.0 Myr). The BEAST model (Figure 2) generally produced intermediate estimates of divergence time relative to the other two methods. Only the BEAST values were used for subsequent rate analyses, so the uncertainty in divergence times should be considered when interpreting absolute substitution rate estimates.

**Mitochondrial rate variation**

Branch lengths in terms of both synonymous ($d_S$) and non-synonymous ($d_N$) substitutions per site for each mitochondrial gene are shown in Figure 3. All four genes show little divergence at non-synonymous sites across the entire tree (Table 2). In addition, they all share a pattern of extreme synonymous divergence in six *Silene* species that can be divided into two clear clades: (1) the previously characterized *S. noctiflora* along with its close relative *S. turkestanica* and (2) *S. ammophila*, *S. conica*, *S. conoidea*, and *S. macrodonta*, which all belong to section *Conoimorpha*. Beyond those similarities, the four mitochondrial genes differ markedly in synonymous branch lengths (Figure 3). Very little divergence is observed in *nad9* outside of the aforementioned six species. Synonymous divergence is similarly low throughout much of the *cox3* and *atp1* trees, but there are a number of species that exhibit substantial divergence, particularly within subgenus *Silene*. This group

includes *S. nutans* which, despite showing no sign of abnormal divergence in *cox3* and *nad9*, is highly divergent for *atp1*. Finally, synonymous divergence in *atp9* is extreme and highly variable throughout most of the genus *Silene*, although many of the outgroup genera exhibit typically low levels of divergence. The total synonymous tree length is approximately 9-fold larger for *atp9* than the slowly-evolving *nad9*. This gap widens to 41-fold if the six taxa that have accelerated rates across all genes are excluded from the analysis (Table 2).

As expected given the enormous variation in mitochondrial divergence across species, absolute synonymous substitution rates ($R_S$) differ dramatically throughout the tribe *Sileneae* (Figure 4, [see Additional files 3 and 4]). The outgroups to *Silene* tend to have $R_S$ values of less than 0.5 substitutions per site per billion years (SSB), and certain branches have an estimated $R_S$ of 0 because they lack a single synonymous substitution (Figure 4). Many lineages within *Silene* have maintained these low rates. At the other extreme, the rapidly-evolving *Silene* lineages have $R_S$ values that are more than two orders of magnitude greater than the low rates of *Beta vulgaris* and other outgroups. The fastest rate estimates observed in the entire dataset were found in the *atp9* tree. The internal branch subtending the minimally inclusive clade that contains *S. succulenta* and *S. imbricata* had an estimated $R_S$ value of 392 SSB. The fastest terminal branch in the *atp9* tree was that of *S. schafta* with a rate of 292 SSB, although it should be noted that the error associated with *atp9* rate estimates for individual branches was generally large [see Additional file 4]. $R_S$ and $R_N$ values were both positively correlated across *atp1*, *cox3* and *nad9*, but this correlation broke down in comparisons with *atp9* and *matK* (Table 3). In addition, $R_S$ and $R_N$ values

were significantly correlated with each other within genes for *atp1*, *cox3* and *nad9* but not for the other two loci.

**Evolutionary congruence between mitochondrial and chloroplast genomes**

We utilized a constrained topology derived from cpDNA to analyze evolutionary rates in mtDNA, reflecting the assumption that the two organelle genomes share a single genealogy. Although the mitochondrial genes often yielded limited phylogenetic signal because of the dual problems of low variation and long branch attraction, there was some evidence to support phylogenetic congruence between these genomes [see Additional file 5]. For example, *S. hookeri* and *S. menziesii* were consistently paired by both chloroplast and mitochondrial genes, suggesting that the allopolyploid *S. hookeri* inherited both of its cytoplasmic genomes from the *S. menziesii* parental lineage [33]. In addition, the rapidly evolving *atp9* gene produced a tree that generally agreed with the chloroplast *matK* topology at younger nodes, which are presumably less susceptible to saturation at synonymous sites.

There were a large number of incongruencies between chloroplast and mitochondrial trees, but they were generally lacking in support. Perhaps the most suspicious example of conflict between mitochondrial and chloroplast topologies was the placement of *S. samojedora*, *S. seoulensis*, *S. zawadzkii* and the major accelerated species from subgenus *Behenantha* in a clade otherwise populated by subgenus *Silene* in the *cox3* tree [see Additional file 5]. Although this clade was supported by 74% of bootstrap replicates, inspection of the alignments showed that the grouping was based entirely on a single 6 bp region with 5 substitutions, raising doubts about the independence of those characters. Overall, we found no overwhelming evidence of conflicts between mitochondrial and chloroplast topologies, but the lack of

mitochondrial divergence in many lineages gave us little statistical power. Therefore, it is possible that topological inaccuracies in our constraint tree could have led to misidentification of small mitochondrial rate accelerations, but given the phylogenetic scale of our analysis, it is unlikely that any of the major rate changes was an artifact of topological conflicts.

To date, studies of angiosperms with major increases in plant mitochondrial substitution rates (*e.g. Plantago*, *Pelargonium* and *Silene*) have concluded that the observed accelerations are largely independent of evolutionary rates in the chloroplast or nuclear genomes (although there is growing evidence for accelerated sequence and structural evolution in the chloroplast genomes of species with high mitochondrial substitution rates [34-36]). In our dataset, we found little substitution rate variation among species for the chloroplast *matK* gene and no significant correlation between mitochondrial and chloroplast rates (Table 3).

# Discussion

### Mutation rate variation among species

*Silene noctiflora* has been shown to have dramatically accelerated rates of mitochondrial evolution relative to its congeners [4, 7]. We examined the phylogenetic distribution of this rate acceleration within the tribe *Sileneae* and identified six *Silene* species grouped into two clades that exhibited major increases in synonymous substitution rates across all four loci examined (Figure 3). As an illustration of the magnitude of these accelerations, we note the average synonymous pairwise divergence between these two closely-related clades within *Silene* subgenus *Behenantha* exceeds the divergence typically observed between flowering plants and

liverworts—the deepest split in the land plant phylogeny [37]. Based on the currently available data in seed plants, the synonymous substitution rates exhibited by these rapidly-evolving lineages (Figure 4) are exceeded only by the fastest lineages of *Plantago* (Figure 1) [5]. In addition, the observed rates are on par with average estimates for mammalian mtDNA, although they still fall well below the fastest mammalian rates [38]. As discussed above (see Background), the observed differences in synonymous substitution rates most likely reflect differences in the underlying mutation rate.

The phylogenetic data remain ambiguous with respect to whether the two clades with rate accelerations represent independent evolutionary events. The *matK* tree does not strongly support or reject a monophyletic relationship between *S. noctiflora*/*S. turkestanica* and section *Conoimorpha* (Figure 2; [see Additional files 6 and 7]). More thorough phylogenetic analyses of these taxa have recently been conducted, utilizing both chloroplast and nuclear loci [39]. These studies have found that, while cpDNA sequences suggest phylogenetic independence between the two clades, at least some nuclear loci support monophyly. Therefore it is possible but inconclusive that both high rate clades are sister taxa that inherited an accelerated mitochondrial substitution rate from a common ancestor. If so, the two clades must have split shortly after that acceleration, because internal branches shared by the two lineages in the mitochondrial gene trees are quite short relative to the divergence between the lineages [see Additional file 5]. Resolving these phylogenetic relationships could prove difficult because previous studies have shown that the evolutionary history of subgenus *Behenantha* may be complicated by reticulation [32, 40], such that relationships differ across genes and genomes

Comparisons of mitochondrial sequences from multiple populations of *S. noctiflora* have revealed very low levels of polymorphism, suggesting that the historically high mutation rates in this lineage may have undergone a reversion to more typical levels ([41] and unpublished data). This conclusion was, at least partially, supported by our phylogenetic data. The terminal branches for *S. noctiflora* and *S. turkestanica* exhibited a marked reduction in $R_S$ values relative to the ancestral rate for that clade (Figure 4). In contrast, the patterns of divergence within section *Conoimorpha* gave little indication of rate reversions.

The genus *Silene* is characterized by great diversity in breeding system and life history, and there has been substantial interest in how these traits may be related to molecular evolution in mitochondrial genomes [14, 26, 41-44]. There is no clear correlation between breeding system/life history and rate acceleration. The species exhibiting rate acceleration across all four mitochondrial genes are all hermaphroditic/gynomonoecious annuals with the exception of *S. turkestanica*, which is perennial. However, there are at least ten additional annual lineages represented in our sampling, and breeding system (hermaphroditic/gynomonoecious or gynodioecious) has yet to be determined for most species.

**Mutation rate variation among genes**

Substitution rates commonly differ among regions within a genome because of variation in selection and/or mutational pressure, and a previous study had already identified substantial rate heterogeneity among *Silene* mitochondrial genes [45]. Nevertheless, the differences in synonymous substitution rates among mitochondrial genes in the current study are surprisingly large. If the six species that show universal acceleration across all four mitochondrial genes are excluded, *atp9* appears to be

evolving more than 40 times faster than *nad9* at synonymous sites, while *cox3* and *atp1* fall in between these extremes.

The extreme elevation in *atp9* substitution rates calls into question whether a biological mechanism other than an increase in the mutation rate might be responsible. The obvious alternatives to explain high levels of divergence include horizontal gene transfer (HGT) from distantly related species [46], maintenance of ancient, trans-specific polymorphism by balancing selection [26, 41, 44, 47], re-localization of the gene to the higher mutation rate environment of the nuclear genome [48], or relaxed selection in a non-functional pseudogene [49].

None of these explanations, however, are fully consistent with the data. To explain the observed levels of divergence based on HGT without an increase in evolutionary rates would require multiple phylogenetically distant donor species (*i.e.* outside the angiosperms). Phylogenetic analysis of *atp9*, however, clearly places these sequences within the Caryophyllaceae ([Additional file 5] and unpublished data; note that this argument also applies to the lineage-specific divergence in *S. noctiflora*/*S. turkestanica* and section *Conoimorpha*). Likewise, in the absence of rate acceleration, an explanation based on balancing selection alone would require that polymorphism be maintained for hundreds of millions of years. Such a model seems extremely unlikely and even still could not explain the retention of partial phylogenetic congruence between *atp9* and *matK*. Of course, the fact that balancing selection alone cannot explain the pattern of divergence in *atp9* does not rule out the possibility that balancing selection has been acting on *atp9* and other mitochondrial genes in *Silene*.

It is unlikely that *atp9* has been functionally transferred to the nucleus in at least four *Silene* species—*S. latifolia*, *S. noctiflora*, *S. vulgaris* and *S. paradoxa*.

Whole mitochondrial genome sequences confirm that *atp9* is mitochondrially encoded in both *S. latifolia* and *S. noctiflora* (Sloan *et al.*, unpublished data). In addition, the gene has been shown to be maternally inherited in *S. vulgaris* [26]. Comparing cDNA and genomic sequence also confirms that *atp9* contains a site that undergoes C-to-U RNA editing in *S. paradoxa*—a process that is characteristic of organellar but not nuclear genes in plants (Sloan *et al.*, unpublished data). Although we cannot definitively rule out the possibility of nuclear transfer, these data strongly suggest that nuclear transfer is not the driving force behind the pattern of elevated substitution rate observed in *atp9*.

It is also clear that *atp9* is functional based on its low $\omega$ values and the absence of internal stop codons. Therefore, we conclude that the most likely explanation for the high levels of divergence is an increased mutation rate that is specific to *atp9* (or a subset of the mitochondrial genome that includes *atp9*).

The molecular evolution of *atp9* could be influenced by the presence of multiple gene copies in at least some species (see Methods). The existence of multiple copies could reflect heteroplasmy resulting from paternal leakage [25], non-functional paralogs in the mitochondria or other genomes [50], or the existence of multiple functional mitochondrial copies [51]. It is conceivable that *atp9* is located in a region of active recombination within the *Silene* mitochondrial genome or is experiencing frequent retroprocessing back into the genome from mRNA. Both of these processes may be mutagenic as well as lead to gene duplication and, therefore, would be consistent with our observations [6, 52, 53]. Alternatively, high mutations rates in *atp9* may have simply increased divergence between heteroplasmic and/or paralogous copies, thereby enhancing our ability to detect multiple copies of *atp9* even though

they exist for other genes as well. Sequencing complete mitochondrial genomes, analyzing relative copy number of *atp9* variants, and sampling multiple individuals per species would help distinguish between these possibilities. In a sample of individuals from 40 different populations of *S. vulgaris*, we found 4 individuals with multiple *atp9* copies, and certain variants were only found in multi-copy individuals (unpublished data). This result suggests there is polymorphism for the presence of a paralogous copy within *S. vulgaris*, although heteroplasmy involving a rare haplotype is also plausible.

The acceleration in *atp9* appears to be common to most of *Silene*/*Lychnis*. In contrast, most of the other *Sileneae* genera exhibit more conventional substitution rates for *atp9*, although their rates are still elevated on average. This pattern is consistent with an *atp9*-specific increase in substitution rate very early in the divergence of *Silene*, which may have been magnified by further accelerations in local areas of the genus.

A previous study of mitochondrial substitution rate variation across the seed plant phylogeny identified a handful of individual species exhibiting elevated divergence in one gene but not others [4]. Our observations of rate variation in *atp9* within *Silene* indicate that such gene-specific effects can be maintained across large clades of species over millions of years. We also found that these effects can occur quite locally. Most notably, *S. nutans* exhibited an $R_S$ value of 80 SSB for *atp1* (a rate that exceeds all other species for that gene), but it showed no sign of acceleration in *nad9* or *cox3* (Figure 3). A number of other species showed more modest rate increases in *atp1* and/or *cox3* without correlated accelerations in other genes. These patterns may reflect local mutational effects within the genome. Alternatively, given

the mounting evidence for recombination in plant mtDNA [41, 46, 54] and the existence of rate variation both within and among species [7], rate discrepancies between genes may be the result of recombination between genomes with different mutational histories. Finally, the possibility of nuclear transfer for a gene such as *atp1* in *S. nutans* should also be considered [4].

**Evolution at synonymous and non-synonymous sites**

Despite the massive variation in synonymous substitution rates among genes and species, we found that rates of non-synonymous substitution generally remained low (although there was a positive correlation between $R_N$ and $R_S$ across branches: Figure 3, Table 3). Across genes, $R_S$ values vary by 9 to 41-fold (depending on whether the six species with apparent genome-wide accelerations are included), while $R_N$ values vary by only 2 to 3-fold (Table 2). As a result, there is an apparent negative relationship between $R_S$ and the ratio of non-synonymous to synonymous changes (ω). While $R_S$ is commonly interpreted as a measure of the mutation rate, ω is used as an estimate of the intensity/efficacy of purifying selection (*i.e.* "the selective sieve" [55]). Under these interpretations, our data would suggest that genes experiencing high mutations rates also face greater purifying selection. In contrast, the opposite pattern has been observed in comparisons of nuclear genes in mammals [56].

The relationship between $R_S$ and $\omega$ among mitochondrial genes in *Silene* should be confirmed in a larger sample, because we have examined only 4 loci in the present study, and *atp9* may generally be subject to strong purifying selection [57]. Whether $R_S$ and $\omega$ can be reliably interpreted as measures of mutation rate and purifying selection depends on the distribution of fitness effects for mutations at synonymous and non-synonymous sites, which are not well understood in plant

mitochondrial genomes. These distributions will dictate how the synonymous and non-synonymous substitution rates scale with the mutation rate.

In a comparison of sequence divergence in 15 protein-coding mitochondrial genes between angiosperms and the liverwort *Marchantia,* Laroche *et al.* [57] found much greater variation among genes in $d_N$ than in $d_S$—the opposite of what we observed. This discrepancy highlights the importance of phylogenetic scale in these studies. Across deep nodes in the land plant phylogeny, local differences in gene-specific mutation rates are apparently averaged out, and variation in the magnitude of purifying/positive selection among genes becomes the primary determinant of evolutionary rates. In contrast, at the local phylogenetic scale of our study, the signature of gene-specific differences in mutation rate is apparently maintained.

Because of the high variance and abundance of 0 values associated with short branches in our analysis, it is difficult to test for the same relationship between $R_S$ and $\omega$ across lineages that we observed across genes. We did see, however, that removing the rapidly evolving branches from each tree raised the $\omega$ ratio for all four genes, suggesting that the same pattern may hold. Mower *et al.* [4] conducted a broad phylogenetic survey of seed plants in which short branch lengths would be expected to be less of a problem. Their data showed a strong negative relationship between $R_S$ and $\omega$ across lineages (see also [53]). Therefore, it appears that major increases in mitochondrial synonymous substitution rate—either gene or taxon-specific—are accompanied by a less than proportional increase in non-synonymous substitution rate such that the effects of an apparent increased mutational pressure on amino acid sequences are greatly dampened.

**Uncertainty in divergence time estimates**

We used molecular clock based methods to estimate divergence times in our *matK* gene tree. The estimated ages were generally older than estimates from two previous studies [4, 30]. The discrepancy between these studies is likely attributable to two major differences. First, there is simple difference in calibration age between our study and the analysis of Mower *et al*. (2007), which utilized an age of 38 Myr for the *Beta*/*Silene* divergence derived from a broader molecular clock analysis of the angiosperms [58]. This date appears to be in conflict with our fossil calibration point, as all three of our analyses estimate the age of the *Beta*/*Silene* split to be at least 52 Myr old. This distinction, however, cannot explain the contrasting results between our study and that of Frajman *et al.* [30], because we used essentially the same calibration point. Instead, we note that there was a significant difference in sampling schemes between these two studies. Our focus on the genus *Silene* produced a very imbalanced topology with much denser branching in certain parts of the tree than others. In contrast, Frajman *et al.* [30] utilized a much more balanced phylogenetic sampling. Because it is easier to detect multiple substitutions at the same site in regions with lots of branching, there is a tendency to estimate longer branch lengths in species-rich parts of a phylogeny (the "node density effect" [59]). This effect may contribute to our older age estimates within the tribe *Sileneae*. Because of this potential bias, the uncertainty over calibration points and the many assumptions associated with molecular clock based dating, it is important to stress that the divergence times used in this analysis should be considered only as approximations.

Divergence time estimates are necessary to calculate absolute substitution rates, so dating uncertainty should be considered in comparing absolute rates across

studies. For example, re-calibrating node ages within *Silene* to correspond with the 12 Myr divergence time for the genus estimated by Frajman *et al*. [30] would increase our substitution rate estimates by approximately 50%. In relative terms, however, our three dating analyses were quite consistent within *Sileneae*. Therefore, our estimates of proportional variation in substitution rate across species and genes are less sensitive to dating method.

## Conclusions

Based on our analysis of mitochondrial divergence within the tribe *Sileneae,* we conclude that mutational acceleration is not restricted to a single species nor is it completely confined to a small number of high rate lineages. The patterns of divergence in *atp9* illustrated that elevated rates have been maintained throughout much of the genus *Silene* for at least one mitochondrial gene, highlighting a complex gene x species interaction in the distribution of rate variation. The diversity in phylogenetic and genomics scale suggests that there is no simple rule or single mechanism underlying mutation rate variation in plant mitochondrial genomes. Elucidating the mechanistic forces that shape mutation rate variation should represent a high priority in the field of plant mitochondrial genomics. *Silene* was targeted for this in-depth sampling of species-level mitochondrial divergence because of *a priori* knowledge of the rate acceleration in *S. noctiflora*. Determining whether the patterns of rate variation among species and among genes in *Silene* are broadly representative of angiosperm genera or represent something unique about the molecular evolution of *Silene* will require similar levels of sampling in taxa that currently show no evidence of rate increase.

# Methods

## Study species

*Silene* (Caryophyllaceae) comprises approximately 700 predominantly herbaceous species that vary substantially in life history and breeding system [60]. The genus has become a model system for diverse areas of research with a particular focus on the molecular evolution of organelle genomes, including studies of population genetics [61, 62], organelle transmission [25, 27], evolutionary rates [4, 7, 35, 45], and cytoplasmic male sterility [26, 44]. *Silene* belongs to the tribe *Sileneae*, which has been the subject of extensive and ongoing phylogenetic analysis [30-32, 63-65]. Taxa were selected so as to represent major groups that will appear in a forthcoming revised taxonomy of the genus (Oxelman *et al*. in prep). For this study, we used a combination of field collected samples and preserved herbarium specimens along with previously published sequence data. Sample collection and voucher information are summarized in Table 1.

## DNA extraction, PCR and Sequencing

We extracted total genomic DNA from each sample. For silica-dried samples and herbarium specimens, we followed the protocol described by Oxelman *et al.* [31] and performed subsequent purification using the Qiagen QIAquick Purification Kit protocol, Ultra Silica Bead kit (ABgene), or GFX PCR DNA and Gel Band Purification Kit (Amersham Biosciences). For fresh tissue samples, extractions were performed using the Qiagen Plant DNeasy Kit.

We PCR amplified the full-length coding sequence of the chloroplast gene *maturase K* (*matK*) and portions of four mitochondrial protein coding genes: *ATP*

*synthase subunit 1* (*atp1*), *ATP synthase subunit 9* (*atp9*), *cytochrome c oxidase subunit 3* (*cox3*) and *NADH dehydrogenase subunit 9* (*nad9*). [See Additional file 8 for PCR primer sequences.]

PCR products were cleaned with Exonuclease I and shrimp alkaline phosphatase (USB Corporation), cycle sequenced with BigDye v3.1 (Applied Biosystems), and analyzed on an ABI 3130xl capillary sequencer. Automated basecalls were edited manually using published *Beta vulgaris* sequences as a reference for reading frame, and sequences were assembled into contigs using Sequencher v4.5 (Gene Codes). All sequences obtained for *S. sorensenis* were identical to those from *S. involucrata*, so *S. sorensenis* was excluded to simplify subsequent analysis. DNA sequences have been submitted to GenBank [see Additional file 9 for accession numbers]. Sequence alignments were generated using the Clustal function imbedded in MEGA v4.0 [66] and edited manually [see Additional file 10].

### *matK* phylogenetic analysis and dating

We estimated the phylogeny of our sample based on the *matK* dataset, using both maximum likelihood (ML) and maximum parsimony (MP) criteria in PAUP* v4.0b10 [67]. In addition to the species listed in Table 1, our analysis also included *matK* sequences from GenBank for the following outgroups: *Beta vulgaris* (Amaranthaceae), *Illecebrum verticillatum* (Caryophyllaceae) and *Scleranthus perennis* (Caryophyllaceae). Our ML search employed a GTR+Γ substitution model with fixed parameter values identified based on an analysis of our full *matK* dataset (including outgroups) using the AIC method in ModelTest v3.7 [68]. The ML topology was identified with a heuristic search using the TBR branch swapping

algorithm, the MULTREES option in effect, and random addition of sequences with 10 replicates. A MP bootstrap analysis based on 1000 replicate datasets was performed using the same heuristic search settings except with MULTREES off. We performed ML and MP analyses in the same fashion for each mitochondrial gene.

We used three different techniques to estimate divergence times from our *matK* gene tree: (1) a Bayesian relaxed clock model implemented in BEAST v1.4.8 [69], (2) a penalized likelihood (PL) method, and (3) the Langley-Fitch (LF) method. The latter two methods were both performed in r8s v1.71 [70]. The LF model is a maximum likelihood strict molecular clock method that enforces a constant substitution rate over the entire tree. The other two methods allow for rate variation among branches. The PL approach assumes that rates are correlated across adjacent branches and penalizes models that require rapid rate changes within the tree. In contrast, the BEAST analysis constrained the rate variation among branches to a lognormal distribution but placed no restriction on correlations between adjacent branches. All dating analyses incorporated an extra outgroup, *Nepenthes glabrata* (Nepenthaceae), which was added solely to determine the position of the root along the *Beta vulgaris* branch. It was pruned from the resulting trees and discarded from all subsequent analyses. We used a calibration time of 34 million years (Myr) for the split between *Scleranthus* and *Sileneae*, which corresponds to the recent analysis of Frajman *et al.* [30] and the fossil evidence described therein.

The BEAST analysis was conducted with a GTR+Γ model of substitution with 4 rate categories, empirical base frequencies and a birth-death process tree prior. We defined a monophyletic ingroup to include all species except *Nepenthes*, *Beta* and *Illecebrum*. The calibration date was effectively fixed by specifying a normal

distribution with mean of 34 and standard deviation of 0.00001 as the prior for the time to most recent common ancestor (TMRCA) of the pre-defined ingroup. We ran 3 MCMC chains of length 50 million each with trees saved every 25,000 iterations. The first 1000 trees (50%) from each chain were discarded as burn in, and chains were combined after verifying convergence among runs. We generated a maximum credibility tree with mean node heights as well as a tree that was constrained to the 70% bootstrap consensus topology from our parsimony analysis. BEAST reported the estimated node ages along with the associated 95% high probability densities (HPDs), as well as the posterior probability for each node.

For the PL and LF analyses in r8s, we used the 70% parsimony bootstrap consensus topology with branch lengths optimized under ML in PAUP*. We fixed the root age of the tree to an arbitrary value of 1 and calibrated the resulting output tree with the 34 Myr fossil age for the *Scleranthus*/*Sileneae* split. Both analyses utilized the TN search algorithm with 10 restarts, 10 time guesses, and the checkgradient option on. The cross-validation procedure was used to determine an optimal smoothing parameter of 0.0022 for the PL analysis. Error in divergence times for both methods was estimated based on the distribution of 100 bootstrap replicate datasets, following the recommendations in the r8s documentation.

**Estimating $d_N$ and $d_S$**

We estimated the branch lengths of *matK* and all 4 mitochondrial genes individually in terms of synonymous ($d_S$) and non-synonymous ($d_N$) substitutions per site, using a codon-based model of substitution within the codeml application in PAML v4.0 [71]. We also analyzed a concatenation of all 4 mitochondrial genes and a concatenation of *atp1*, *cox3* and *nad9* only. Tree topologies were constrained based on the *matK* 70%

parsimony bootstrap consensus. Codon frequencies were determined by an F1x4 model. The parameters values for $\omega$ and transition/transversion ratio were estimated from the data with initial values of 0.4 and 2 respectively. Separate $\omega$ values were estimated for each branch. As in the dating analysis, *Nepenthes glabrata* was used as an outgroup in the *matK* dataset to identify the position of the root along the *Beta vulgaris* branch. *Arabidopsis thaliana* served a similar purpose for the mitochondrial genes. These outgroups were pruned from the resulting trees and not considered further. Because the process of C-to-U RNA editing can bias the estimation of $d_N$ and $d_S$ values in plant mitochondrial genomes [72], we excluded all codons known to undergo RNA editing in *Beta vulgaris* [73].

In the *matK* dataset, 6 species (*Heliosperma pusillum*, *S. conica*, *S. paradoxa*, *S. samojedora*, *S. seoulensis* and *S. conoidea*) produced sequences with an apparent frameshift indel in one of two homopolymer regions, raising the possibility that we sequenced pseudogenes in these species. In addition, *S. otites* did not have a start codon at the conserved position in the *matK* alignment. These sequences showed little indication of elevated rates or other abnormal substitution patterns, and their phylogenetic placement was consistent with *a priori* information. Therefore, they were retained in the dataset, and codons that contained frameshift indels were removed to keep all sequences in frame.

For the mitochondrial dataset, we obtained sequences for *atp1*, *cox3*, and *nad9* from all sampled species, but only a subset of *atp9* sequences were successfully generated. For 5 of the 74 species in our sample, we failed to successfully amplify and sequence *atp9*, and an additional 9 species appeared to have multiple *atp9* copies. In the latter group, sequencing electropherograms indicated the presence of two

different nucleotides in between 1.2 and 19.5% of sites. These samples were excluded from the analysis.

**Estimates of absolute substitution rates ($R_N$ and $R_S$)**

Following the basic methodology described by Cho *et al.* [5], absolute substitution rates (substitutions per site per year) can be obtained by dividing branch lengths (defined in terms of substitution per site) by the age of the branch. We calculated branch ages using the divergence times estimated by BEAST. We divided the $d_N$ and $d_S$ values reported by PAML for each branch by the respective branch age to obtain absolute substitution rates in terms of non-synonymous and synonymous sites ($R_N$ and $R_S$, respectively). Standard errors for $R_N$ and $R_S$ were calculated as described by Parkinson *et al*. [6] where the standard errors for node ages were approximated as one quarter of the 95% HPD. Pearson correlations coefficient for $R_N$ and $R_S$ values across and within genes were calculated with PROC CORR in SAS Software v9.1.

## Authors' contributions

DBS conceived of the study, participated in its design, conducted the bulk of the sequencing and data analysis, and drafted the manuscript. BO participated in design of the study, collected and identified specimens, and helped with data analysis and drafting of the manuscript. AR collected and identified specimens and helped with sequencing, data analysis, and drafting of the manuscript. DRT participated in design of the study and helped with data analysis and drafting the manuscript. All authors read and approved the final manuscript.

## Acknowledgements

We thank Stephanie Goodrich for providing *Agrostemma* seeds and Nahid Heidari and Vivian Aldén for assistance in the lab. We also appreciate comments on an earlier version of this manuscript from Steve Keller, Magnus Lidén, Matt Olson, and the members of the Taylor lab. This study was supported by NSF DEB-0808452 (to DBS and DRT), NSF DEB-0349558 (to DRT) and grants from the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (to BO).

## References

1. Wolfe KH, Li WH, Sharp PM: **Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs.** Proc Natl Acad Sci 1987, **84**(24):9054-9058.

2. Palmer JD, Herbon LA: **Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence.** J Mol Evol 1988, **28**(1-2):87-97.

3. Drouin G, Daoud H, Xia J: **Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants.** Mol Phylogenet Evol 2008, **49**(3):827-831.

4. Mower JP, Touzet P, Gummow JS, Delph LF, Palmer JD: **Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants.** BMC Evol Biol 2007, **7**(1):135.

5. Cho Y, Mower JP, Qiu YL, Palmer JD: **Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants.** Proc Natl Acad Sci 2004, **101**(51):17741-17746.

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

7. Sloan DB, Barr CM, Olson MS, Keller SR, Taylor DR: **Evolutionary rate variation at multiple levels of biological organization in plant mitochondrial DNA.** Mol Biol Evol 2008, **25**(2):243-246.

8. Laird CD, McConaughy BL, McCarthy BJ: **Rate of fixation of nucleotide substitutions in evolution.** Nature 1969, **224**(5215):149-154.

9. Britten RJ: **Rates of DNA sequence evolution differ between taxonomic groups.** Science 1986, **231**(4744):1393-1398.

10. Martin AP, Palumbi SR: **Body size, metabolic rate, generation time, and the molecular clock.** Proc Natl Acad Sci 1993, **90**(9):4087-4091.

11. Gaut BS, Morton BR, McCaig BC, Clegg MT: **Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL.** Proc Natl Acad Sci 1996, **93**(19):10274-10279.

12. Sniegowski PD, Gerrish PJ, Johnson T, Shaver A: **The evolution of mutation rates: separating causes from consequences.** Bioessays 2000, **22**(12):1057-1066.

13. Foury F, Hu J, Vanderstraeten S: **Mitochondrial DNA mutators.** Cell Mol Life Sci 2004, **61**(22):2799-2811.

14. Smith SA, Donoghue MJ: **Rates of molecular evolution are linked to life history in flowering plants.** Science 2008, **322**(5898):86-89.

15. Kimura M: *The Neutral Theory of Molecular Evolution:* Cambridge: Cambridge University Press; 1983.

16. Chamary JV, Parmley JL, Hurst LD: **Hearing silence: non-neutral evolution at synonymous sites in mammals.** Nat Rev Genet 2006, **7**(2):98-108.

17. Denver DR, Morris K, Lynch M, Vassilieva LL, Thomas WK: **High direct estimate of the mutation rate in the mitochondrial genome of *Caenorhabditis elegans*.** Science 2000, **289**(5488):2342-2344.

18. Muse SV: **Examining rates and patterns of nucleotide substitution in plants.** Plant Mol Biol 2000, **42**(1):25-43.

19. Drake JW, Charlesworth B, Charlesworth D, Crow JF: **Rates of spontaneous mutation.** Genetics 1998, **148**(4):1667-1686.

20. Lynch M, Koskella B, Schaack S: **Mutation pressure and the evolution of organelle genomic architecture.** Science 2006, **311**(5768):1727-1730.

21. Baer CF, Miyamoto MM, Denver DR: **Mutation rate variation in multicellular eukaryotes: causes and consequences.** Nat Rev Genet 2007, **8**:619-631.

22. Bergsten J: **A review of long-branch attraction.** Cladistics 2005, **21**(2):163-193.

23. Desplanque B, Viard F, Bernard J, Forcioli D, Saumitou-Laprade P, Cuguen J, Van Dijk H: **The linkage disequilibrium between chloroplast DNA and mitochondrial DNA haplotypes in *Beta vulgaris* ssp. *maritima* (L.): the usefulness of both genomes for population genetic studies.** Mol Ecol 2000, **9**(2):141-154.

24. Olson MS, McCauley DE: **Linkage disequilibrium and phylogenetic congruence between chloroplast and mitochondrial haplotypes in *Silene vulgaris*.** Proc R Soc Lond B 2000, **267**(1454):1801-1808.

25. McCauley DE, Bailey MF, Sherman NA, Darnell MZ: **Evidence for paternal transmission and heteroplasmy in the mitochondrial genome of *Silene vulgaris*, a gynodioecious plant.** Heredity 2005, **95**(1):50-58.

26. Houliston GJ, Olson MS: **Nonneutral evolution of organelle genes in *Silene vulgaris*.** Genetics 2006, **174**(4):1983-1994.

27. McCauley DE, Sundby AK, Bailey MF, Welch ME: **Inheritance of chloroplast DNA is not strictly maternal in *Silene vulgaris* (Caryophyllaceae): evidence from experimental crosses and natural populations.** Am J Bot 2007, **94**(8):1333.

28. Hilu KW, Liang H: **The *matK* gene: sequence variation and application in plant systematics.** Am J Bot 1997, **84**(6):830-830.

29. Barthet MM, Hilu KW: **Evaluating evolutionary constraint on the rapidly evolving gene *matK* using protein composition.** J Mol Evol 2008, **66**(2):85-97.

30. Frajman B, Eggens F, Oxelman B: **Hybrid origins and homoploid reticulate evolution within *Heliosperma* (*Sileneae*, Caryophyllaceae) – a multigene phylogenetic approach with relative dating.** Systematic Biology 2009, **58**(3):328-345.

31. Oxelman B, Liden M, Berglund D: **Chloroplast *rps16* intron phylogeny of the tribe *Sileneae* (Caryophyllaceae).** Plant Syst Evol 1997, **206**(1-4):393-410.

32. Erixon P, Oxelman B: **Reticulate or tree-like chloroplast DNA evolution in *Sileneae* (Caryophyllaceae)?** Mol Phylogenet Evol 2008, **48**(1):313-325.

33. Popp M, Oxelman B: **Origin and evolution of North American polyploid *Silene* (Caryophyllaceae).** Am J Bot 2007, **94**(3):330.

34. Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK: **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 2006, **23**(11):2175.

35. Erixon P, Oxelman B: **Whole-gene positive selection, elevated synonymous substitution rates, duplication, and indel evolution of the chloroplast *clpP1* gene.** PLoS ONE 2008, **3**(1):e1386.

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

37. Qiu YL, Li L, Wang B, Chen Z, Knoop V, Groth-Malonek M, Dombrovska O, Lee J, Kent L, Rest J: **The deepest divergences in land plants inferred from phylogenomic evidence.** Proc Natl Acad Sci 2006, **103**(42):15511-15516.

38. Nabholz B, Glemin S, Galtier N: **Strong variations of mitochondrial mutation rate across mammals--the longevity hypothesis.** Mol Biol Evol 2008, **25**(1):120-130.

39. Rautenberg A: **Phylogenetic relationships of *Silene* sect. *Melandrium* and allied taxa (Caryophyllaceae), as deduced from multiple gene trees.** Ph.D. thesis, Uppsala University; Uppsala**,** 2009

40. Rautenberg A, Filatov D, Svennblad B, Heidari N, Oxelman B: **Conflicting phylogenetic signals in the *SlX1/Y1* gene in *Silene*.** BMC Evol Biol 2008, **8**(1):299.

41. Touzet P, Delph LF: **The effect of breeding system on polymorphism in mitochondrial genes of Silene.** Genetics 2009, **181**(2):631-644.

42. Lenaz G: **Role of mitochondria in oxidative stress and ageing.** Biochim Biophys Acta 1998, **1366**(1-2):53-67.

43. Ingvarsson PK, Taylor DR: **Genealogical evidence for epidemics of selfish genes.** Proc Natl Acad Sci 2002, **99**(17):11265-11269.

44. Stadler T, Delph LF: **Ancient mitochondrial haplotypes and evidence for intragenic recombination in a gynodioecious plant.** Proc Natl Acad Sci U S A 2002, **99**(18):11730-11735.

45. Barr CM, Keller SR, Ingvarsson PK, Sloan DB, Taylor DR: **Variation in mutation rate and polymorphism among mitochondrial genes in *Silene vulgaris*.** Mol Biol Evol 2007, **24**(8):1783-1791.

46. Richardson AO, Palmer JD: **Horizontal gene transfer in plants.** J Exp Bot 2007, **58**(1):1-9.

47. Ioerger TR, Clark AG, Kao T: **Polymorphism at the self-incompatibility locus in Solanaceae predates speciation.** Proc Natl Acad Sci 1990, **87**(24):9732-9735.

48. Adams KL, Qiu YL, Stoutemyer M, Palmer JD: **Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution.** Proc Natl Acad Sci 2002, **99**(15):9905-9912.

49. Li WH, Gojobori T, Nei M: **Pseudogenes as a paradigm of neutral evolution.** Nature 1981, **292**:237-239.

50. Lin X, Kaul S, Rounsley S, Shea TP, Benito MI, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M: **Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*.** Nature 1999, **402**:761-768.

51. Marienfeld J, Unseld M, Brandt P, Brennicke A: **Genomic recombination of the mitochondrial *atp6* gene in *Arabidopsis thaliana* at the protein processing site creates two different presequences.** DNA Research 1996, **3**(5):287-290.

52. Lercher MJ, Hurst LD: **Human SNP variability and mutation rate are higher in regions of high recombination.** Trends in Genetics 2002, **18**(7):337-340.

53. Bakker FT, Breman F, Merckx V: **DNA sequence evolution in fast evolving mitochondrial DNA *nad1* exons in Geraniaceae and Plantaginaceae.** Taxon 2006, **55**(4):887-896.

54. Pearl SA, Welch ME, McCauley DE: **Mitochondrial heteroplasmy and paternal leakage in natural populations of *Silene vulgaris*, a gynodioecious plant.** Mol Biol Evol 2009, **26**(3):537-545.

55. Lynch M, Blanchard JL: **Deleterious mutation accumulation in organelle genomes.** Genetica 1998, **103**:29-39.

56. Wyckoff GJ, Malcom CM, Vallender EJ, Lahn BT: **A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate.** Trends in Genetics 2005, **21**(7):381-385.

57. Laroche J, Li P, Maggia L, Bousquet J: **Molecular evolution of angiosperm mitochondrial introns and exons.** Proc Natl Acad Sci 1997, **94**(11):5722-5727.

58. Wikström N, Savolainen V, Chase MW: **Evolution of the angiosperms: calibrating the family tree.** Proc R Soc Lond B 2001, **268**(1482):2211.

59. Venditti C, Meade A, Pagel M: **Detecting the node-density artifact in phylogeny reconstruction.** Syst Biol 2006, **55**(4):637-643.

60. Brach AR, Song H: **eFloras: New directions for online floras exemplified by the Flora of China Project.** Taxon 2006, **55**(1):188.

61. McCauley DE: **Contrasting the distribution of chloroplast DNA and allozyme polymorphism among local populations of *Silene alba*: implications for studies of gene flow in plants.** Proc Natl Acad Sci 1994, **91**(17):8127-8131.

62. Taylor DR, Keller SR: **Historical range expansion determines the phylogenetic diversity introduced during contemporary species invasion.** Evolution 2007, **61**(2):334-345.

63. Oxelman B, Liden M: **Generic boundaries in the tribe *Sileneae* (Caryophyllaceae) as inferred from nuclear rDNA sequences.** Taxon 1995, **44**(4):525-542.

64. Popp M, Oxelman B: **Evolution of a RNA polymerase gene family in *Silene* (Caryophyllaceae) - Incomplete concerted evolution and topological congruence among paralogues.** Syst Biol 2004, **53**(6):914-932.

65. Frajman B, Oxelman B: **Reticulate phylogenetics and phytogeographical structure of *Heliosperma* (*Sileneae*, Caryophyllaceae) inferred from chloroplast and nuclear DNA sequences.** Mol Phylogenet Evol 2007, **43**(1):140-155.

66. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.** Mol Biol Evol 2007, **24**(8):1596-1599.

67. Swofford DL: *PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4:* Sunderland, MA: Sinauer Associates; 1998.

68. Posada D, Crandall KA: **MODELTEST: testing the model of DNA substitution.** Bioinformatics 1998, **14**(9):817-818.

69. Drummond AJ, Rambaut A: **BEAST: Bayesian evolutionary analysis by sampling trees.** BMC Evol Biol 2007, **7**:214.

70. Sanderson MJ: **r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock.** Bioinformatics 2003, **19**(2):301-302.

71. Yang Z: **PAML 4: Phylogenetic Analysis by Maximum Likelihood.** Mol Biol Evol 2007, **24**(8):1586-1591.

72. Lu MZ, Szmidt AE, Wang XR: **RNA editing in gymnosperms and its impact on the evolution of the mitochondrial *coxI* gene.** Plant Mol Biol 1998, **37**(2):225-234.

73. Mower JP, Palmer JD: **Patterns of partial RNA editing in mitochondrial genes of *Beta vulgaris*.** Mol Genet Genomics 2006, **276**(3):285-293.

74. Wolfe KH, Sharp PM, Li WH: **Rates of synonymous substitution in plant nuclear genes.** J Mol Evol 1989, **29**(3):208-211.

75. Holmgren PK, Holmgren NH, Barnett LC: *Index Herbariorum: Part 1. The herbaria of of the world:* New York: New York Botanical Garden; 1990.

# Figures

**Figure 1 - Diversity in substitution rates**

Synonymous substitution rates per site per billion years (SSB) for different organisms and genomes plotted on a log scale. Black bars represent seed plant mitochondrial genomes. Average rates for animal taxa from Lynch *et al.* [20]; angiosperm chloroplast and nuclear estimates from Wolfe *et al.* [74]; mitochondrial rates for individual plant species taken from Cho *et al.* [5] and Mower *et al.* [4].

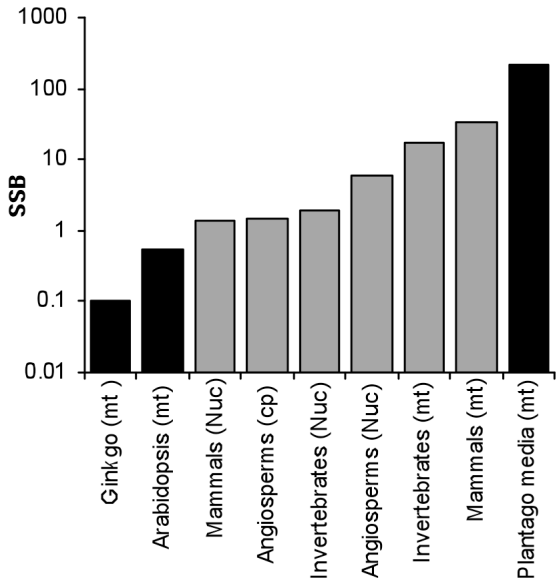


**Figure 2 - Chronogram showing divergence times estimated in BEAST based on full-length *matK* coding sequence.**

Time scale is in millions of years. Error bars at each node show 95% HPD for node age. Values to the right of each node show parsimony bootstrap support and Bayesian posterior probability (in that order) for the corresponding clade. Tree topology was constrained based on 70% parsimony bootstrap consensus.

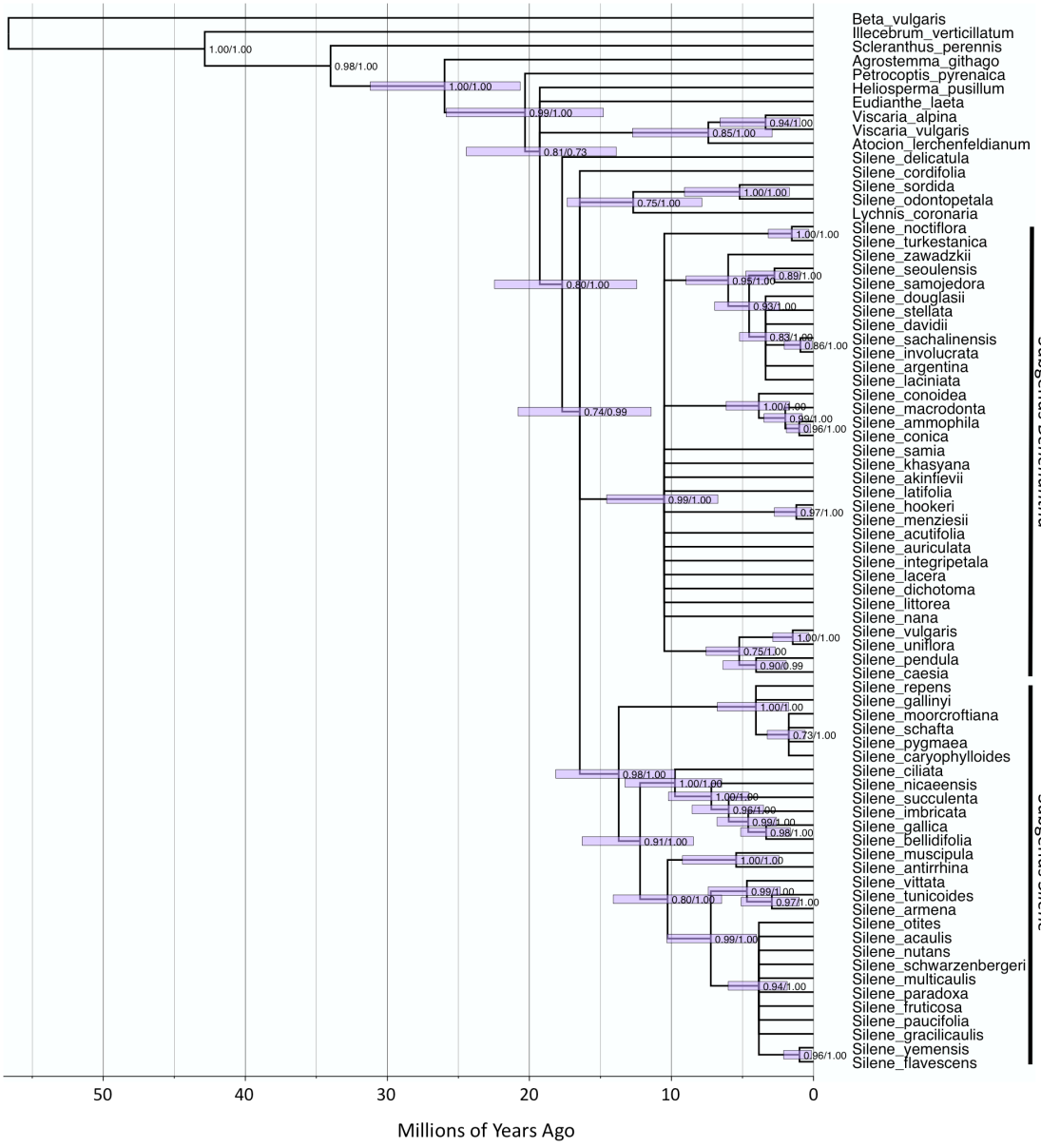


**Figure 3 - $d_N$ and $d_S$ trees for mitochondrial genes.**

Branch lengths are in terms of non-synonymous ($d_N$) or synonymous ($d_S$) substitutions per site as estimated by PAML under a constrained topology. The scale is the same for all trees.

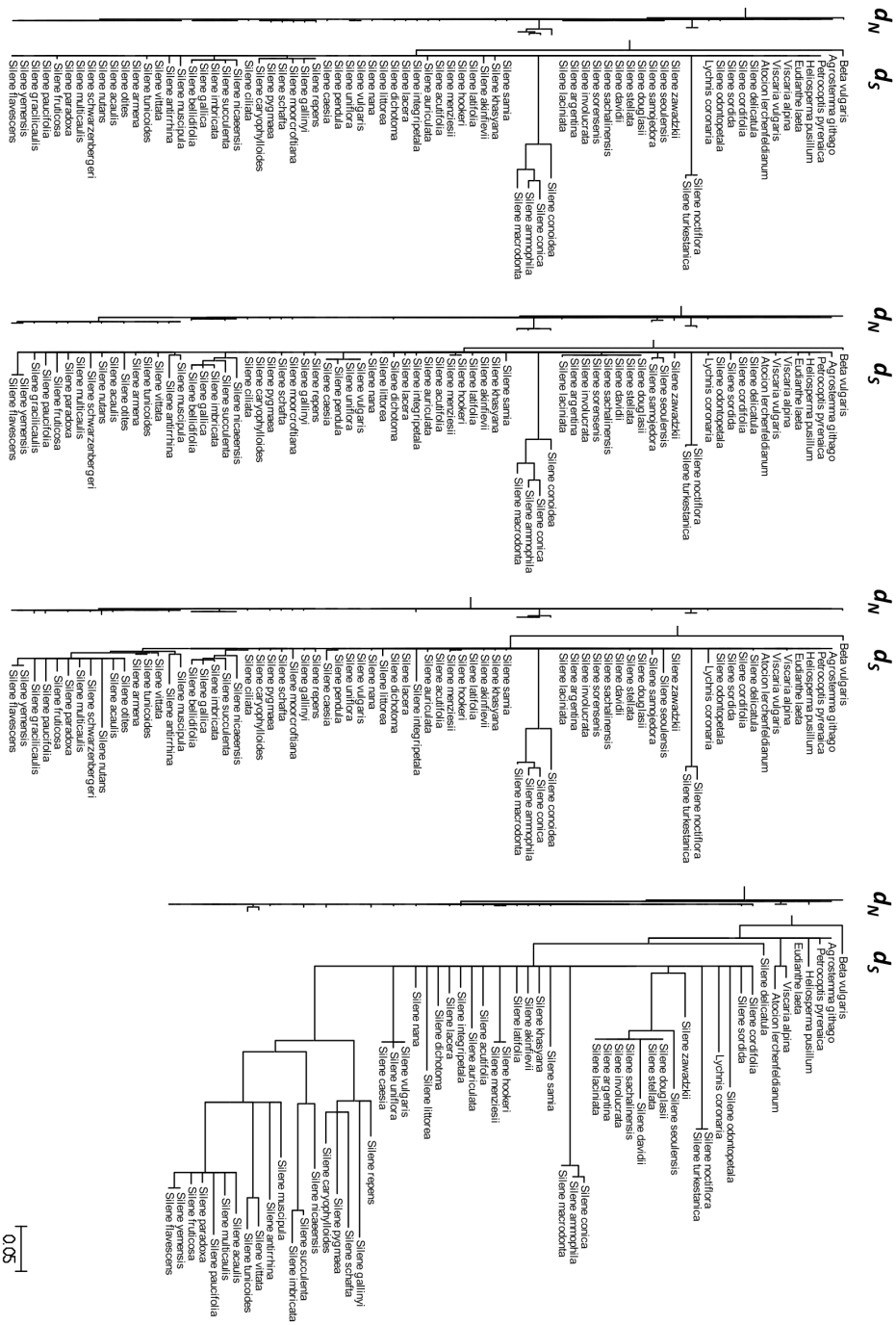
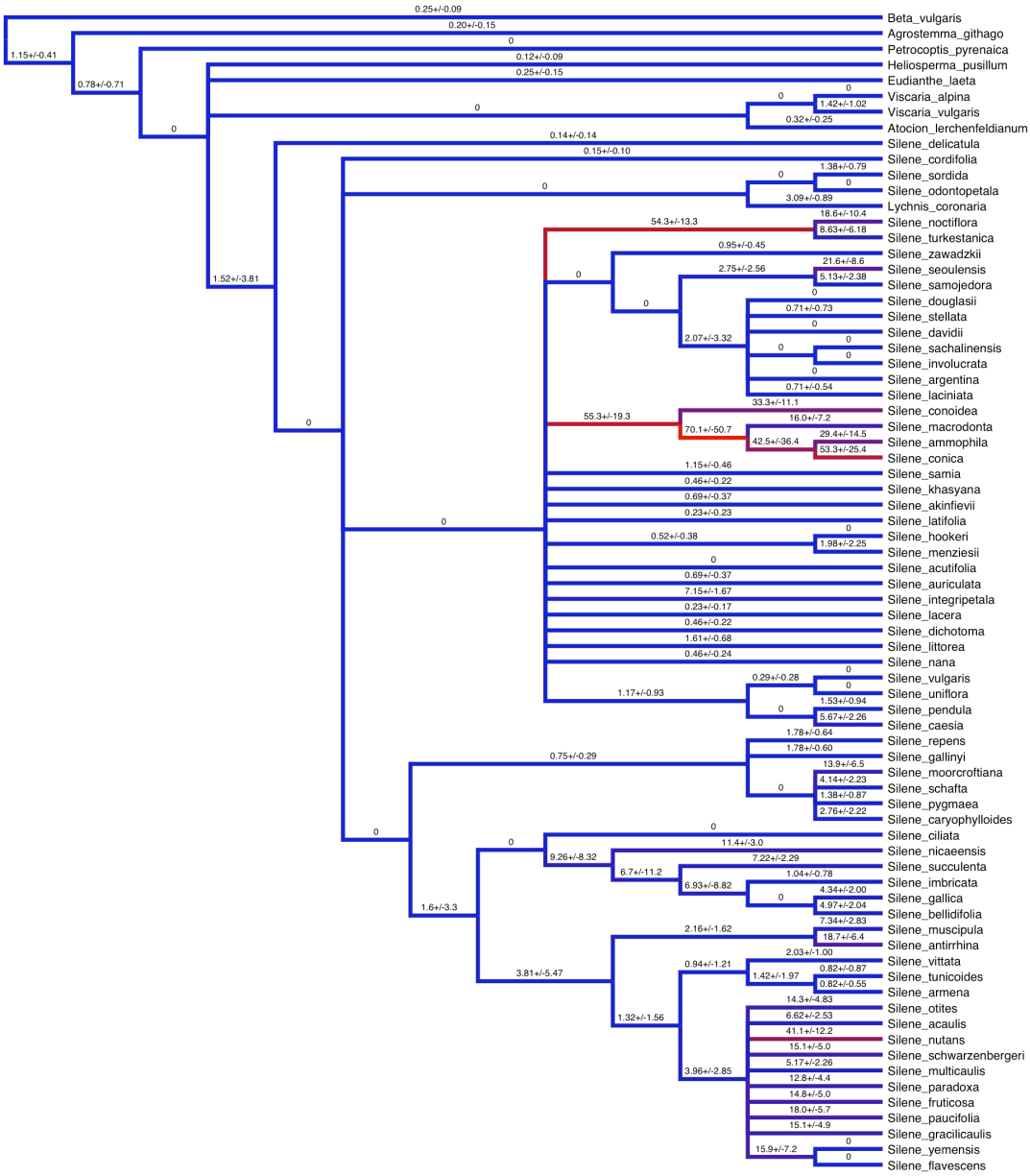


**Figure 4 - Phylogenetic variation in $R_S$.**

Branches labelled with absolute synonymous substitution rates and approximate standard errors based on concatenation of *nad9*, *cox3* and *atp1*. Branch colors indicate fast (red) and slow (blue) rates.



# Tables

## Table 1 - Sampled species and voucher information.

| Species | Voucher |
|---|---|
| *Agrostemma githago* L. | D. Sloan 001 (VPI) |
| *Atocion lerchenfeldianum* (Baumg.) M.Popp | Strid 24875 (GB) |
| *Eudianthe laeta* (Aiton) Rchb. ex Wilk. | Strandhede *et al.* 690 (GB) |
| *Heliosperma pusillum* (Waldst. & Kit.) Rchb. | E. Zogg ZH 1438 (Z) |
| *Lychnis coronaria* (L.) Desr. | N/A. Collected by D. Sloan. Charlottesville, VA, USA |
| *Petrocoptis pyrenaica* A.Br. | Schneeweiss *et al.* 6549 (WU) |
| *Silene acaulis* (L.) Jacq. | *Schneeweiss 5315 (WU) |
| *Silene acutifolia* Link ex Rohrb. | Rothmaler 13691 (S) |
| *Silene akinfievii* Schmalh. | Portenier 3814 (LE) |
| *Silene ammophila* Boiss. & Heldr. | Raus 7631 (GB) |
| *Silene antirrhina* L. | N/A. Collected by D. Sloan. Kellog, MN, USA |
| *Silene argentina* (Pax) Bocquet | M. Popp 2005-11-11 (GB) |
| *Silene armena* Boiss. | B. Oxelman 2436 (GB) |
| *Silene auriculata* Sibth. & Sm. | Baden & Franzén 795 (Strid) |
| *Silene bellidifolia* Jacq. | Strid *et al.* 35179 (Strid) |
| *Silene caesia* Sm. | Baden 1114 (Strid) |
| *Silene caryophylloides* (Poir) Otth | Görk *et al.* 2436 (Strid) |
| *Silene ciliata* Pourr. | Franzén *et al.* 822 (Strid) |
| *Silene conica* L. | P. Erixon 70 (UPS) |
| *Silene conoidea* L. | A. Rautenberg 290 (GB) |
| *Silene cordifolia* All. | Lippert & Merxmüller 17265 (Strid) |
| *Silene davidii* (Franch.) Oxelman & Lidén | F. Eggens 85 (UPS) |
| *Silene delicatula* Boiss. | B. Oxelman 2456 (GB) |
| *Silene dichotoma* Ehrh. | W. Till 17.7.2004 (WU) |
| *Silene douglasii* var. *oraria* (M.Peck) C.L.Hitchc. & Maguire | *N/A. Collected by S. Kephart. Cascade Head, OR, USA |
| *Silene flavescens* Waldst. & Kit. | Strid & Papanicolaou 15820 (Strid) |
| *Silene fruticosa* L. | B. Oxelman & L. Tollsten 934 (GB) |
| *Silene gallica* L. | D. Sloan 002 (VPI) |
| *Silene gallinyi* Heuff. ex Rchb. | Strid & Hansen 9283 (Strid) |
| *Silene gracilicaulis* C.L.Tang | Smith 11346 (UPS) |
| *Silene hookeri* Nutt. subsp. *hookeri* | F. Schwartz 107 (WTU) |
| *Silene imbricata* Desf. | B. Oxelman 1881 (GB) |
| *Silene integripetala* Bory & Chaub. | B. Oxelman 1902 (GB) |
| *Silene involucrata* (Cham. & Schltdl.) Bocquet | F. Eggens 7 (UPS) |
| *Silene khasyana* Rohrb. | Einarsson et.al 3025 (UPS) |
| *Silene lacera* (Stev.) Sims | Schönswetter & Tribsch Iter Georgicum 51 (WU) |
| *Silene laciniata* subsp. *californica* (Durand) J.K.Morton | Schwartz 102-2 (WTU) |
| *Silene latifolia* Poir. | *N/A. Collected by J. Greimler. Vienna, Austria |
| *Silene littorea* Brot. | P. Erixon 74 (UPS) |
| *Silene macrodonta* Boiss. | B. Oxelman 2441 (GB) |
| *Silene menziesii* Hook. | Kruckeberg 3436 (WTU) |
| *Silene moorcroftiana* Wall. ex Benth | B. Dickoré 17783 (Dickoré) |
| *Silene multicaulis* Guss. | Strid & Hansen 9954 (Strid) |

| | |
|---|---|
| *Silene muscipula* subsp. deserticola Murb. | *Chevalier 548 (WU) |
| *Silene nana* Kar. & Kir. | Kereverzova & Mekeda 1976.V.5 (LECB) |
| *Silene nicaeensis* All. | D. Sloan 005 (VPI) |
| *Silene noctiflora* L. | D. Sloan 003 (VPI) |
| *Silene nutans* L. | *Larsen, Larsen & Jeppesen 196 (S) |
| *Silene odontopetala* Fenzl | Görk *et al.* 23817 (Strid) |
| *Silene otites* (L.) Wibel | A. Rautenberg 83 (UPS) |
| *Silene paradoxa* L. | W. & S. Till 21 July 2002 (WU) |
| *Silene paucifolia* Ledeb. | H. Solstad & R. Elven 04/1384 (O) |
| *Silene pendula* L. | A. Rautenberg 289 (GB) |
| *Silene pygmaea* Adams | Amirkhanov 22.VI-1977 MW) |
| *Silene repens* Patrin | Argus 1068 (UPS) |
| *Silene sachalinensis* F.Schmidt | Popov 1949.VII.8 (LE) |
| *Silene samia* Melzh. & Christod | B. Oxelman 2208 (UPS) |
| *Silene samojedora* (Sambuk) Oxelman | H. Solstad, R. Elven SUP-04-3871 (O) |
| *Silene schafta* S.G.Gmel. ex Hohen. | M. Popp 1053 (UPS) |
| *Silene schwarzenbergeri* Halácsy | Hartvig & Christiansen 8167 (Strid) |
| *Silene seoulensis* Nakai | Hong & Han 13420001 (UPS) |
| *Silene sordida* Hub.-Mor. & Reese | B. Oxelman 2206 (GB) |
| *Silene sorensenis* (B.Boivin) Bocquet | F. Eggens 48 (UPS) |
| *Silene stellata* (L.) W.T.Aiton | N/A. Collected by D. Sloan. Giles County, VA, USA |
| *Silene succulenta* Forssk. | Strid & Kit Tan 55028 (Strid) |
| *Silene tunicoides* Boiss. | Carlström 5970 (Strid) |
| *Silene turkestanica* Regel | K. Kiseleva 20.VI.1970 (MW) |
| *Silene uniflora* Roth | P. Erixon 73 (UPS) |
| *Silene vittata* Stapf | B. Oxelman 2390 (UPS) |
| *Silene vulgaris* (Moench) Garcke | *N/A. Collected by M. Dzhus. Minsk, Belarus |
| *Silene yemensis* Deflers | Hepper 5792 (WU) |
| *Silene zawadzkii* Herbich | B. Oxelman 2241 (GB) |
| *Viscaria alpina* (L.) G.Don | B. Frajman & P. Schönswetter 11415 (LJU) |
| *Viscaria vulgaris* Bernh. | P. Schönswetter & B. Frajman 11097 (LJU) |

Herbaria abbreviations are from Holmgren *et al.* [75], except Dickoré (private herbarium Bernhard Dickoré, Göttingen, Germany), and Strid (private herbarium Arne Strid, Ørbaek, Denmark). Asterisks indicate that a second specimen was used for one or more loci [see Additional file 9].

**Table 2 - Absolute substitution rates by gene (SSB).**

| | $R_N$ | $R_S$ | $\omega$ |
|---|---|---|---|
| *nad9* (378 bp) | 0.36 (0.18) | 2.62 (0.51) | 0.137 (0.357) |
| *cox3* (588 bp) | 0.38 (0.22) | 3.43 (1.64) | 0.110 (0.133) |
| *atp1* (960 bp) | 0.20 (0.13) | 4.25 (2.38) | 0.048 (0.055) |
| *atp9* (162 bp) | 0.39 (0.41) | 22.66 (20.75) | 0.017 (0.020) |

Values represent tree-wide average rates (total branch length divided by total branch time) based on the subset of 61 species for which we have sequence for all 4 mitochondrial genes. The values in parentheses are the rates estimated after excluding the two clades with highly elevated rates across all 4 genes.

**Table 3 - Pairwise $R_N$ and $R_S$ correlation coefficients within and among genes across phylogenetic lineages.**

| | *nad9* | *cox3* | *atp1* | *atp9* | *matK* |
|---|---|---|---|---|---|
| *nad9* | **0.86** | **0.88** | **0.73** | 0.04 | 0.17 |
| *cox3* | **0.52** | **0.49** | **0.77** | 0.11 | 0.15 |
| *atp1* | **0.39** | **0.65** | **0.70** | 0.13 | 0.13 |
| *atp9* | -0.11 | -0.12 | -0.12 | 0.12 | -0.05 |
| *matK* | 0.03 | -0.03 | -0.02 | 0.13 | 0.28 |

Values above and below diagonal are from pairwise comparisons between genes for $R_S$ and $R_N$, respectively. Values on the diagonal are correlation coefficients between $R_N$ and $R_S$ within each gene. Bold values are significantly different from 0 based on a Bonferroni corrected $\alpha$ of 0.05/25 = 0.002.

# Additional files

**Additional file 1**

File Format: XLS

**Title: Estimated divergence times (in millions of years) from three different dating methods.**

Description: [See Additional file 2 for definitions of node names].

**Additional file 2**

File Format: PDF

**Title: Names for internal nodes**

Description: The labels to the right of each node correspond to the names used in Additional files 1, 3, and 4.

**Additional file 3**

File Format: XLS

**Title: Detailed data on $d_N$, $d_S$, $R_N$, $R_S$ and associated error for each gene and concatenated dataset (all species).**

Description: Each row corresponds to a phylogenetic branch defined by its basal node and derived node/tip. $R_N$ and $R_S$ values are in terms of SSB. Approximated standard errors (SEs) are provided for $R_N$ and $R_S$. Note that SE approximations are undefined and reported as 0 for any absolute rate estimate of 0. This table does not include *atp9* because it was not sequenced in all species. [See Additional file 2 for definitions of node names].

**Additional file 4**

File Format: XLS

**Title: Detailed data on $d_N$, $d_S$, $R_N$, $R_S$ and associated error for each gene and concatenated dataset (*atp9* subset).**

Description: Same setup as Additional file 2. Only the 61 species for which *atp9* was sequenced are included, so that data for *atp9* and the concatenation of all four mitochondrial genes could be presented. [See Additional file 2 for definitions of node names].

**Additional file 5**

File Format: PDF

**Title: Maximum likelihood trees for each of the 4 mitochondrial genes (generated without topological constraint).**

Description: Parsimony bootstrap values are noted to the left of the corresponding node. Only values > 0.5 are shown. Branch lengths are in terms of substitutions per site.

**Additional file 6**

File Format: PDF

**Title: Maximum likelihood tree for *matK* dataset.**

Description: Branch lengths are in terms of substitutions per site.

**Additional file 7**

File Format: PDF

**Title: BEAST analysis of *matK* dataset with unconstrained topology.**

Description: Time scale is in millions of years. Posterior support is shown to the right of each node.

**Additional file 8**

File Format: XLS

**Title: Sequences and references for PCR primers**

Description: Nucleotide sequences (5' to 3') of primers used for PCR amplification and DNA sequencing.

**Additional file 9**

File Format: XLS

**Title: GenBank accession numbers for all sequences.**

Description: Accession numbers in bold are not from the voucher listed in Table 1.

**Additional file 10**

File Format: ZIP

**Title: TableS3**

Description: Alignments for each gene in FASTA format.

# Chapter 5.

# Extensive loss of translational genes in the structurally dynamic mitochondrial genome of *Silene latifolia*[1]

[1]Formatted as a co-authored manuscript and published as:

Sloan DB, Alverson AJ, Štorchová H, Palmer JD, Taylor DR. 2010. *BMC Evol. Biol.* 10:274.

Referenced supplementary material is available online at:

http://www.biomedcentral.com/1471-2148/10/274

# Abstract


## Background

Mitochondrial gene loss and functional transfer to the nucleus is an ongoing process in many lineages of plants, resulting in substantial variation across species in mitochondrial gene content. The Caryophyllaceae represents one lineage that has experienced a particularly high rate of mitochondrial gene loss relative to other angiosperms.

## Results

In this study, we report the first complete mitochondrial genome sequence from a member of this family, *Silene latifolia*. The genome can be mapped as a 253,413 bp circle, but its structure is complicated by a large repeated region that is present in 6 copies. Active recombination among these copies produces a suite of alternative genome configurations that appear to be at or near "recombinational equilibrium". The genome contains the fewest genes of any angiosperm mitochondrial genome sequenced to date, with intact copies of only 25 of the 41 protein genes inferred to be present in the common ancestor of angiosperms. As observed more broadly in angiosperms, ribosomal proteins have been especially prone to gene loss in the *S. latifolia* lineage. The genome has also experienced a major reduction in tRNA gene content, including loss of functional tRNAs of both native and chloroplast origin. Even assuming expanded wobble-pairing rules, the mitochondrial genome can support translation of only 17 of the 61 sense codons, which code for only 9 of the 20 amino acids. In addition, genes encoding 18S and, especially, 5S rRNA exhibit exceptional sequence divergence relative to other plants. Divergence in one region of 18S rRNA

appears to be the result of a gene conversion event, in which recombination with a homologous gene of chloroplast origin led to the complete replacement of a helix in this ribosomal RNA.

**Conclusions**

These findings suggest a markedly expanded role for nuclear gene products in the translation of mitochondrial genes in *S. latifolia* and raise the possibility of altered selective constraints operating on the mitochondrial translational apparatus in this lineage.

## Background

The mitochondrial genomes of flowering plants exhibit a number of characteristics that distinguish them from the mitochondrial genomes of other eukaryotes [1]. They are large and variable in size with ample non-coding content [2], including substantial amounts of "promiscuous" DNA of nuclear and chloroplast origin [3, 4] as well as sequences of horizontal origin acquired from the mitochondrial genomes of other land plants [5, 6]. Angiosperm mitochondrial genomes also contain numerous introns, some of which have been split such that the resulting gene fragments must be transcribed separately and then *trans*-spliced together [7]. Gene expression also relies on extensive C-to-U (and sometimes U-to-C) RNA editing, in which substitution of specific pyrimidines in the mRNA sequence restores phylogenetically conserved codons [8]. Plant mitochondrial genomes generally experience some of the slowest documented rates of nucleotide substitution [9, 10] but are subject to rapid structural evolution [11]. High frequency intra- and intermolecular recombination among large repeated sequences is the rule, generating a heterogeneous pool of genome configurations within a single individual [12-14]. The size and complexity of plant mitochondrial genomes, especially when compared with animals and fungi, make them powerful models for exploring the forces affecting eukaryotic genome structure and evolution.

The genomes of plant mitochondria, like any organelle genome, depend on highly integrated functional coordination with the nucleus. For example, translation of mitochondrially-encoded genes requires a mix of nuclear and mitochondrially encoded components. Plant mitochondrial genomes contain genes for their own rRNA

subunits as well as for some of the ribosomal proteins and tRNAs required for translation (Figure 1), but many necessary ribosomal protein and tRNA genes are located in the nuclear genome, so their gene products must be imported into the mitochondrion [15]. The tRNA population within plant mitochondria represents a particularly complex assemblage derived from at least 3 anciently divergent classes of genes [15-17]: 1) "native" tRNAs encoded in the mitochondrial genome and inherited from the α-proteobacterial progenitor of mitochondria, 2) chloroplast-like tRNAs, which are also encoded in the mitochondrial genome but which were acquired by functional gene transfer from the chloroplast genome during land plant evolution, and 3) nuclear-encoded tRNAs imported from the cytosol.

This mixture of tRNA genes is phylogenetically dynamic. Sequenced plant mitochondrial genomes differ in both the number and the identity of tRNA genes that they contain (Figure 1) [4, 18]. Likewise, ribosomal protein gene content in the mitochondrial genome is highly variable among plant lineages. The process of mitochondrial gene loss and functional transfer to the nucleus is active and ongoing in plants, and 15 of the 17 protein genes that have been subject to frequent loss across the angiosperm phylogeny encode ribosomal proteins [19-21].

The Caryophyllaceae represents one angiosperm lineage with a relatively high rate of mitochondrial gene loss/transfer. Adams et al. [19] used Southern blots to show that 2 genera from this family (*Dianthus* and *Stellaria*) lack most mitochondrial protein genes outside the core set of 24 genes that are nearly universally conserved throughout angiosperms, and we recently reported that 2 species from a third genus (*Silene*) are similarly reduced in gene content [22]. The genus *Silene* is of particular interest with respect to mitochondrial genome evolution and transmission [23, 24].

This large genus exhibits substantial diversity in breeding system, including a high frequency of gynodioecy (mixed populations of hermaphrodites and females), which is often the result of mitochondrial mutations that induce cytoplasmic male sterility [25]. Furthermore, *Silene* species differ markedly in mitochondrial mutation rate [10, 26-28] and in the amount of mitochondrial sequence polymorphism that they maintain [26, 27, 29, 30]. Previous analyses of *Silene* mitochondrial genomes, however, have been limited to individual gene sequences.

In this study, we report the complete mitochondrial genome sequence of *Silene latifolia*, which confirms earlier findings of reduced mitochondrial protein gene content in the Caryophyllaceae. We also found a reduction in tRNA gene content that is unprecedented in plants as well as a major increase in the substitution rate for some rRNA genes. In addition, we use paired-end sequence data and Southern blot hybridizations to analyze the complex structural dynamics of this genome, which are driven by a large recombining repeat sequence that is present in 6 copies. These methods could be used more broadly to explore the complex dynamics of mitochondrial genomes in established plant model systems.

# Methods

### Study Species and Plant Material

*Silene latifolia* Poir. (Caryophyllaceae) is a short-lived, herbaceous perennial that is widespread in its native Eurasia [31]. Frequently associated with human disturbance, it is also introduced and invasive in other regions, including North America [32]. Like other members of *Silene* section *Elisanthe*, *S. latifolia* has a dioecious breeding system with XY chromosomal sex determination [25, 33].

We grew seeds from a single maternal family in the greenhouse. These seeds were collected by D.R. Sowell from a common garden experiment in Oxford, England, but the maternal plant was derived from seed originally collected on the Apple Orchard Falls Trail in Bedford County, Virginia, USA. A voucher specimen from this family was deposited in the Massey Herbarium at Virginia Polytechnic Institute and State University (D. Sloan #004). Fifteen weeks after the seeds were sown, we harvested 500 g of flowers and fresh green leaves from a total of 550 plants.

**Mitochondrial DNA Extraction, Sequencing, Assembly, and Finishing**

We followed previously published protocols for plant organelle DNA extraction [34, 35], which yielded approximately 4 μg of mitochondrial DNA. We confirmed the purity of the DNA by digesting a 100 ng sample with PstI and observing a well-defined electrophoretic banding pattern on an agarose gel.

Library construction, cloning, shotgun sequencing, and genome assembly were performed by the Genome Center at Washington University in St. Louis. The genomic DNA was fragmented using a Hydroshear (Digilab; Holliston, MA), end polished, and run on a 0.8% agarose gel. A fraction of that gel corresponding to a 4-6.5 kb size range was excised, purified and ligated into the pSMART vector system (Lucigen; Middleton, WI). After transformation, 2688 subclones were purified, cycle sequenced from both ends with BigDye v3.1 (Applied Biosystems; Foster City, CA), and analyzed on an ABI 3730 capillary sequencer, providing an average of 7x genome sequence coverage.

Shotgun sequence data were assembled with Phrap followed by manual sorting in Consed to resolve misassemblies [36]. Assembly gaps were closed by sequencing subclones with paired-end reads that mapped to the ends of adjacent

contigs. Regions with low quality or single read coverage were augmented by PCR and Sanger sequencing of total cellular DNA.

**Genome Annotation**

Protein, rRNA, and tRNA genes as well as regions of chloroplast origin were identified using BLAST and tRNAscan-SE as described previously [37]. Regions that were not annotated as belonging to one of these categories were used to search against the NCBI non-redundant nucleotide and protein databases (nt/nr) with BLASTN (r = 5, q = -4, G = 8, E = 6, W = 7, and e = 0.001) and BLASTX (netblast v2.2.19 default parameters except e = 0.001). Perfectly repeated sequences were identified with REPuter [38]. The annotated genome sequence was deposited in GenBank (HM562727).

**Sequence Analysis**

Previous studies have shown substantial variation in substitution rates among mitochondrial genes within the genus *Silene* [26, 28]. To quantify differences in substitution rate, we analyzed individual protein and rRNA genes in a phylogenetic context with PAML v4.1 [39]. For each gene, we included sequences from 18 seed plant species for which complete mitochondrial genome sequences are available. In these analyses, phylogenetic relationships among the species were constrained according to previous studies [40, 41]. For protein genes, branch lengths were estimated in terms of both synonymous and non-synonymous substitutions per site with the program codeml as described previously [28]. For rRNA genes, branch lengths were estimated in terms of substitutions per site with the program baseml. We employed a K80 (Kimura 2-parameter) model of substitution for *rrn5* (5S rRNA) and

an HKY model for *rrn18* (18S rRNA) and *rrn26* (26S rRNA). For all 3 genes, we modeled rate variation among sites with a gamma distribution. These substitution models were chosen based on the results of likelihood ratio tests between pairs of competing models. Because the annotated boundaries of rRNA genes differ slightly across species, we trimmed all sequences to the shortest annotated length.

Our analysis revealed a substantial elevation in substitution rate for *rrn5* in *S. latifolia*. To determine the structural consequences of these substitutions we used the RNAeval program within the Vienna RNA Package v1.8.4 [42] to calculate the free energy of the predicted secondary structure for plant mitochondrial 5S rRNA [43, 44]. To test for selection for conservation of secondary structure in *S. latifolia*, we generated 10,000 sequences by randomly placing 16 substitutions (the number observed in *S. latifolia*) into the *Beta vulgaris rrn5* sequence. *Beta vulgaris* was chosen because it is the most closely related species with an available *rrn5* sequence, and it appears to have maintained the ancestral sequence of core eudicots. We compared the free energy of the conserved 5S rRNA secondary structure for *S. latifolia* to the distribution of values from the 10,000 simulated sequences to determine whether the *S. latifolia* structure was more highly conserved than expected by chance.

**Southern Blot Hybridizations**

We used Southern blots to assess the existence and relative abundance of alternative genome conformations resulting from intramolecular recombination between large repeated sequences. Total cellular DNA was purified from individual fresh leaves using a sorbitol extraction method [45]. Samples were taken from 2 individuals from each of 2 full-sib families. Each of these families was generated by crossing a female

from the family used for genome sequencing with a male from an unrelated family. Between 0.5 and 1 μg of genomic DNA was digested with EcoRI (HF enzyme, New England BioLabs), electrophoresed overnight on a 0.9% agarose gel, and transferred to a positively charged nylon membrane (Roche) by capillary blotting. Two probes were generated to target single copy regions flanking large repeated sequences. The probes correspond to genomic positions 140,389-141,463 nt ("left") and 5636-6500 nt ("right") and were generated with the following PCR primers: LeftF1 5'- AGTCTGCCTTTGTCCGACTG; LeftR1 5'-TCCCCTTGGGGTTCTTATCT; RightF2 5'-TCTTTCTTTGCGCTTTCGAT; RightR2 5'-CATTGGCCTTTGCTTCCTT. The probes were labeled with digoxigenin (DIG) using Roche's PCR labeling kit. The genomic blots were hybridized in an EasyHyb buffer (Roche) with the DIG-labeled probe at 42° C overnight, washed at high stringency (0.1x SSC, 65° C), and detected using CDPStar (Roche). An exposure time of 5 to 20 minutes was sufficient to achieve clear bands on ECL film (Kodak). Preliminary data showed that, when amplified directly from genomic DNA, the "right" probe yielded non-specific hybridization, so the PCR fragment was cloned in pGEM T Easy vector (Promega). The resulting plasmid was used as a template to generate the probe with the same PCR primers. The "left" probe was amplified directly from genomic DNA.

# Results

### Genome Size and Organization

The sequenced *S. latifolia* mitochondrial genome can be mapped as a 253,413 bp "master" circle with a total complexity of 244,058 bp if only a single copy of perfectly repeated sequences greater than 100 bp is included (Figure 2). The majority of the repeated sequence in the genome is represented by a 1362 bp "core" repeat sequence that is present in 6 identical copies, all of which are in the same (i.e., direct) orientation relative to each other. Most of the remaining repeated sequence is found in "extensions" of the core repeat. The extensions are identical stretches of sequence between 12 and 1593 bp shared by 2 or more (but not all 6) of the flanking sequences on either the "left" or "right" side of the core repeat (Figure 3). With the exception of this 6-copy repeat and its extensions, the *S. latifolia* mitochondrial genome is relatively devoid of repeated sequences, containing only 2 other repeat families greater than 100 bp (123 bp and 167 bp). Each of these is a 2-copy repeat.

The master circle depicted in Figure 2 represents only one of many possible genome conformations. No single circle is fully consistent with all the sequencing reads because there are numerous paired-end conflicts, i.e., cases where 2 reads from the same subclone map too far apart or in the wrong orientation. With default filtering settings in Consed, these conflicts are exclusively associated with the large 6-copy repeat sequence, suggesting active intra- and intermolecular recombination among repeats [13, 14].

With 6 copies of the core repeat, there are 36 possible pairs of flanking sequences if all core repeats recombine with each other. Of these, 26 pairs are

supported by multiple subclones from our shotgun sequence data. The lack of evidence for the remaining 10 flanking pairs could reflect a reduced frequency or complete absence of these recombination products, but it may also be the result of stochastic sampling and/or cloning bias given our relatively low (7x) sequencing coverage. To distinguish between these possibilities, we first performed (non-quantitative) PCR with all possible pairwise combinations of primers designed for the left and right single-copy regions that flank each core repeat plus its repeat extensions. We detected all 36 possible flanking sequence pairs in DNA extracted from a single leaf (data not shown). We then utilized Southern blots to assess the relative abundance of the various recombination products and confirm that the results from the PCR experiment were not simply an artefact of PCR-mediated recombination [46]. We separately hybridized probes representing one "left" single-copy flanking sequence and one "right" single-copy flanking sequence (Figure 3) to genomic DNA digested with EcoRI. In each case, we detected 6 strong bands, corresponding to the expected sizes of the 6 possible recombination products (Figure 4; Additional File 1). All 6 bands are of similar intensity, suggesting that the alternative conformations of the *S. latifolia* mitochondrial genome exist at relatively equal frequencies. The "right" probe also unexpectedly hybridized to a 1.8 kb fragment, producing a seventh, fainter band that was present in a subset of the individuals (Figure 4). Studies are ongoing to assess the possibility that this seventh band reflects the existence of sublimons and substoichiometric shifting in *S. latifolia* [47, 48].

**Gene Content**

*Protein Genes.* The *S. latifolia* mitochondrial genome contains intact and putatively functional copies for all 24 of the protein genes that are nearly universally conserved

across the large sample of angiosperm mitochondrial genomes examined to date (Figure 1) [19]. In contrast, the genome appears to lack functional copies for most of the 17 other protein genes that were ancestrally present in angiosperm mitochondrial genomes, but which have been subsequently lost, and for the most part, functionally transferred to the nucleus, many times during the course of angiosperm evolution [19, 20].

Eleven of these 17 genes have little or no remnant in the genome, while most of the other 6 genes (*rpl5*, *rps3*, *rps4*, *rps13*, *rps14*, and *sdh3*) appear to be pseudogenes. Of this group, only *rpl5* is fully intact relative to other angiosperms. It is possible that *rps3* and *rps14* are functional, but both of these genes show evidence of degeneration. The first exon (75 bp) of *rps3* has been lost, and the much larger second exon (1773 bp) exhibits a substantial 3' extension before the first in-frame stop codon relative to other angiosperms. The 5' portion of *rps14* is altered by a frameshift mutation that is corrected after 45 bp by a second frameshift indel. The remaining genes either lack substantial regions that are conserved in other angiosperms (*rps4*) or are truncated by internal stop codons (*rps13* and *sdh3*). Based on these results, we have identified putatively functional genes and pseudogenes in Figure 1, though a more definitive classification will require detailed analysis of gene expression and function. Regardless, it is apparent that the *S. latifolia* mitochondrial genome has lost a large fraction of the protein genes that were part of the ancestral angiosperm mitochondrial genome.

*tRNA Genes*. The *S. latifolia* mitochondrial genome contains substantially fewer tRNA genes than any angiosperm mitochondrial genome sequenced to date. A search

of the genome with BLASTn and tRNAscan-SE identified only 11 tRNA genes, and at least 2 of these (*trnP-cp* and *trnM-cp*) are potential pseudogenes based on the presence of multiple substitutions and insertions in their anticodon loops. A third gene (*trnfM*) shows an elevated substitution rate, but its anticodon and secondary structure appear largely intact (Additional File 2). Five of the 11 genes (including both potential pseudogenes) are of chloroplast origin, representing apparently ancient cpDNA transfers that pre-date the divergence between *Silene* and *Beta*. Collectively, the genes encode a set of tRNAs that, even after including the potential pseudogenes and assuming expanded wobble pairing rules [49, 50], can translate only 17 of the 61 sense codons, encoding only 9 of the 20 amino acids (Table 1). By comparison, mitochondrially-encoded tRNAs in *Beta vulgaris* (the most closely related species with a complete mitochondrial genome sequence) can potentially recognize 35 codons, encoding 16 amino acids. Therefore, it is likely that an unusually large fraction of the *S. latifolia* mitochondrial tRNA population is encoded in the nuclear genome and imported from the cytosol.

*rRNA Genes*. Like other angiosperm mitochondrial genomes, the *S. latifolia* genome contains genes encoding 3 ribosomal RNA species (*rrn5*, *rrn18*, and *rrn26*). Two divergent copies of the *rrn5* gene are present, although one is likely non-functional, exhibiting 3 substantial insertions (7, 9, and 16 bp) and multiple substitutions that greatly reduce the stability of the widely conserved 5S rRNA secondary structure (see below).

**Intron and RNA Editing Content**

The *S. latifolia* mitochondrial genome contains a total of 19 group II introns, 6 of which are *trans*-spliced. All 19 introns are found in protein genes, and all but one occur in genes that encode subunits of complex I (NADH dehydrogenase). The *S. latifolia* lineage has lost the second *nad4* intron and both of the *cox2* introns found in other angiosperms [51]. It also lacks the group I intron in *cox1*, which has been widely distributed across the angiosperm phylogeny by numerous horizontal transfer events [52]. A previous study identified a total of 287 C-to-U RNA editing sites within the genome's protein genes, which is fewer than typically found in angiosperm mitochondrial genome but substantially more than observed in the rapidly evolving congeners *S. noctiflora* and *S. conica* [22].

**Intergenic Regions**

A BLAST search of intergenic regions from the *S. latifolia* mitochondrial genome found that 46.2 kb (23.1%) of this sequence exhibits significant similarity to other land plant mitochondrial genomes (after excluding sequences of clear chloroplast origin). Much of this conserved sequence is directly flanking annotated genes and likely represents regulatory elements, UTRs and *trans*-spliced introns [37]. The genome also contains 2 open reading frames (ORFs) related to the DNA and RNA polymerase genes found on linear mitochondrial plasmids in angiosperms and other eukaryotes [53]. These polymerase genes have also been integrated into the mitochondrial genomes in a number of other angiosperms [54, 55].

By searching the complete mitochondrial genome sequence against a collection of diverse chloroplast genomes, we identified a total of 2462 bp of apparent chloroplast origin distributed in 9 fragments ranging in size from 43 to 588 bp. The total chloroplast contribution represents 1.0% of the genome, which is on the low end

of the range of approximately 1 to 12% detected in other sequenced angiosperm mitochondrial genomes [37, 56]. As found in other angiosperms, the *S. latifolia* mitochondrial genome also contains numerous sequences of apparent nuclear origin, including many regions with homology to (presumably inactivated) angiosperm transposable elements. Nevertheless, based on our search criteria, more than 143.5 kb of intergenic sequence (a full 56.6% of the genome) lacks detectable homology with any DNA or protein sequence in the NCBI nt/nr databases.

**Nucleotide Composition and Codon Usage**

The *S. latifolia* mitochondrial genome has a 42.6% GC content, which is slightly below the range of 42.8% to 45.2% observed in other sequenced angiosperm mitochondrial genomes [37, 57]. The patterns of codon usage in protein genes (Additional File 3) are very similar to other angiosperm mitochondrial genomes [57], despite the significant changes in tRNA gene content in the *S. latifolia* genome.

**Substitution Rates**

Based on a phylogenetic analysis of 18 complete plant mitochondrial genomes, *Silene latifolia* consistently shows higher substitution rates than its sister lineage, *Beta vulgaris* (Figure 5). For the most part, these differences are minor, and the substitution rates in *S. latifolia* are consistent with the low rates that generally characterize plant mitochondrial genomes [10, 28]. There are, however, 2 notable outliers with more extreme elevations in substitution rate: the protein gene *atp9* and the putatively functional copy of the ribosomal rRNA gene *rrn5* (Figure 6). Elevated substitution rates for *atp9* have previously been reported throughout *Silene* [28], but this study represents the first analysis of *rrn5* in the genus.

Despite their elevated substitution rates, both *atp9* and *rrn5* exhibit evidence of purifying selection, suggesting that they are still functionally expressed in the mitochondria. While the observed synonymous substitution rate in *atp9* is more than 5-fold higher than in any other protein gene in *S. latifolia*, this is the only gene without a single inferred non-synonymous substitution, suggesting strong purifying selection on amino acid sequence (Figure 6; note that *atp9*, at 225 nt in length, is the shortest protein gene in the genome). In the case of the ribosomal rRNA gene *rrn5* (ca. 111 nt), 13 of the 16 inferred substitutions occur in loops within the conserved secondary structure (Figure 7) [43, 44]. Moreover, the 3 substitutions within helices are structurally conservative. Two of those substitutions compensate for each other by altering both bases in a single pairing, resulting in a C:G to G:C change at positions 27:56 (Figure 6). The third substitution found at a conserved helix position (A-to-G at position 98) should still allow for base pairing (G:U instead of A:U). The one predicted change in secondary structure in *S. latifolia* results from a T-to-G substitution at position 34. This position normally represents the first base of the terminal loop on that branch, but the substitution should allow it to pair with $C_{46}$ and extend the preceding helix (Figure 7). As a result, the predicted secondary structure is slightly more stable in *S. latifolia* ($\Delta G$ = -40.80) than in other angiosperms (e.g., *Beta vulgaris*; $\Delta G$ = -39.16). A simulation test that randomly placed mutations in *rrn5* showed that, given the number of substitutions in *S. latifolia*, the conservation of secondary structure is much stronger than expected by chance ($p < 0.0001$). Therefore, it appears that, despite its elevated substitution rate in *S. latifolia*, *rrn5* is still under selection to maintain folding stability. In contrast, a second *rrn5* copy in *S.*

*latifolia* is likely a pseudogene, as it contains 3 insertions as well as 3 nucleotide substitutions that disrupt conserved base pairing in helices ($\Delta G$ = -17.58).

### Gene Conversion Between Mitochondrial and Chloroplast Sequences

The distribution of substitutions contributing to the elevated *rrn18* divergence in *S. latifolia* is noticeably clustered (Figure 8a). One cluster of substitutions is likely the result of a gene conversion event in which a segment of at least 47 bp of *rrn18* sequence was converted by a homologous chloroplast *rrn16* gene (Figure 8b). The boundaries of this apparent conversion tract correspond precisely to the beginning and end of helix 240 (domain I) in the secondary structure model for 16S rRNA in *Escherichia coli* [58]. Therefore, the result of the gene conversion appears to have been a clean exchange of the entirety of this helix. The region appears to have been further modified by multiple substitutions and indels since the conversion event. Evidence of this conversion is also present in *S. vulgaris*, but not in *S. acaulis*, indicating that it occurred after the split between the two *Silene* subgenera but before the divergence of the major lineages in subgenus *Behenantha* [28]. We did not find evidence of cpDNA-mediated conversion in any other *Silene* mitochondrial genes, including the rapidly evolving *rrn5* and *atp9* genes.

## Discussion

### Mitochondrial Gene Loss

The vast majority of genes in plant mitochondrial genomes can be placed into one of two functional categories: 1) bioenergetics, i.e., oxidative phosphorylation and ATP

synthesis (*atp*, *ccm*, *cob*, *cox*, *nad*, and *sdh* genes) and 2) translational machinery (ribosomal protein, rRNA, and tRNA genes). Analysis of the phylogenetic distribution of protein genes across seed plants has clearly shown that ribosomal proteins are subject to more rapid rates of loss than genes involved in bioenergetics [19]. The complete sequence of the *S. latifolia* mitochondrial genome provides the first evidence that mitochondrial tRNA genes, another component of the organelle's translational machinery, can also be lost rapidly and in large numbers in plants. This finding is consistent with broader patterns in eukaryotic evolution, as numerous independent lineages have experienced the loss of most or even all of their mitochondrially-encoded tRNAs [59]. The present study also extends earlier work that found reduced protein gene content in 2 other genera in the Caryophyllaceae [19]. The similar reduction in protein gene content in these 3 taxa suggests that much of the observed protein gene loss probably occurred prior to the diversification of this family, although some degree of parallel loss within the family is also possible.

Protein genes that are lost from mitochondrial genomes can experience a variety of fates. For example, the evolutionary history of eukaryotes has been characterized by a massive physical transfer of genes from the mitochondrial genome to the nucleus. This process is ongoing in plants, and there are a number of well-established cases of such endosymbiotic gene transfer that have occurred since the divergence of angiosperms [60-63]. Losses can also occur when a gene is functionally replaced by an anciently divergent homolog [20, 21, 64, 65], and when a protein or even an entire multi-subunit complex is no longer functionally required (e.g., the loss of the NADH dehydrogenase complex I in apicomplexans and at least 2 yeast lineages [66, 67]). In *Silene*, an analysis of the *S. vulgaris* transcriptome (unpublished data)

revealed evidence of nuclear copies for at least 9 of the protein genes that appear to have been functionally lost from the *S. latifolia* mitochondrial genome.

To the best of our knowledge, a functional transfer of a mitochondrial tRNA gene to the nucleus has never been documented. Instead, mitochondrial tRNA gene loss is typically offset by importing tRNAs of eukaryotic nuclear origin from the cytosol [17, 59]. Therefore, it is likely that *Silene* mitochondria import a greatly expanded set of nuclear tRNAs relative to other plants—a prediction that could be tested by purifying and sequencing *Silene* organelle tRNAs.

In some specific cases, however, more complex evolutionary changes may be required to explain the loss of mitochondrially-encoded tRNAs. For example, in plant mitochondria, the function of tRNA-Gln is dependent on coordinated enzymatic processes. Aminoacyl tRNA synthetases play an essential role in translation by matching tRNAs with their corresponding amino acids, but plant organelles generally lack a Gln tRNA synthetase. Instead, tRNA-Gln is typically aminoacylated by a Glu tRNA synthetase followed by a chemical modification (amidation) to convert Glu to Gln [68].

The gene encoding tRNA-Gln (*trnQ*) is present in all sequenced seed plant mitochondrial genomes with the exception of *S. latifolia*. The loss of the mitochondrially-encoded copy of tRNA-Gln in *S. latifolia* raises several possibilities. First, it is conceivable that aminoacylation and amidation are carried out in the same fashion with an imported cytosolic tRNA-Gln. This may be unlikely, however, because it would require associated changes in tRNA recognition for multiple enzymes. Second, it is possible that, unlike other plants, *S. latifolia* imports the cytosolic Gln tRNA synthetase into its mitochondria, allowing for direct

aminoacylation of an imported tRNA-Gln without the use of a Glu intermediate. Finally, it is possible that *S. latifolia* has experienced an unprecedented transfer of a functional tRNA gene (*trnQ*) from the mitochondrial genome to the nucleus, where it is expressed and its product targeted back to the mitochondria. All of these possibilities should be investigated to better understand the mechanisms involved in the co-evolution of organellar and nuclear gene content.

It is intriguing that extensive gene loss in two components of *Silene* mitochondrial translation machinery has been associated with accelerated evolutionary rates in a third component, rRNA genes. This pattern raises the possibility of a correlated reduction in functional constraint across these 3 translational components. A general relaxation of selection on organelle translation has been observed in cases such as the chloroplasts of non-photosynthetic plants where the organelle's functional role has been greatly reduced [69]. However, we have no *a priori* reason to expect relaxed selection on mitochondrial gene expression in *Silene*, and the distribution of substitutions in *rrn5* suggests that its secondary structure is under strong selection to maintain function. Broader comparative and functional analyses would be of value in assessing the extent to which correlated evolutionary pressures act on these 3 components of mitochondrial translation machinery.

An alternative interpretation of our results is that, rather than being lost, certain genes have been functionally retained in the mitochondrial genome but escaped detection by our annotation methods. For example, cryptic genes could result from accelerated rates of evolution or the proliferation of introns and RNA editing sites [18, 70]. Although these explanations are unlikely given the generally slow rate

of plant mtDNA sequence evolution and the trend towards a reduced frequency of introns and RNA editing in *Silene* [22], they certainly cannot be ruled out. Likewise, it is possible that some of the gene fragments that we have classified as pseudogenes are functional. Mitochondrial tRNAs often exhibit aberrant or non-canonical secondary structures, making detection of genes and the assessment of functionality more difficult [71, 72]. Under any of these scenarios, however, it is still evident that the *S. latifolia* lineage has experienced a period of significant evolutionary change in its mitochondrially-encoded translation machinery.

**Mitochondrial Substitution Rates and Gene Conversion with Chloroplast Genes**

Given that the divergence between mitochondria (proteobacteria) and chloroplasts (cyanobacteria) spans billions of years of evolution [73], the notion that gene conversion is occurring between their respective genomes is rather astonishing. Nevertheless, examples of conversion between the mitochondrial *atp1* and chloroplast *atpA* genes have been documented in multiple angiosperm lineages [74]. The *S. latifolia* mitochondrial genome sequence provides compelling evidence for a similar history of conversion in an rRNA gene. Evidence of recombination between divergent rRNA sequences has also been found in free-living bacteria and archaea [75-77], including one other example of a chimeric proteobacterial/cyanobacterial small subunit rRNA [78].

In all documented cases of apparent conversion between mitochondrial and chloroplast genes, the mitochondrial gene acted as the recipient, which may reflect the propensity of angiosperm mitochondrial genomes to acquire and retain “promiscuous sequences”, including those of chloroplast origin. If a conversion event in *Silene* did result from a copy of chloroplast *rrn16* that had been incorporated into the

mitochondrial genome, the promiscuous sequence must have been subsequently lost, because it is no longer present in the *S. latifolia* mitochondrial genome.

The history of gene conversion in *S. latifolia rrn18* was readily detectable because the conversion tract (47 to 60 bp in length) introduced a distinct cluster of 14 substitutions (although 2 of these appear to have been obscured by subsequent mutations; Figure 8). These changes contributed to an accelerated *rrn18* substitution rate in *Silene* (Figure 5). Although we did not identify other clusters of substitutions that could be readily explained by gene conversion with homologous chloroplast sequence, it is conceivable that more localized conversion events occurred but escaped detection. It would be difficult if not impossible to distinguish conversion events that introduce only 1 or 2 substitutions from *de novo* point mutations. It has been hypothesized that increases in the frequency of gene conversion with reverse transcribed mitochondrial mRNA ("mutagenic retroprocessing") might explain elevated evolutionary rates in some angiosperm mitochondrial genomes [79]. Given the evidence for gene conversion between mitochondrial and chloroplast genes, the role of DNA-mediated conversion between divergent homologs (or even non-homologous sequences that share small regions of similarity) should be investigated as another potential source of mutational input in plant mitochondrial genomes.

**Repeats, Recombination, and Genome Structure**

With rare exception [80], the structure of angiosperm mitochondrial genomes is characterized by the presence of large repeated sequences that facilitate intra-and intermolecular recombination [12, 14]. These repeats are generally present in 2 or sometimes 3 copies. In this study, we identified an unprecedented 6-copy family of large, actively recombining repeats in the *S. latifolia* mitochondrial genome. Given a

repeat family of this size and recombinational activity, there are 120 different possible conformations for the idealized "master circle", which differ in the precise order of the 6 single-copy regions. The genome structure depicted in Figure 2 represents one of these possible conformations. However, the genome organization is much more complex than any single circular representation for at least 2 reasons. First, a 6-copy family of recombining repeats will potentially generate hundreds of possible subgenomic circles containing anywhere from 1 to 5 repeat loci, as well as a theoretically infinite number of supergenomic circles through multimerization. Second, plant mitochondrial genomes have been shown to exist *in vivo* as a complex assemblage of linear, circular and branched molecules [81, 82].

As observed in cases of repeat families with lower copy number [12, 83-89], our Southern blot hybridizations confirm the co-existence of multiple alternative genome conformations. The similar intensity of each band (Figure 4) suggests that recombination among the repeats is sufficiently frequent that the many possible pairs of flanking sequences occur at relatively equal levels, a condition defined as "recombinational equilibrium" [13]. Moreover, the repeat copies appear to be completely identical in sequence, providing further evidence for a high rate of homogenization through recombination/gene conversion.

For this study, we utilized Southern blots and *in silico* predictions from a completely sequenced plant mitochondrial genome to provide a semi-quantitative assessment of recombination activity. Extending these methods to other sequenced genomes that differ in the number and size of repeat families could provide valuable comparative data on recombination activity in plant mitochondria. Moreover, the advent of DNA sequencing technologies (e.g., 454 and Illumina) that produce deep

sequencing coverage of large span paired-end libraries can provide an opportunity to generate quantitative estimates of the relative abundance of alternative genome conformations.

## Conclusions

Overall, the patterns of gene loss and divergence in the *S. latifolia* mitochondrial genome suggest a markedly expanded role for nuclear gene products in the translation of mitochondrial genes. Furthermore, the novel, recombinationally active repeat structure of this genome represents a complex elaboration of one of the long list of unique features that distinguish plant mitochondrial genomes. With ongoing efforts to sequence the mitochondrial genomes of other *Silene* species that differ profoundly in mitochondrial mutation rates and breeding system, the *S. latifolia* mitochondrial genome should provide a valuable comparative model for investigating the evolutionary forces that shape genome organization.

## Authors' contributions

DBS planned the study, extracted DNA, performed genome finishing and most of the data analysis, and drafted the manuscript. AJA extracted DNA and helped plan the study, analyze the data and draft the manuscript. HŠ performed the Southern blot analysis and helped draft the manuscript. JDP and DRT helped plan the study, analyze the data and draft the manuscript. All authors read and approved the final manuscript.

## Acknowledgements

We gratefully acknowledge the sequencing work of Lucinda Fulton and the WUSTL Genome Center. We would also like to thank Janis Antonovics, Peter Fields, and two anonymous reviewers for helpful comments on an earlier version of this manuscript. This study was supported by NSF DEB-0808452 (to DBS and DRT), NIH RO1-GM-70612 (to JDP), and MŠMT KONTAKT ME09035 and LC06004 (to HŠ). AJA was supported by an NIH Ruth L. Kirschstein NRSA Postdoctoral Fellowship (1F32GM080079).

## References

1. Gray MW, Burger G, Lang BF: **Mitochondrial evolution.** Science 1999, **283**(5407):1476-1481.
2. Ward BL, Anderson RS, Bendich AJ: **The mitochondrial genome is large and variable in a family of plants (Cucurbitaceae).** Cell 1981, **25**(3):793-803.
3. Ellis J: **Promiscuous DNA--chloroplast genes inside plant mitochondria.** Nature 1982, **299**(5885):678-679.

4. Kubo T, Newton KJ: **Angiosperm mitochondrial genomes and mutations.** Mitochondrion 2008, **8**(1):5-14.

5. Richardson AO, Palmer JD: **Horizontal gene transfer in plants.** J Exp Bot 2007, **58**(1):1-9.

6. Bock R: **The give-and-take of DNA: horizontal gene transfer in plants.** Trends Plant Sci 2010, **15**(1):11-22.

7. Malek O, Brennicke A, Knoop V: **Evolution of trans-splicing plant mitochondrial introns in pre-Permian times.** Proc Natl Acad Sci 1997, **94**:553-558.

8. Gray MW, Covello PS: **RNA editing in plant mitochondria and chloroplasts.** FASEB J 1993, **7**(1):64-71.

9. Wolfe KH, Li WH, Sharp PM: **Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs.** Proc Natl Acad Sci 1987, **84**(24):9054-9058.

10. Mower JP, Touzet P, Gummow JS, Delph LF, Palmer JD: **Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants.** BMC Evol Biol 2007, **7**(1):135.

11. Palmer JD, Herbon LA: **Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence.** J Mol Evol 1988, **28**(1-2):87-97.

12. Palmer JD, Shields CR: **Tripartite structure of the *Brassica campestris* mitochondrial genome.** Nature 1984, **307**:437-440.

13. Lonsdale DM, Brears T, Hodge TP, Melville SE, Rottmann WH: **The plant mitochondrial genome: homologous recombination as a mechanism for**

**generating heterogeneity.** Philos Trans R Soc Lond B Biol Sci 1988, **319**(1193):149-163.

14. Marechal A, Brisson N: **Recombination and the maintenance of plant organelle genome stability.** New Phytol 2010, **186**:299-317.
15. Dietrich A, Small I, Cosset A, Weil JH, Marechal-Drouard L: **Editing and import: strategies for providing plant mitochondria with a complete set of functional transfer RNAs.** Biochimie 1996, **78**(6):518-529.
16. Small I, Akashi K, Chapron A, Dietrich A, Duchene AM, Lancelin D, Maréchal-Drouard L, Menand B, Mireau H, Moudden Y: **The strange evolutionary history of plant mitochondrial tRNAs and their aminoacyl-tRNA synthetases.** J Hered 1999, **90**(3):333-337.
17. Glover KE, Spencer DF, Gray MW: **Identification and structural characterization of nucleus-encoded transfer RNAs imported into wheat mitochondria.** J Biol Chem 2001, **276**(1):639-648.
18. Grewe F, Viehoever P, Weisshaar B, Knoop V: **A trans-splicing group I intron and tRNA-hyperediting in the mitochondrial genome of the lycophyte *Isoetes engelmannii* .** Nucleic Acids Res 2009, **37**(15):5093-5104.
19. Adams KL, Qiu YL, Stoutemyer M, Palmer JD: **Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution.** Proc Natl Acad Sci 2002, **99**(15):9905-9912.
20. Mower JP, Bonen L: **Ribosomal protein L10 is encoded in the mitochondrial genome of many land plants and green algae.** BMC Evol Biol 2009, **9**(1):265.

21. Kubo N, Arimura S: **Discovery of the rpl10 gene in diverse plant mitochondrial genomes and its probable replacement by the nuclear gene for chloroplast RPL10 in two lineages of angiosperms.** DNA Res 2010, **17**(1):1-9.

22. Sloan DB, MacQueen AH, Alverson AJ, Palmer JD, Taylor DR: **Extensive loss of RNA editing sites in rapidly evolving *Silene* mitochondrial genomes: Selection vs. retroprocessing as the driving force.** Genetics 2010, **185**(4):1369-1380.

23. McCauley DE, Bailey MF, Sherman NA, Darnell MZ: **Evidence for paternal transmission and heteroplasmy in the mitochondrial genome of *Silene vulgaris*, a gynodioecious plant.** Heredity 2005, **95**(1):50-58.

24. Bernasconi G, Antonovics J, Biere A, Charlesworth D, Delph LF, Filatov D, Giraud T, Hood ME, Marais GAB, McCauley D: ***Silene* as a model system in ecology and evolution.** Heredity 2009, **103**:5-14.

25. Desfeux C, Maurice S, Henry JP, Lejeune B, Gouyon PH: **Evolution of reproductive systems in the genus *Silene*.** Proc R Soc Lond B 1996, **263**(1369):409-414.

26. Barr CM, Keller SR, Ingvarsson PK, Sloan DB, Taylor DR: **Variation in mutation rate and polymorphism among mitochondrial genes in *Silene vulgaris* .** Mol Biol Evol 2007, **24**(8):1783-1791.

27. Sloan DB, Barr CM, Olson MS, Keller SR, Taylor DR: **Evolutionary rate variation at multiple levels of biological organization in plant mitochondrial DNA.** Mol Biol Evol 2008, **25**(2):243-246.

28. Sloan DB, Oxelman B, Rautenberg A, Taylor DR: **Phylogenetic analysis of mitochondrial substitution rate variation in the angiosperm tribe *Sileneae* (Caryophyllaceae).** BMC Evol Biol 2009, **9**:260.

29. Stadler T, Delph LF: **Ancient mitochondrial haplotypes and evidence for intragenic recombination in a gynodioecious plant.** Proc Natl Acad Sci U S A 2002, **99**(18):11730-11735.

30. Touzet P, Delph LF: **The effect of breeding system on polymorphism in mitochondrial genes of *Silene*.** Genetics 2009, **181**(2):631-644.

31. Jalas J, Suominen J: **Atlas florae Europaeae. Distribution of vascular plants in Europe. III. Caryophyllaceae.** 1987.

32. Taylor DR, Keller SR: **Historical range expansion determines the phylogenetic diversity introduced during contemporary species invasion.** Evolution 2007, **61**(2):334-345.

33. Nicolas M, Marais G, Hykelova V, Janousek B, Laporte V, Vyskot B, Mouchiroud D, Negrutiu I, Charlesworth D, Moneger F: **A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants.** PLoS Biol 2005, **3**(1):47-56.

34. Kolodner R, Tewari KK: **Physicochemical characterization of mitochondrial DNA from pea leaves.** Proc Natl Acad Sci 1972, **69**(7):1830-1834.

35. Palmer JD: **Physical and gene mapping of chloroplast DNA from *Atriplex triangularis* and *Cucumis sativa* .** Nucleic Acids Res 1982, **10**(5):1593-1605.

36. Gordon D, Abajian C, Green P: **Consed: a graphical tool for sequence finishing.** Genome Res 1998, **8**(3):195-202.

37. Alverson AJ, Wei X, Rice DW, Stern DB, Barry K, Palmer JD: **Insights into the evolution of plant mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae).** Mol Biol Evol 2010, **27**(6):1436-1448.

38. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R, Journals O: **REPuter: the manifold applications of repeat analysis on a genomic scale.** Nucleic Acids Res 2001, **29**(22):4633-4642.

39. Yang Z: **PAML 4: Phylogenetic Analysis by Maximum Likelihood.** Mol Biol Evol 2007, **24**(8):1586-1591.

40. Buckler ES, Holtsford TP: ***Zea* systematics: ribosomal ITS evidence.** Mol Biol Evol 1996, **13**(4):612-622.

41. Soltis DE, Soltis PS, Chase MW, Mort ME, Albach DC, Zanis M, Savolainen V, Hahn WH, Hoot SB, Fay MF: **Angiosperm phylogeny inferred from 18S rDNA, rbcL, and atpB sequences.** Bot J Linn Soc 2000, **133**(4):381-461.

42. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P: **Fast folding and comparison of RNA secondary structures.** Monatshefte für Chemie 1994, **125**(2):167-188.

43. Spencer DF, Bonen L, Gray MW: **Primary sequence of wheat mitochondrial 5S ribosomal ribonucleic acid: functional and evolutionary implications.** Biochemistry 1981, **20**(14):4022-4029.

44. Brennicke A, Möller S, Blanz PA: **The 18S and 5S ribosomal RNA genes in *Oenothera* mitochondria: Sequence rearrangments in the 18S and 5S rRNA genes of higher plants.** Mol General Genet 1985, **198**(3):404-410.

45. Štorchová H, Hrdličková R, Chrtek Jr J, Tetera M, Fitze D, Fehrer J: **An improved method of DNA isolation from plants collected in the field and conserved in saturated NaCl/CTAB solution.** Taxon 2000, **49**:79-84.

46. Meyerhans A, Vartanian JP, Wain-Hobson S: **DNA recombination during PCR.** Nucleic Acids Res 1990, **18**(7):1687-1691.

47. Small ID, Isaac PG, Leaver CJ: **Stoichiometric differences in DNA molecules containing the atpA gene suggest mechanisms for the generation of mitochondrial genome diversity in maize.** EMBO J 1987, **6**(4):865-869.

48. Arrieta-Montiel MP, Shedge V, Davila J, Christensen AC, Mackenzie SA: **Diversity of the Arabidopsis mitochondrial genome occurs via nuclear-controlled recombination activity.** Genetics 2009, **183**(4):1261-1268.

49. Crick FH: **Codon--anticodon pairing: the wobble hypothesis.** J Mol Biol 1966, **19**(2):548-555.

50. Cochella L, Green R: **Wobble during decoding: more than third-position promiscuity.** Nat Struct Mol Biol 2004, **11**(12):1160-1162.

51. Joly S, Brouillet L, Bruneau A: **Phylogenetic implications of the multiple losses of the mitochondrial *coxII.i3* intron in the angiosperms.** Int J Plant Sci 2001, **162**(2):359-373.

52. Sanchez-Puerta MV, Cho Y, Mower JP, Alverson AJ, Palmer JD: **Frequent, phylogenetically local horizontal transfer of the *cox1* group I intron in flowering plant mitochondria.** Mol Biol Evol 2008, **25**(8):1762-1777.

53. Shutt TE, Gray MW: **Bacteriophage origins of mitochondrial replication and transcription proteins.** Trends Genet 2006, **22**(2):90-95.

54. Robison MM, Wolyn DJ: **A mitochondrial plasmid and plasmid-like RNA and DNA polymerases encoded within the mitochondrial genome of carrot (*Daucus carota* L.).** Curr Genet 2005, **47**(1):57-66.

55. Goremykin VV, Salamini F, Velasco R, Viola R: **Mitochondrial DNA of *Vitis vinifera* and the issue of rampant horizontal gene transfer.** Mol Biol Evol 2009, **26**(1):99-110.

56. Kubo T, Mikami T: **Organization and variation of angiosperm mitochondrial genome.** Physiol Plantarum 2007, **129**(1):6-13.

57. Sloan DB, Taylor DR: **Testing for selection on synonymous sites in plant mitochondrial DNA: the role of codon bias and RNA editing.** J Mol Evol 2010, **70**:479-491.

58. Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Müller KM: **The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs.** BMC Bioinformatics 2002, **3**:2.

59. Gray MW, Lang BF, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, Brossard N, Delage E, Littlejohn TG, Plante I, Rioux P, Saint-Louis D, Zhu Y, Burger G: **Genome structure and gene content in protist mitochondrial DNAs.** Nucleic Acids Res 1998, **26**(4):865-878.

60. Nugent JM, Palmer JD: **RNA-mediated transfer of the gene coxII from the mitochondrion to the nucleus during flowering plant evolution.** Cell 1991, **66**(3):473-481.

61. Kubo N, Harada K, Hirai A, Kadowaki K: **A single nuclear transcript encoding mitochondrial RPS14 and SDHB of rice is processed by alternative splicing:**

**common use of the same mitochondrial targeting signal for different proteins.** Proc Natl Acad Sci 1999, **96**(16):9207-9211.

62. Adams KL, Daley DO, Qiu YL, Whelan J, Palmer JD: **Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants.** Nature 2000, **408**(6810):354-357.
63. Liu SL, Zhuang Y, Zhang P, Adams KL: **Comparative analysis of structural diversity and sequence evolution in plant mitochondrial genes transferred to the nucleus.** Mol Biol Evol 2009, **26**(4):875-891.
64. Adams KL, Daley DO, Whelan J, Palmer JD: **Genes for two mitochondrial ribosomal proteins in flowering plants are derived from their chloroplast or cytosolic counterparts.** Plant Cell 2002, **14**(4):931-943.
65. Mollier P, Hoffmann B, Debast C, Small I: **The gene encoding *Arabidopsis thaliana* mitochondrial ribosomal protein S13 is a recent duplication of the gene encoding plastid S13.** Curr Genet 2002, **40**(6):405-409.
66. Paquin B, Laforest MJ, Forget L, Roewer I, Wang Z, Longcore J, Lang BF: **The fungal mitochondrial genome project: evolution of fungal mitochondrial genomes and their gene expression.** Curr Genet 1997, **31**(5):380-395.
67. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B: **Genome sequence of the**

**human malaria parasite *Plasmodium falciparum*.** Nature 2002, **419**(6906):498-511.

68. Pujol C, Bailly M, Kern D, Marechal-Drouard L, Becker H, Duchene AM: **Dual-targeted tRNA-dependent amidotransferase ensures both mitochondrial and chloroplastic Gln-tRNAGln synthesis in plants.** Proc Natl Acad Sci 2008, **105**(17):6481-6485.
69. Wolfe KH, Morden CW, Ems SC, Palmer JD: **Rapid evolution of the plastid translational apparatus in a nonphotosynthetic plant: loss or accelerated sequence evolution of tRNA and ribosomal protein genes.** J Mol Evol 1992, **35**(4):304-317.
70. Feagin JE, Abraham JM, Stuart K: **Extensive editing of the cytochrome c oxidase III transcript in *Trypanosoma brucei* .** Cell 1988, **53**(3):413-422.
71. Okimoto R, Wolstenholme DR: **A set of tRNAs that lack either the T psi C arm or the dihydrouridine arm: towards a minimal tRNA adaptor.** EMBO J 1990, **9**(10):3405-3411.
72. Schnare MN, Greenwood SJ, Gray MW: **Primary sequence and post-transcriptional modification pattern of an unusual mitochondrial tRNA(Met) from *Tetrahymena pyriformis* .** FEBS Lett 1995, **362**(1):24-28.
73. Sheridan PP, Freeman KH, Brenchley JE: **Estimated minimal divergence times of the major bacterial and archaeal phyla.** Geomicrobiol J 2003, **20**(1):1-14.
74. Hao W, Palmer JD: **Fine-scale mergers of chloroplast and mitochondrial genes create functional, transcompartmentally chimeric mitochondrial genes.** Proc Natl Acad Sci 2009, **106**(39):16728-16733.

75. Parker MA: **Case of localized recombination in 23S rRNA genes from divergent *Bradyrhizobium* lineages associated with neotropical legumes.** Appl Environ Microbiol 2001, **67**(5):2076-2082.

76. Schouls LM, Schot CS, Jacobs JA: **Horizontal transfer of segments of the 16S rRNA genes between species of the *Streptococcus anginosus* group.** J Bacteriol 2003, **185**(24):7241-7246.

77. Boucher Y, Douady CJ, Sharma AK, Kamekura M, Doolittle WF: **Intragenomic heterogeneity and intergenomic recombination among haloarchaeal rRNA genes.** J Bacteriol 2004, **186**(12):3980-3990.

78. Miller SR, Augustine S, Olson TL, Blankenship RE, Selker J, Wood AM: **Discovery of a free-living chlorophyll d-producing cyanobacterium with a hybrid proteobacterial/cyanobacterial small-subunit rRNA gene.** Proc Natl Acad Sci 2005, **102**(3):850-855.

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

80. Palmer JD, Herbon LA: **Unicircular structure of the *Brassica hirta* mitochondrial genome.** Curr Genet 1987, **11**(6-7):565-570.

81. Oldenburg DJ, Bendich AJ: **Size and Structure of Replicating Mitochondrial DNA in Cultured Tobacco Cells.** Plant Cell 1996, **8**(3):447-461.

82. Backert S, Borner T: **Phage T4-like intermediates of DNA replication and recombination in the mitochondria of the higher plant *Chenopodium album* (L.).** Curr Genet 2000, **37**(5):304-314.

83. Lonsdale DM, Hodge TP, Fauron CMR, Flavell RB: **A predicted structure for the mitochondrial genome from the fertile cytoplasm of maize.** In *Plant molecular biology: UCLA symposia on Molecular and Cellular Biology, New Series. Volume 12.* Edited by Goldberg RB. New York: AR Liss; 1983:445-456.

84. Stern DB, Palmer JD: **Recombination sequences in plant mitochondrial genomes: diversity and homologies to known mitochondrial genes.** Nucleic Acids Res 1984, **12**(15):6141-6157.

85. Palmer JD, Herbon LA: **Tricircular mitochondrial genomes of *Brassica* and *Raphanus*: reversal of repeat configurations by inversion.** Nucleic Acids Res 1986, **14**(24):9755-9764.

86. Stern DB, Palmer JD: **Tripartite mitochondrial genome of spinach: physical structure, mitochondrial gene mapping, and locations of transposed chloroplast DNA sequences.** Nucleic Acids Res 1986, **14**(14):5651-5666.

87. Palmer JD: **Intraspecific variation and multicircularity in Brassica mitochondrial DNAs.** Genetics 1988, **118**(2):341-351.

88. Siculella L, Palmer JD: **Physical and gene organization of mitochondrial DNA in fertile and male sterile sunflower. CMS-associated alterations in structure and transcription of the atpA gene.** Nucleic Acids Res 1988, **16**(9):3787-3799.

89. Folkerts O, Hanson MR: **Three copies of a single recombination repeat occur on the 443 kb master circle of the *Petunia hybrida* 3704 mitochondrial genome.** Nucleic Acids Res 1989, **17**(18):7345-7357.

90. Leon P, Walbot V, Bedinger P: **Molecular analysis of the linear 2.3 kb plasmid of maize mitochondria: apparent capture of tRNA genes.** Nucleic Acids Res 1989, **17**(11):4089-4099.

91. Lohse M, Drechsel O, Bock R: **OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes.** Curr Genet 2007, **52**(5-6):267-274.

92. Darty K, Denise A, Ponty Y: **VARNA: Interactive drawing and editing of the RNA secondary structure.** Bioinformatics 2009, **25**(15):1974-1975.

93. Placido A, Damiano F, Losacco M, Rainaldi G, De Benedetto C, Gallerani R: **Variable structures of promoters regulating transcription of cp-like tRNA genes and of some native genes on the sunflower mitochondrial genome.** Gene 2006, **371**(1):93-101.

# Figures

**Figure 1 - Gene content in seed plant mitochondrial genomes**

Dark gray boxes indicate the presence of an intact reading frame or folding structure and, therefore, a putatively functional gene, while light gray boxes indicate the presence of a putative pseudogene. The numbers at the bottom of each gene group indicate the total number of intact genes for that species. Note that in some cases the presence of an intact gene sequence may not actually reflect functionality. In particular, for tRNA genes of chloroplast origin, it is possible that transferred sequences still appear intact, but nevertheless, are not functionally expressed in the mitochondrion [37, 90]. GenBank accession numbers for each genome are indicated in parentheses.

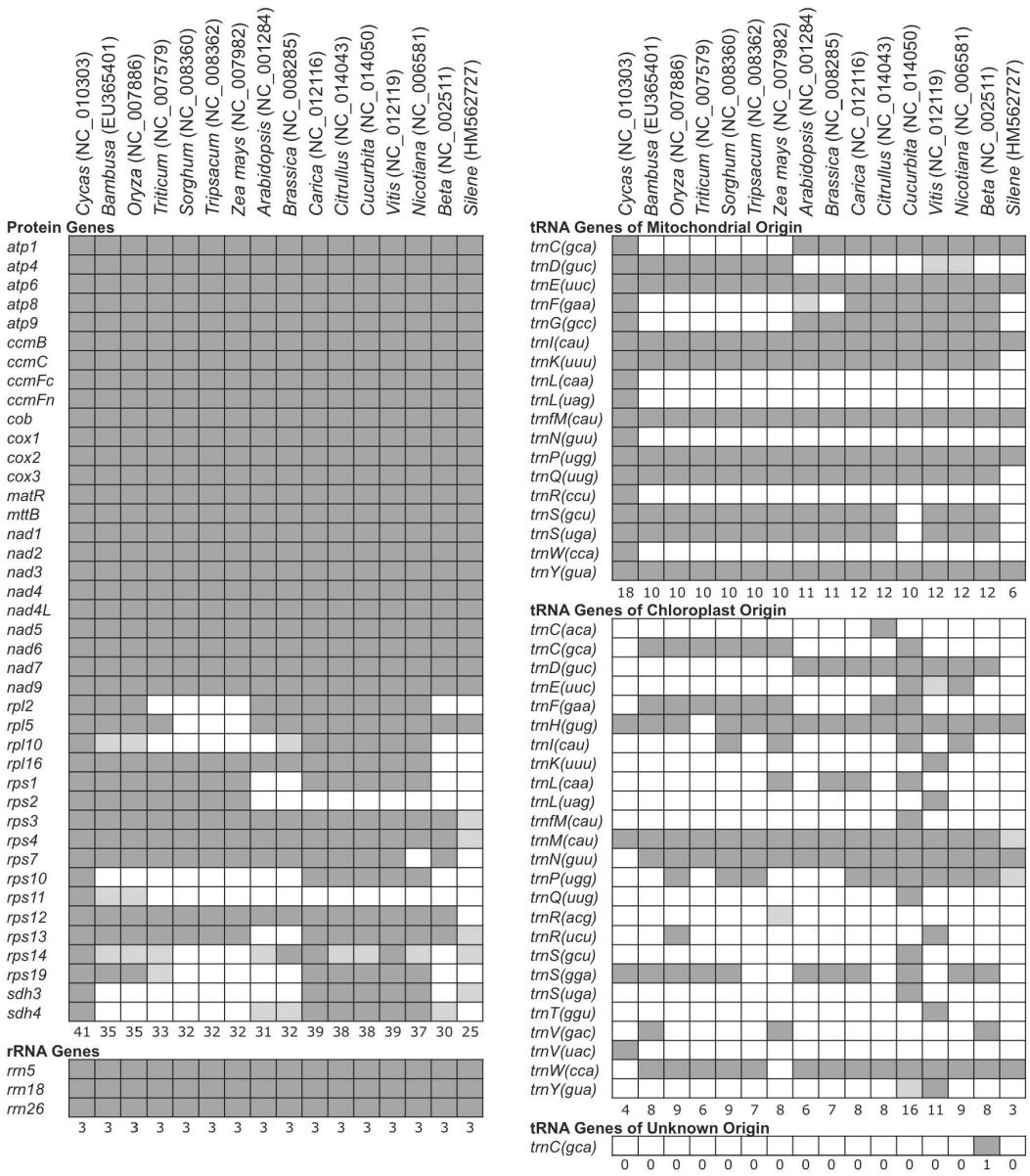

Cycas (NC_010303)
Bambusa (EU365401)
Oryza (NC_007886)
Triticum (NC_007579)
Sorghum (NC_008360)
Tripsacum (NC_008362)
Zea mays (NC_007982)
Arabidopsis (NC_001284)
Brassica (NC_008285)
Carica (NC_012116)
Citrullus (NC_014043)
Cucurbita (NC_014050)
Vitis (NC_012119)
Nicotiana (NC_006581)
Beta (NC_002511)
Silene (HM562727)
Protein Genes
atp1
atp4
atp6
atp8
atp9
ccmB
ccmC
ccmFc
ccmFn
cob
cox1
cox2
cox3
matR
mttB
nad1
nad2
nad3
nad4
nad4L
nad5
nad6
nad7
nad9
rpl2
rpl5
rpl10
rpl16
rps1
rps2
rps3
rps4
rps7
rps10
rps11
rps12
rps13
rps14
rps19
sdh3
sdh4
41 35 35 33 32 32 32 31 32 39 38 38 39 37 30 25
rRNA Genes
rrn5
rrn18
rrn26
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
tRNA Genes of Mitochondrial Origin
trnC(gca)
trnD(guc)
trnE(uuc)
trnF(gaa)
trnG(gcc)
trnI(cau)
trnK(uuu)
trnL(caa)
trnL(uag)
trnfM(cau)
trnN(guu)
trnP(ugg)
trnQ(uug)
trnR(ccu)
trnS(gcu)
trnS(uga)
trnW(cca)
trnY(gua)
18 10 10 10 10 10 10 11 11 12 12 10 12 12 12 6
tRNA Genes of Chloroplast Origin
trnC(aca)
trnC(gca)
trnD(guc)
trnE(uuc)
trnF(gaa)
trnH(gug)
trnI(cau)
trnK(uuu)
trnL(caa)
trnL(uag)
trnfM(cau)
trnM(cau)
trnN(guu)
trnP(ugg)
trnQ(uug)
trnR(acg)
trnR(ucu)
trnS(gcu)
trnS(gga)
trnS(uga)
trnT(ggu)
trnV(gac)
trnV(uac)
trnW(cca)
trnY(gua)
4 8 9 6 9 7 8 6 7 8 8 16 11 9 8 3
tRNA Genes of Unknown Origin
trnC(gca)
0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0

**Figure 2 - Mitochondrial genome map**

One of many possible master circle representations of the *Silene latifolia* mitochondrial genome (although this does not necessarily reflect the *in vivo* structure of the genome; see Discussion). Boxes inside and outside the circle correspond to genes on the clockwise and anti-clockwise strand, respectively. Arrows indicate the orientation of repeats as shown in Figure 3. This figure was generated with OGDraw v1.1 [91].

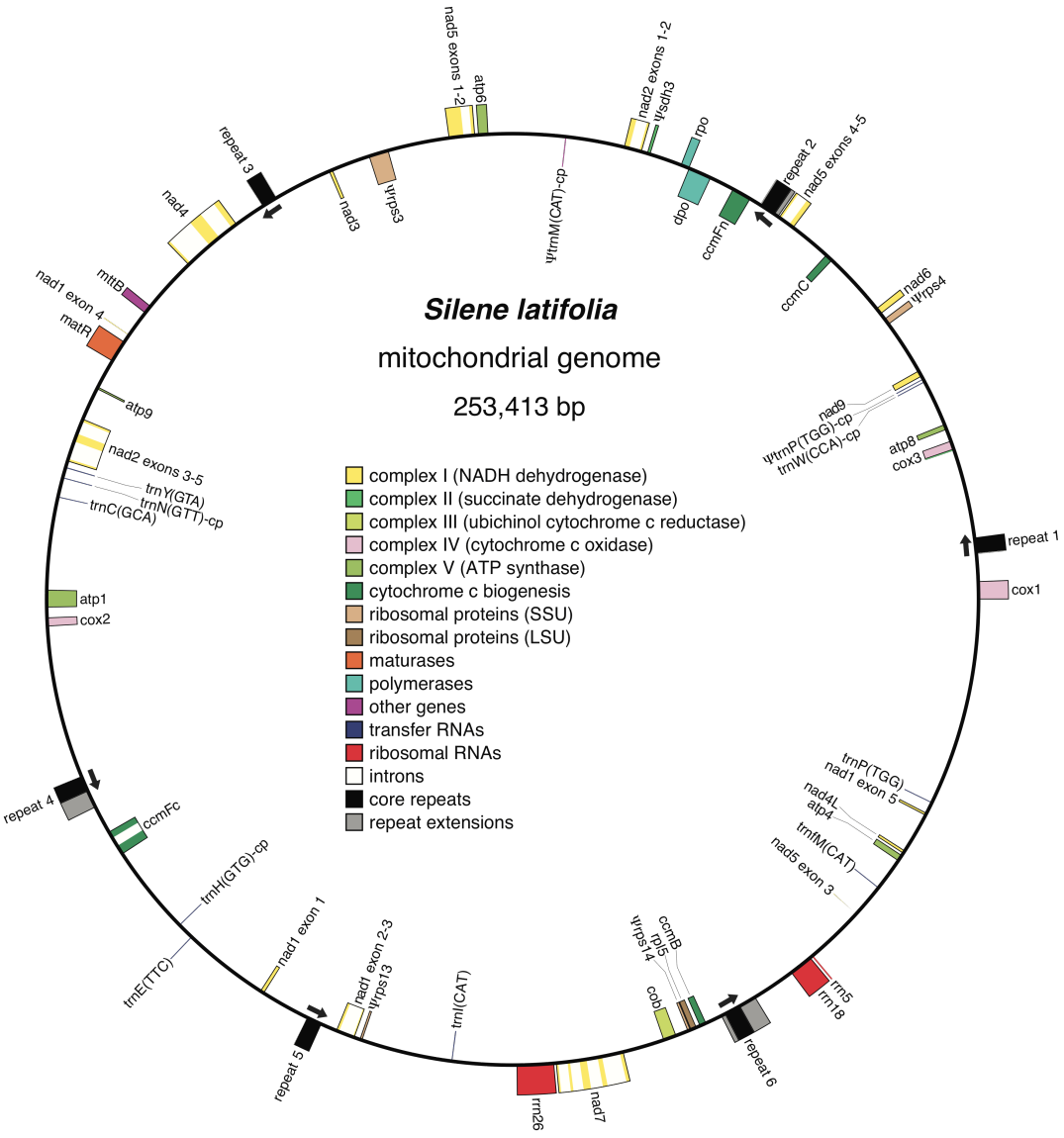


**Figure 3 - Structure of large repeated sequences in the *Silene latifolia* mitochondrial genome**

The genome contains a 1362 bp direct repeat present in 6 copies (white boxes). Additional repeat extensions (gray boxes) of varying length are shared by some but not all of the regions that flank the repeat. Shorter repeat extensions are identical in sequence to the initial portions of longer repeat extensions, with the exception of the "left" flanking regions next to repeat copies 4 and 5, which share a short 12 bp sequence (solid black boxes) that is unique relative to the other flanking sequences. Single copy sequences flanking the repeats are shown by thin black lines. The red bars indicate the location of probes used in Southern blot hybridizations (Figure 4). The values on the left and right side indicate the length of the respective repeat extensions. The order of the repeat copies and their flanking sequences corresponds to the genome conformation shown in Figure 2.

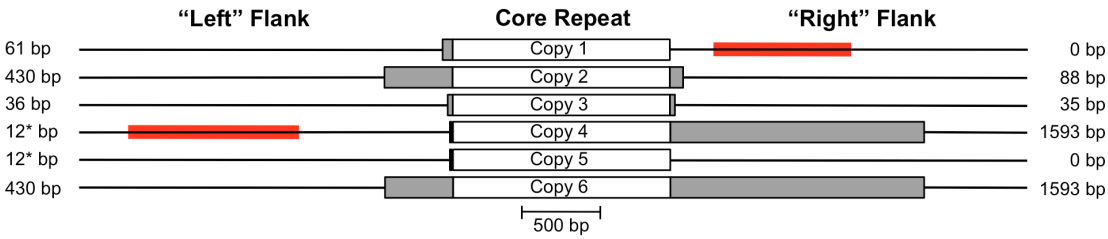


**Figure 4 - Recombining repeats in the *Silene latifolia* mitochondrial genome**

(A) A stylized version of the master circle undergoing one of many possible recombination events. The black boxes represent the 6-copy repeat with numbering corresponding to Figure 2. The lettered sections represent intervening single-copy regions. The "left" and "right" probes used in Southern blot hybridizations are indicated with small gray bars and labeled L and R, respectively. The dotted gray lines indicate a crossover event between repeat copies that produces 2 sub-genomic molecules. Given all possible recombination events, each left flanking sequence has the potential to be paired with 6 different right flanking sequences (and vice versa), and therefore, each probe is expected to hybridize to 6 restriction fragments. (B) Southern blot hybridizations with "left" and "right" probes each show 6 strong bands, corresponding to the sizes predicted based on recombination among the 6 large repeats (see Additional File 1 for a more resolved replicate of the "left" probe blot). The left pair of lanes contain DNA samples from one full-sib family, while the right pair contain DNA samples from a second full-sib family. The size standards are indicated by the values between the two blots. The values on either side represent the predicted fragment sizes with the corresponding single-copy flanking sequence noted in parentheses. The black triangle indicates an unexpected 1.8 kb fragment detected in some but not all individuals with the "right" probe.

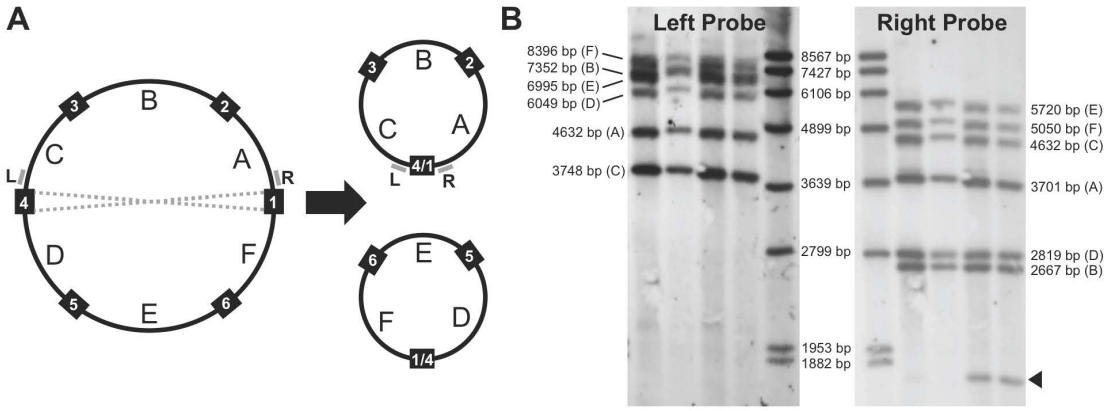

A
B
C
A
D
F
E
L
R
4
1
2
3
5
6
4/1
1/4
B
Left Probe
Right Probe
8396 bp (F)
7352 bp (B)
6995 bp (E)
6049 bp (D)
4632 bp (A)
3748 bp (C)
8567 bp
7427 bp
6106 bp
4899 bp
3639 bp
2799 bp
1953 bp
1882 bp
5720 bp (E)
5050 bp (F)
4632 bp (C)
3701 bp (A)
2819 bp (D)
2667 bp (B)

**Figure 5 - Phylogenetic analysis of substitution rates in seed plant mitochondrial genomes**

rRNA gene branch lengths are in terms of substitutions per site, while protein gene branch lengths reflect synonymous substitutions per site based on a concatenated dataset of 25 genes present in the mitochondrial genomes of all 18 species. All analyses used a constrained topology.

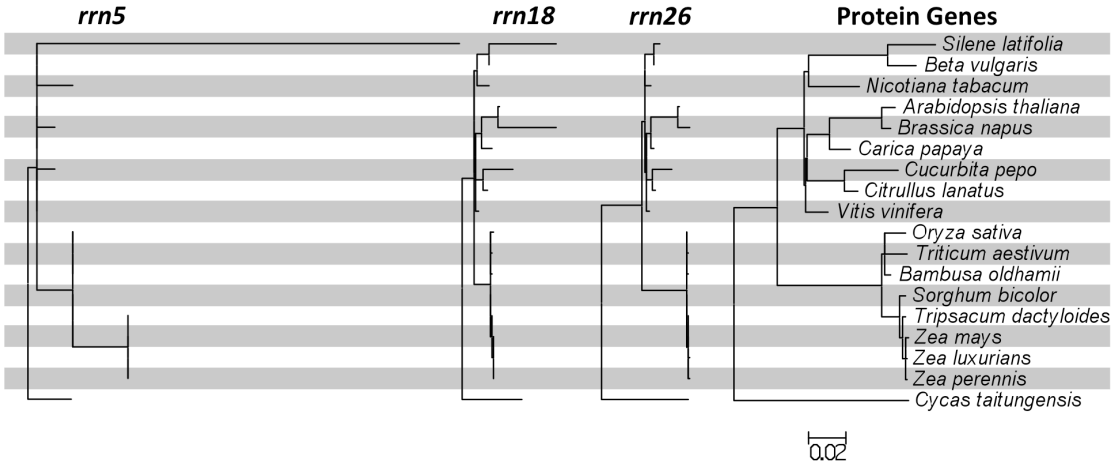


**Figure 6 - Substitution rate variation among genes in *Silene latifolia***

Each bar represents the terminal branch length for *S. latifolia* based on a phylogenetic analysis of 18 land plant species with fully sequenced mitochondrial genomes. For protein genes, branch lengths were estimated in terms of non-synonymous substitutions (black bars) or synonymous substitutions (white bars) per site. For rRNA genes, branch lengths were estimated in terms of substitutions per site (gray bars). Error bars represent standard errors, which were calculated as described by Parkinson et al. [79].

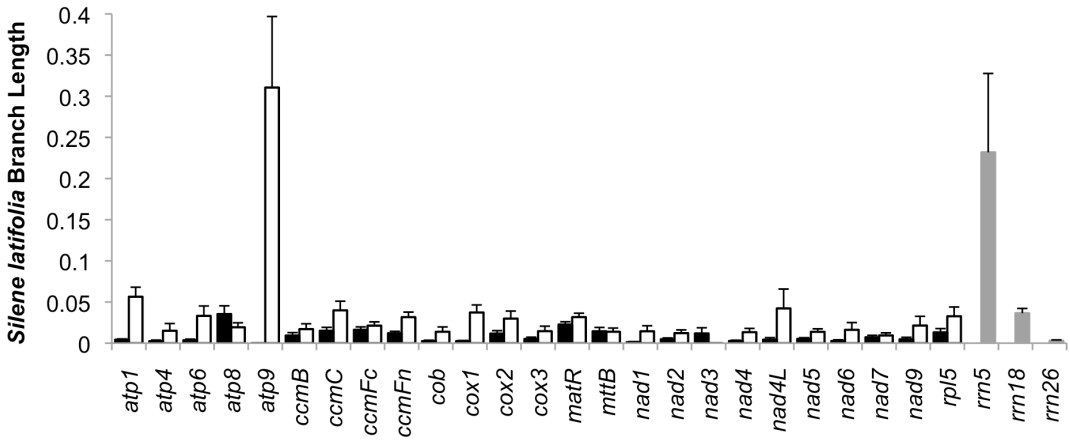


**Figure 7 - Predicted secondary structure for *Silene latifolia* 5S ribosomal RNA (*rrn5*)**

Sites that have experienced a substitution in the *S. latifolia* lineage are highlighted in black. The black arrow indicates the one predicted change in secondary structure resulting from nucleotide substitution (a novel base pairing between positions 34 and 46). The figure was generated with VARNA v3.6 [92].

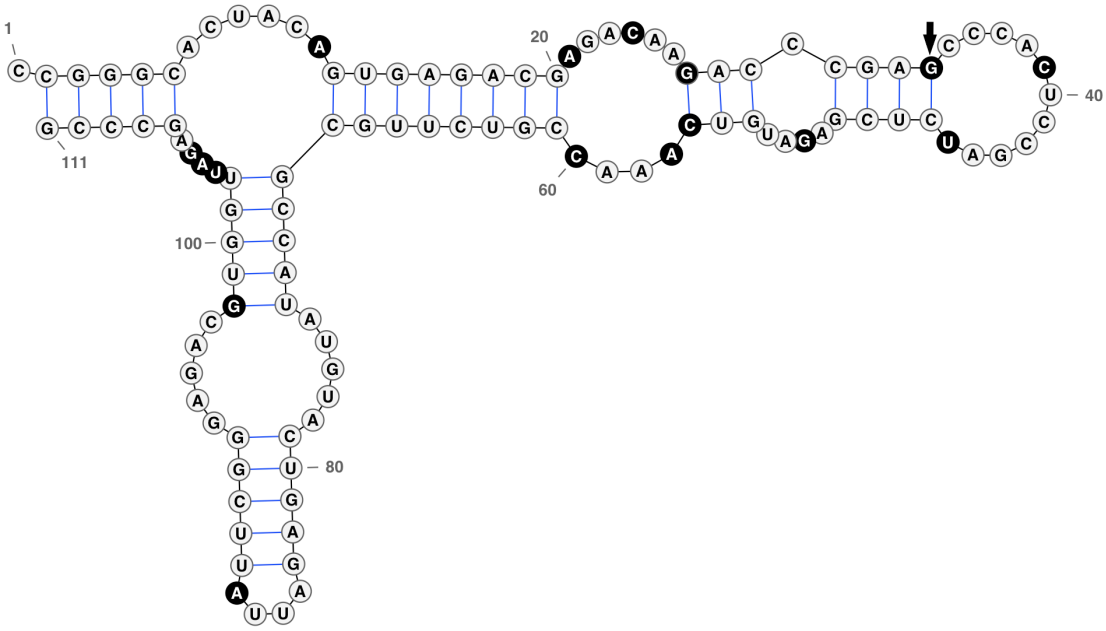
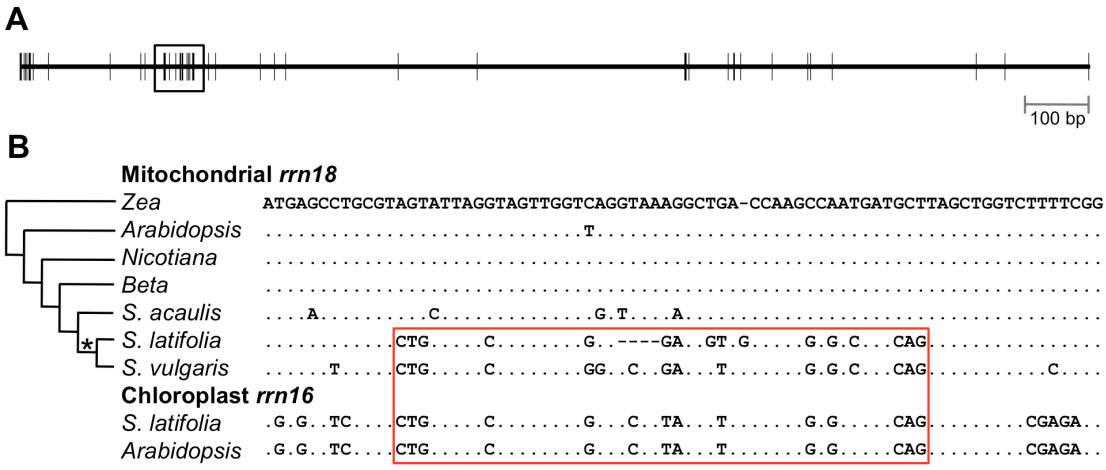


**Figure 8 - Gene conversion between mitochondrial and chloroplast small subunit rRNA genes**

(A) The spatial distribution of substitutions (vertical lines) in mitochondrial *rrn18* that distinguish *Silene latifolia* from *Beta vulgaris* (regions that could not be reliably aligned in a multiple species alignment were excluded). The black box indicates the region shown in detail below. (B) Aligned sequences of angiosperm mitochondrial *rrn18* and chloroplast *rrn16*. Dots in the alignment indicate sequence identity with the *Zea* reference sequence. The red box shows the minimal extent of the region inferred to have experienced a gene conversion event, which also corresponds to the position of helix 240 in *E. coli* 16S rRNA [58]. Analysis of these sequences with GENECONV v1.81a using a mismatch cost of 1 found highly significant evidence for gene conversion in this region ($p < 0.0001$). The asterisk indicates the inferred phylogenetic timing of that event. Gene sequences were taken from published genomes (see Figure 1) with the exception *S. acaulis* and *S. vulgaris rrn18* (GenBank EF547249 and HM562728) and *S. latifolia rrn16* (AB189069).



# Tables

**Table 1 - Translational capacity of mitochondrially encoded tRNAs in *Silene latifolia***

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| UUU | Phe | -- | UCU | Ser | -- | UAU | Tyr | wobble | UGU | Cys | wobble |
| UUC | Phe | -- | UCC | Ser | -- | UAC | Tyr | trnY(gua) | UGC | Cys | trnC(gca) |
| UUA | Leu | -- | UCA | Ser | -- | UAA | * | -- | UGA | * | -- |
| UUG | Leu | -- | UCG | Ser | -- | UAG | * | -- | UGG | Trp | trnW(cca)-cp |
| CUU | Leu | -- | CCU | Pro | wobble | CAU | His | wobble | CGU | Arg | -- |
| CUC | Leu | -- | CCC | Pro | wobble | CAC | His | trnH(gug)-cp | CGU | Arg | -- |
| CUA | Leu | -- | CCA | Pro | trnP(ugg)[3] | CAA | Gln | -- | CGA | Arg | -- |
| CUG | Leu | -- | CCG | Pro | wobble | CAG | Gln | -- | CGG | Arg | -- |
| AUU | Ile | -- | ACU | Thr | -- | AAU | Asn | wobble | AGU | Ser | -- |
| AUC | Ile | -- | ACC | Thr | -- | AAC | Asn | trnN(guu)-cp | AGC | Ser | -- |
| AUA | Ile | trnI(cau)[1] | ACA | Thr | -- | AAA | Lys | -- | AGA | Arg | -- |
| AUG | Met | trnfM(cau)[2] | ACG | Thr | -- | AAG | Lys | -- | AGG | Arg | -- |
| GUU | Val | -- | GCU | Ala | -- | GAU | Asp | -- | GGU | Gly | -- |
| GUC | Val | -- | GCC | Ala | -- | GAC | Asp | -- | GGC | Gly | -- |
| GUA | Val | -- | GCA | Ala | -- | GAA | Glu | trnE(uuc) | GGA | Gly | -- |
| GUG | Val | -- | GCG | Ala | -- | GAG | Glu | wobble | GGG | Gly | -- |

[1]The C in the first anticodon position of *trnI(cau)* is assumed to be post-transcriptionally converted to lysidine, which pairs with A.

[2]*trnfM(cau)* transfers formylmethionine and therefore is assumed to recognize only the AUG start codon. The genome also contains a gene with homology to the chloroplast *trnM* gene (which recognizes internal AUG codons), but it has experienced a large expansion in its anticodon loop, making its functionality (and its anticodon) uncertain (Additional File 2).

[3]The genome contains a native *trnP(ugg)* gene as well as a homolog of the chloroplast *trnP(ugg)* gene. The chloroplast-derived copy is likely a pseudogene in *S. latifolia*, as it is believed to be in some other angiosperm mitochondrial genomes [90, 93]. Its anticodon loop has experienced multiple substitutions, including one that converts the ancestral UGG anticodon to AGG (Additional File 2).

# Additional files

**Additional file 1 – Southern blot hybridizations**

**Additional file 2 – Predicted secondary structures of mitochondrially-encoded tRNAs**

**Additional file 3 – Summary of codon usage**

Chapter 6.

# Extensive loss of RNA editing sites in rapidly evolving *Silene* mitochondrial genomes: selection vs. retroprocessing as the driving force[1]

[1]Formatted as a co-authored manuscript and published as:

Sloan DB, MacQueen AH, Alverson AJ, Palmer JD, Taylor DR. 2010. *Genetics* 185:1369-1380.

Referenced supplementary material is available online at:

http://www.genetics.org/cgi/content/abstract/185/4/1369

ABSTRACT

Theoretical arguments suggest that mutation rates influence the proliferation and maintenance of RNA editing. We identified RNA editing sites in five species within the angiosperm genus *Silene* that exhibit highly divergent mitochondrial mutation rates. We found that mutational acceleration has been associated with rapid loss of mitochondrial editing sites. In contrast, we did not find a significant difference in the frequency of editing in chloroplast genes, which lack the mutation rate variation observed in the mitochondrial genome. As found in other angiosperms, the rate of substitution at RNA editing sites in *Silene* greatly exceeds the rate at synonymous sites, a pattern that has previously been interpreted as evidence for selection against RNA editing. Alternatively, we suggest that editing sites may experience higher rates of C-to-T mutation than other portions of the genome. Such a pattern could be caused by gene conversion with reverse transcribed mRNA (i.e., retroprocessing). If so, the genomic distribution of RNA editing site losses in *Silene* suggests that such conversions must be occurring at a local scale such that only one or two editing sites are affected at a time. Because preferential substitution at editing sites appears to occur in angiosperms regardless of the mutation rate, we conclude that mitochondrial rate accelerations within *Silene* have "fast-forwarded" a pre-existing pattern but have not fundamentally changed the evolutionary forces acting on RNA editing sites.

## INTRODUCTION

In the organelle genomes of land plants, a variable but often large number of sites undergo C-to-U RNA editing in which a cytidine is converted to uridine by deamination (Yu and Schuster. 1995; Giege and Brennicke. 1999). A generally much smaller number of sites undergo "reverse" U-to-C editing (Steinhauser *et al*. 1999). RNA editing is believed to be essential for organelle gene function in plants. Editing sites are preferentially located in protein genes and, within them, at first and second codon positions (Gray. 2003). Editing at these sites generally results in the restoration of phylogenetically conserved (and presumably functionally constrained) amino acids in mitochondrial and chloroplast protein sequences (Gray and Covello. 1993; Mower. 2005; Yura and Go. 2008). Therefore, there are obvious selective pressures for plants to maintain RNA editing in the short-term. In contrast, the origin and long-term maintenance of RNA editing pose an evolutionary puzzle, as it is unclear what if any benefit this seemingly cumbersome process confers over direct encoding of the edited sequence in the genomic DNA. This puzzle mirrors broader evolutionary questions about the origin, maintenance, and function of a number of major features of gene and genome architecture (Lynch. 2007).

Various adaptive effects of RNA editing have been proposed, including a role in gene regulation (Hirose *et al*. 1999; Farajollahi and Maas. 2010), maintenance of alternative functional protein isoforms (Gott. 2003; Farajollahi and Maas. 2010), generation of genetic variation (Tillich *et al*. 2006; Gommans *et al*. 2009), optimization of genomic GC content (Jobson and Qiu. 2008), nuclear control of selfish organelle genes (Burt and Trivers. 2006), and mutational buffering (Borner *et al*. 1997). In humans, there

is evidence for divergent functional roles for products of the edited and unedited forms of *apolipoprotein B* (Powell *et al*. 1987), but there is little evidence for the aforementioned adaptive mechanisms in plant organelles. Moreover, replacement of edited C's with T's at the genomic level appears to occur readily across lineages with no obvious detrimental effect (Shields and Wolfe. 1997; Mower. 2008). Accordingly, neutral and non-adaptive models for the proliferation of RNA editing have also been proposed (Covello and Gray. 1993; Fiebig *et al*. 2004; Lynch *et al*. 2006).

C-to-U RNA editing appears to have evolved in a recent common ancestor of land plants, but the frequency of editing varies dramatically across lineages and between genomes (Turmel *et al*. 2003; Salone *et al*. 2007; Grewe *et al*. 2009). Seed plants experience relatively frequent editing in mitochondrial genes with approximately 400 sites per species (>1% of all coding sequence) in the few angiosperms examined so far, (e.g., Giege and Brennicke. 1999) and even higher rates inferred in gymnosperms (Lu *et al*. 1998; Chaw *et al*. 2008; Ran *et al*. 2010). The rate of editing in seed plant chloroplast genomes is over an order of magnitude lower (Tillich *et al*. 2006). Outside seed plants, the moss *Physcomitrella patens* has only 11 edited sites in the mitochondrial genome, and the liverwort *Marchantia polymorpha* appears to have lost RNA editing altogether (Rüdinger *et al*. 2008; Rüdinger *et al*. 2009). In contrast, other bryophytes, lycophytes, and ferns exhibit frequent mitochondrial editing (Malek *et al*. 1996; Grewe *et al*. 2009; Li *et al*. 2009), while the only two non-seed plant chloroplast genomes examined (from one hornwort and one fern) also have high levels of editing (Kugita *et al*. 2003; Wolf *et al*. 2004).

Mutation rate variation is one proposed mechanism to explain phylogenetic variation in the occurrence and frequency of RNA editing. Plant mitochondria tend to have very slow point mutation rates as evidenced by their low rate of synonymous substitution (Wolfe *et al*. 1987; Palmer and Herbon. 1988; Mower *et al*. 2007; Drouin *et al*. 2008). Lynch *et al.* (2006) have argued that high mutation rates intensify selection against RNA editing because of the increased probability of disrupting sequences necessary for editing site recognition (the mutational burden hypothesis; Lynch. 2006). This led them to conclude that the low mutation rates in plant mitochondrial genomes have created a permissive environment in which RNA editing sites have been able to proliferate by non-adaptive processes. Recent studies have documented a handful of angiosperm lineages that have experienced dramatic accelerations in synonymous substitution rate (and presumably the underlying mutation rate; Cho *et al*. 2004; Parkinson *et al*. 2005; Mower *et al*. 2007; Sloan *et al*. 2008; Sloan *et al*. 2009). In one extreme example, RNA editing sites appear (based on a sampling of relatively few genes) to have been almost completely eliminated in *Pelargonium hortorum*, which has experienced a series of two successive ~10-fold increases in mitochondrial mutation rate (Parkinson *et al*. 2005). While these results are grossly consistent with the role of selection proposed by Lynch *et al.*, the data have also been interpreted as evidence for potentially frequent conversion of reverse-transcribed mRNA back into the genome. Parkinson *et al*. (2005) suggest that such "mutagenic retroprocessing" could explain both the high overall rate of mutation (because of the very high error rates in reverse transcription) and the loss of RNA editing (because the edited sequence is being incorporated into the genome).

This existing body of work raises two questions that we address in the present study: (1) Do major losses of RNA editing sites always accompany mutation rate accelerations in angiosperms, and (2) if so, to what extent can this correlation be explained by selection (Shields and Wolfe. 1997; Lynch *et al*. 2006; Mower. 2008) as opposed to the mutational bias introduced by retroprocessing (RNA-mediated gene conversion; Parkinson *et al*. 2005; Mulligan *et al*. 2007)?

The genus *Silene* (Caryophyllaceae) offers a particularly valuable system for investigating the consequences of mitochondrial mutation rate variation. Multiple species within the genus exhibit increases of greater than 100-fold in synonymous substitution rate (Figure 1; Sloan *et al*. 2009). These changes appear to have occurred very recently (less than 5-10 million years ago), allowing for close phylogenetic comparisons between accelerated and non-accelerated species (Mower *et al*. 2007; Sloan *et al*. 2009). The recent timing of these accelerations enables tests of molecular evolution based on synonymous site divergence, which has yet to approach saturation despite the extreme substitution rates. Interestingly, a previous study found little difference in the predicted frequency of mitochondrial RNA editing sites between accelerated and non-accelerated *Silene* species (Mower *et al*. 2007). This analysis, however, was restricted to a small sample of five relatively lightly edited *Silene* mitochondrial genes.

We performed cDNA sequencing to empirically identify RNA editing sites for all mitochondrial protein-coding genes in *S. latifolia*, which exhibits a slow rate of mitochondrial substitution typical of most plants, and a close relative, *S. noctiflora*, which has experienced a mitochondrial rate acceleration of >100-fold. We also analyzed a smaller sample of genes for three additional species including *S. conica*, which has a

mitochondrial substitution rate comparably high to that of *S. noctiflora*. We find that the mitochondrial mutational accelerations in *Silene* have been associated with substantial loss of RNA editing sites via C-to-T substitutions. We explore these results along with previously published data in the context of competing models to infer the potential role of both adaptive and non-adaptive forces in the evolution of RNA editing.

## MATERIALS AND METHODS

### Plant Material and RNA and DNA Extraction

*Silene* contains hundreds of predominantly herbaceous species, some of which are widely used in studies of ecology and evolution (Bernasconi *et al*. 2009). Seeds for each of five *Silene* species (*S. latifolia*, *S. noctiflora, S. conica, S. vulgaris* and *S. paradoxa*) were grown in the greenhouse under a 16 hour-8 hour light-dark cycle. Fresh leaf tissue was collected from plants that were >10 weeks old for both RNA and DNA extraction. Total cellular RNA was purified using the RNeasy Mini Kit (Qiagen). Contaminating genomic DNA was removed by digestion with RNase-free DNase I (Qiagen) for 30 minutes at room temperature, and post-reaction clean-up was performed with a second RNeasy spin column. Genomic DNA was extracted using a modified CTAB method (Doyle and Doyle. 1987).

### Reverse Transcription, PCR, and Sequencing

cDNA was produced by reverse transcription of purified RNA primed with random hexamers, using M-MulV reverse transcriptase (New England Biolabs) in accordance with manufacturer's recommendations.

PCR primers were designed to amplify overlapping fragments of up to 700 bp for all mitochondrial protein genes based on draft mitochondrial genome assemblies for *S. latifolia* and *S. noctiflora* (unpublished data). We also designed additional primers that were conserved across *Silene* species for portions of seven mitochondrial genes (*ccmFn, cob, nad2, nad5, nad6, nad7,* and *nad9*) and five chloroplast genes (*ndhB, psbL, rpoB, rpoC1,* and *rps2*). Primer sequences were chosen to exclude predicted RNA editing sites to avoid enriching for unedited transcripts (Mower. 2005; Mower and Palmer. 2006). cDNA was amplified using standard PCR techniques. We also PCR amplified genomic DNA when the corresponding genomic sequence was not already available. PCR products were cycle sequenced directly on both strands as described previously (Barr *et al*. 2007). Primer sequences are provided in File S1. Sequences generated in this study were deposited to GenBank under accessions HM099771-HM099885.

**Identification of RNA Editing Sites**

RNA editing sites were identified by comparing aligned genomic and cDNA sequences and verified by manual inspection of sequencing electropherograms. Partially edited sites were identified and scored only when peaks for both the edited and unedited bases were substantially above background on both strands. In some genes, all edited sites exhibited a degree of incomplete editing (see Files S2 and S3), which could reflect true partial editing but may also result from preferential amplification of immature transcripts or

contaminating genomic DNA (Mower and Palmer. 2006). In such cases, only editing sites that clearly differed from the observed baseline level of incomplete editing were classified as partially edited. This conservative approach is likely to underestimate the number of partially edited sites in favor of unedited or completely edited classifications (Mower and Palmer. 2006).

**Analysis**

The mitochondrial coding sequences of all five *Silene* species as well as *Beta vulgaris*, *Arabidopsis thaliana*, and *Oryza sativa* were aligned using ClustalW and adjusted manually. Regions that could not be reliably aligned in frame were excluded from subsequent analyses. RNA editing data for outgroup species were taken from REDIdb (Picardi *et al*. 2007). RNA editing data for *matR* in *Oryza sativa* were not available, so another monocot*, Triticum aestivum*, was used in its place.

To determine whether C-to-T substitutions occur preferentially at RNA editing sites, we compared the frequency of C-to-T substitutions at both RNA editing sites and 2-fold synonymous (non-edited) sites. The latter were chosen as a basis for comparison because they should be under the same pattern of selective constraint as editing sites with respect to protein sequence. In each case, C-to-A and C-to-G substitutions result in amino acid replacement, but C-to-T substitutions are silent. We analyzed two pairs of species, *Silene latifolia-Silene noctiflora* and *Arabidopsis thaliana-Beta vulgaris*. In each case, we used an outgroup (*Beta vulgaris* and *Oryza sativa*, respectively) and an unweighted parsimony criterion to infer the ancestral state for each site. C-to-T substitutions were then identified by comparing the sequence of each species against the inferred ancestral

state. *Beta* and *Arabidopsis* were chosen as a slowly evolving species pair to contrast with the rapid divergence observed within *Silene*. As members of the Caryophyllales and rosids respectively, *Beta* and *Arabidopsis* span one of the deepest splits in the eudicot phylogeny (Wikström *et al*. 2001). Therefore, they were expected to exhibit relatively substantial sequence divergence despite their slow substitution rates. Codons for which the amino acid was not conserved between the two ingroup species were excluded from this analysis. We used Fisher's exact test to assess the statistical significance of differences between frequencies of C-to-T substitutions at edited and synonymous sites.

To determine whether C-to-T substitutions at RNA editing sites are clustered within the *S. noctiflora* mitochondrial genome, we employed the method previously described by Roy and Gilbert (2005) for analyzing spatial patterns of intron loss. We calculated the probability distribution for the expected number of pairs of adjacent losses (i.e., C-to-T substitutions) given the number of lost and retained sites in each gene and the assumption that each loss occurs as an independent event. We compared the observed number of pairs of adjacent losses to this distribution to assess whether the lost editing sites were significantly clustered relative to a random expectation. For this analysis, we excluded 24 editing sites that were inferred to be lost in *S. noctiflora* by a change other than a C-to-T substitution.

Retroprocessing is expected to result in the loss of both introns and RNA editing sites. To determine whether observed intron losses were significantly associated with losses of flanking editing sites, we compared the observed number of losses to the following null distribution:

$$P[n] = (n+1)\left(\frac{r}{r+l}\right)\left(\frac{r-1}{r+l-1}\right)\prod_{i=0}^{n-1}\frac{l-i}{r+l-i-2} \qquad \text{(Eq. 1)}$$

where $P[n]$ is the probability of observing $n$ lost flanking editing sites around a given intron, and $r$ and $l$ are the total number of retained and lost editing sites in the genome, respectively. The $n + 1$ term in Equation 1 represents the number of possible ways to obtain $n$ flanking editing site losses. For example, there are three ways to lose two flanking sites (i.e., with the intron positioned either to the left, to the right, or in between the two sites). The rest of the expression on the right side of Equation 1 represents the probability of a single one of those patterns occurring.

## RESULTS

### Mitochondrial RNA Editing in *Silene*

The mitochondrial genomes of *S. latifolia* and *S. noctiflora* contain 26 and 27 putatively functional protein-coding genes, respectively (Table 1). The sole difference in intact gene content between the two species is *rps13*, which is intact in *S. noctiflora* but contains an internal stop codon in *S. latifolia*. Comparison of mitochondrial cDNA and genomic sequences revealed a total of 287 C-to-U edited sites in *S. latifolia* but only 189 sites in the rapidly-evolving *S. noctiflora* genome—a 34% difference (Table 1; Files S2 and S3). No other types of editing (e.g., U-to-C) were detected. The total number of editing sites in either *Silene* genome is lower than in any other angiosperm mitochondrial genome analyzed to date (Giege and Brennicke. 1999; Notsu *et al*. 2002; Handa. 2003; Mower and Palmer. 2006; Alverson *et al*. 2010; Picardi *et al*. 2010). This pattern results from

both a reduced number of mitochondrial genes in *Silene* and a lower average density of editing sites within each gene (Table 1).

By contrast, the substantial difference in RNA editing content between *S. latifolia* and *S. noctiflora* is caused entirely by a change in the density of editing sites rather than gene loss. The difference in editing site density derives mostly from sites that have been lost in *S. noctiflora* rather than gained in *S. latifolia*. There are 111 sites that are edited in *S. latifolia* but not edited in *S. noctiflora*. Of these, 106 sites (95.5%) are edited in *Beta* and/or *Arabidopsis*, suggesting that editing was the ancestral state. In contrast, there are only 13 sites that were edited in *S. noctiflora* but not in *S. latifolia*, and only 3 of these (23.1%) were edited in *Beta* and/or *Arabidopsis*. By far the most common pattern of editing site loss was C-to-T substitution at the genomic level, which obviates the need for editing (Figure 2). Although the total number of editing sites is reduced in *S. noctiflora*, the proportional distribution among genes is quite similar to *S. latifolia* and other angiosperms, indicating a genome-wide loss of mitochondrial RNA editing sites (Figure 3, Table 1).

Based on the divergence between *S. noctiflora* and *S. latifolia*, we found a significantly higher C-to-T substitution rate at editing sites ($d_{RE}$) than at two-fold synonymous sites ($d_{S2}$) ($d_{RE}/d_{S2} = 3.05$; $p < 0.001$). This pattern, however, was not unique to these two rapidly diverging *Silene* species. Comparison of two slowly evolving lineages (*Beta vulgaris* and *Arabidopsis thaliana*) also revealed a faster C-to-T substitution rate at RNA editing sites ($d_{RE}/d_{S2} = 3.55$; $p < 0.001$), and a similar analysis of monocot-eudicot divergence previously showed a comparable (~3.9-fold) excess of C-to-T substitutions at editing sites (Shields and Wolfe. 1997). When we restricted our

analysis to synonymous RNA editing sites ($d_{SRE}$), we still found rate elevation, but it was less pronounced. With so few synonymous editing sites, the rate increase was not significant in either the *Silene* comparison ($d_{SRE}/d_{S2}$ = 2.29; $p$ = 0.096) or the *Beta-Arabidopsis* comparison ($d_{SRE}/d_{S2}$ = 1.84; $p$ = 0.27). A similar (~2.4-fold) excess was previously found based on monocot-eudicot divergence (Shields and Wolfe. 1997).

We did not find any evidence for an increase in the frequency of incomplete editing associated with the history of mutational acceleration in the *S. noctiflora* mitochondrial genome. Both *Silene* species exhibited similar frequencies of partial editing (7.7% and 6.9% of editing sites in *S. latifolia* and *S. noctiflora*, respectively). As previously observed (Mower and Palmer. 2006), partial editing was common at synonymous sites (Files S2 and S3). Synonymous editing sites were 49 times more likely to be partially edited than non-synonymous sites in *S. latifolia* and 24 times more likely in *S. noctiflora*.

Analysis of other angiosperms has shown that editing sites are distributed in a clustered fashion along the length of mitochondrial genes (Mulligan *et al*. 2007). This pattern is also evident in *Silene*, but the observed *losses* of editing sites in *S. noctiflora* are not highly clustered (after taking into account the ancestrally clustered distribution of editing sites; Figure 4). We found only 21 adjacent pairs of lost editing sites specific to *S. noctiflora* (counting only sites that were lost by C-to-T substitution), which is not significantly different from the random expectation ($p$ = 0.31; Roy and Gilbert. 2005). This test for clustering of lost editing sites was similarly non-significant when the data were analyzed as individual exons ($p$ = 0.40) or as separately transcribed and *trans*-spliced fragments ($p$ = 0.42) rather than as whole genes.

However, there are two genes in *Silene* that exhibit clear examples of intron loss, and in each of these cases, losses are associated with multiple C-to-T substitutions at RNA editing sites in flanking exons. First, *S. latifolia* and *S. noctiflora* lack both *cox2* introns as well as the five surrounding editing sites present in *Beta* (Figure 5A). Although the second *cox2* intron is also absent in *Beta*, it is retained in other species within the Amaranthaceae (unpublished data), suggesting that it has been lost independently in the *Beta* and *Silene* lineages. Therefore, additional RNA editing sites surrounding the second *cox2* intron may have been lost in parallel between these lineages (Figure 5A). Overall, the *cox2* introns appear to have been lost independently numerous times during angiosperm evolution (Figure S1; Joly *et al*. 2001). In the second case of intron loss, *S. noctiflora* lacks the third *nad7* intron along with two adjacent editing sites (Figure 5b). These two cases suggest that retroprocessing led to the simultaneous loss of introns and flanking editing sites. For both genes, the observed number of lost flanking sites exceeds the random expectation. This difference is significant for *cox2* ($p < 0.001$) but not for *nad7* ($p = 0.23$).

*Silene conica*, another species exhibiting dramatic mitochondrial rate acceleration, contains even fewer RNA editing sites (28) than *S. noctiflora* (34) in a sample of seven mitochondrial genes (Figure 6, Table 2). The number of shared losses between *S. noctiflora* and *S. conica* (13) was significantly higher than expected at random ($p = 0.042$; Fisher's exact test), perhaps reflecting a shared phylogenetic history between the two species and/or variable selection and mutation pressures across sites. On the other hand, *S. noctiflora* and *S. conica* also exhibited a large number of unique losses (7 and

10, respectively), indicating that the specific sites being lost were largely (if not entirely) independent between the two lineages.

In contrast to their rapidly-evolving congeners, *S. vulgaris* (51) and *S. paradoxa* (53) maintain a set of editing sites that is nearly identical to that of *S. latifolia* (53) for the same seven-gene sample. These three slowly evolving species share an identical core set of 51 editing sites. In addition, *S. latifolia* and *S. paradoxa* share a *nad5* editing site that has been lost from *S. vulgaris*, while *S. latifolia* and *S. paradoxa* each have a unique site not found in the other two species. In comparison, *Beta*, *Arabidopsis*, and *Oryza* respectively maintain a total of 56, 62, and 67 editing sites in this same gene sample.

**Chloroplast RNA Editing in *Silene***

We sequenced both cDNA and genomic DNA for portions of the chloroplast genes *ndhB, psbL, rpoB, rpoC1,* and *rps2*. These loci represented all chloroplast genes that were both predicted to undergo RNA editing in *Spinacia* (Tsudzuki *et al*. 2001) and had genomic sequence available for at least one species of *Silene* to aid in primer design. In contrast to the mitochondrial pattern, *S. noctiflora* and *S. conica* did not exhibit higher synonymous substitution rates than their congeners for these chloroplast genes (Figure 1). *Silene latifolia*, *S. vulgaris* and *S. paradoxa* share an identical set of 14 chloroplast RNA editing sites across these five genes (Figure 6, Table 2). *Silene noctiflora* and *S. conica* each lost one site by C-to-T substitution (in *rpoB* and *psbL*, respectively). The loss of 1 out of 14 chloroplast editing sites (7%) represents a significantly lower rate of loss than observed in the mitochondrial genome for both *S. noctiflora* (35.8%; $p = 0.040$; Fisher's exact test)

and *S. conica* (45.1%; $p = 0.011$; Fisher's exact test). All five *Silene* species lacked the RNA editing site in *rpoC1* that is present in other eudicots.

## DISCUSSION

We found that recent increases in mitochondrial synonymous substitution rates within the genus *Silene* have been associated with substantial reductions in the frequency of RNA editing, adding further evidence for a relationship between mutation rate and the maintenance of RNA editing. These findings provide insight into the forces that guide the evolution of RNA editing. For example, as noted previously (Lynch *et al*. 2006), the negative relationship between mutation rate and the frequency of editing runs directly counter to predictions arising from the hypothesis that editing acts as a mutational buffer (Borner *et al*. 1997; Horton and Landweber 2002). In addition, these findings raise central questions about the relative importance of adaptive and non-adaptive processes in the evolution of RNA editing, which we discuss below.

### Elevated C-to-T Substitution Rates at RNA Editing Sites: The Role of Selection vs. Retroprocessing

More than a decade ago, Shields and Wolfe (1997) found that mitochondrial RNA editing sites undergo elevated rates of divergence among angiosperms. Mower (2008) subsequently showed that the high divergence rates are the result of frequent loss (rather than gain) of editing sites by C-to-T substitution. This loss appears to be driving a general decline in the frequency of mitochondrial RNA editing across the angiosperm phylogeny. This pattern has been interpreted as evidence for selection acting against RNA editing by

preferentially favoring the fixation of C-to-T mutations at editing sites. This interpretation, however, rests on the assumption that RNA editing sites experience the same mutation rates as other sites in the genome (Shields and Wolfe. 1997). Homologous recombination (gene conversion) with edited cDNA intermediates, i.e., retroprocessing, could violate this assumption by disproportionately affecting editing sites (Parkinson *et al*. 2005; Mulligan *et al*. 2007). There is only limited biochemical evidence of reverse transcriptase activity in angiosperm mitochondria (Moenne *et al*. 1996), but numerous convincing examples of the incorporation of reverse transcribed mRNA back into the mitochondrial genome, reflected by either the simultaneous loss of introns and editing sites in the flanking exons or the loss of an entire suite of editing sites across much or all of a mitochondrial gene. Examples of the former include the loss of introns and neighboring editing sites from *nad4* in both the Caryophyllales and Asterales (Geiss *et al*. 1994; Itchoda *et al*. 2002), two separate cases involving *rps3* in conifers (Ran *et al*. 2010), and the additional cases involving *cox2* and *nad7* reported in this study. Examples of editing site losses across much or all of a gene have been reported in numerous and diverse seed plant lineages (Krishnasamy *et al*. 1994; Lu *et al*. 1998; Petersen *et al*. 2006; Lopez *et al*. 2007). Moreover, it has been suggested that the clustered distribution of RNA editing sites in angiosperm mitochondrial genes reflects the elimination of stretches of previously intervening editing sites by retroprocessing (Mulligan *et al*. 2007).

The finding of both clustered (Figure 5) and dispersed (Figure 4) losses of editing sites in *Silene* raises intriguing possibilities with respect to the relative importance of selection and retroprocessing as forces acting to purge RNA editing sites. At one extreme, both forces could be operative but with selection predominating, i.e., with

retroprocessing occurring relatively rarely and only across large regions. At the other extreme, retroprocessing could be the predominant if not sole force driving the loss of RNA editing, acting across a wide range of spatial scales. The latter scenario would require that gene conversion acts on relatively short stretches of nucleotides in plant mitochondrial genomes, because we observed numerous cases where editing sites were lost by C-to-T substitution with no change at nearby sites. So-called "microconversions" are known to occur between short stretches of homologous sequence in other genomes and organisms (Wheeler *et al*. 1990; Semple and Wolfe. 1999; Palmer *et al*. 2003), while recent evidence indicates that microconversion (presumably DNA-mediated) readily occurs in plant mitochondrial genomes (Hao and Palmer. 2009; unpublished data). Single nucleotide substitutions at editing sites (independent of retroprocessing) and gene conversion with cDNA from partially edited transcripts would be expected to further reduce the clustering of lost editing sites.

Exactly how selection would act against RNA editing is not well established, but it would presumably involve one or both of the following selective pressures: 1) avoiding the deleterious consequences of failed editing (Lynch *et al*. 2006) and 2) eliminating the need for potentially costly site-specific editing machinery (Kotera *et al*. 2005; Zehrmann *et al*. 2009; Hammani *et al*. 2009). Both of these assume that editing is functionally important. Otherwise, there would be no cost to failed editing and no need to maintain editing machinery. Therefore, if selection has been acting to eliminate RNA editing, we would expect its effects to vary across sites depending upon the functional consequences of editing at those sites. In particular, we would predict that synonymous sites that undergo RNA editing should be relatively free from selection, because the editing

process has no effect on the resulting protein sequence. Therefore, if selection is the predominant force in the preferential C-to-T substitution at RNA editing sites, we would predict this pattern to be absent at synonymous editing sites. In contrast, retroprocessing should not discriminate between synonymous and non-synonymous edits in an mRNA transcript.

Consistent with previous studies (Shields and Wolfe. 1997; Mower. 2008), we found a trend suggesting that synonymous editing sites also experience preferential C-to-T substitutions, albeit at a lower rate than at non-synonymous editing sites. At first glance, the rate difference between synonymous and non-synonymous editing sites might seem inconsistent with a neutral model driven by retroprocessing. However, because synonymous editing sites are prone to both partial editing and frequent evolutionary reversion to unedited status—*unaccompanied* by C-to-T substitution—reverse transcription is less likely to capture the edited state at synonymous sites than at non-synonymous sites. Indeed, using a 5-state maximum likelihood model that accounted for the transitions between edited and unedited cytidines within a phylogenetic lineage, Mower (2008) found that rates of C-to-T substitution at synonymous edited sites are at least as high as at non-synonymous editing sites. Overall then, empirical RNA editing data at synonymous and non-synonymous sites are consistent with a neutral model based on retroprocessing. Although a role of selection certainly cannot be dismissed, any argument invoking selection to account for the elevated C-to-T substitution rates at RNA editing sites must explain why such selection similarly affects both non-synonymous and synonymous sites.

Of course, synonymous editing could have important effects on codon usage, mRNA stability or regulatory sequence conservation (Chamary *et al*. 2006). The pattern of RNA editing at synonymous sites does not, however, support such adaptive mechanisms. There are relatively few synonymous editing sites in *Silene* and other angiosperms (Table 1; Gray and Covello. 1993), most of which are only partially edited (Files S2 and S3; Mower and Palmer. 2006). The extent of editing at these sites is much more likely to vary among individuals and tissue types than at non-synonymous sites (Bentolila *et al*. 2008), and editing is frequently lost without the C-to-T substitution needed to conserve its effects (Shields and Wolfe. 1997; Mower. 2008). Overall, synonymous editing has all the hallmarks of a relatively neutral misfiring of the RNA editing machinery (Rüdinger *et al*. 2009).

The potential for selection on synonymous sites is also important because, as in previous studies, we have used divergence at (non-edited) synonymous sites as a baseline for comparison. The underlying assumption is that these sites are relatively free from selection and, therefore, provide a measure of the neutral substitution rate. If instead synonymous sites are subject to strong selection pressures based on mRNA function, it is possible that the apparent excess of substitutions at RNA editing sites in angiosperms is actually the result of negative selection acting on synonymous sites rather than positive selection or retroprocessing acting on RNA editing sites. In other words, the rate of C-to-T substitutions at RNA editing sites could better reflect the true neutral rate, because such changes are silent at the mRNA level (unlike substitutions at non-edited synonymous sites). Although the available evidence for selection on translational efficiency in plant

mitochondrial DNA is limited (Sloan and Taylor. 2010), this possibility should be considered as an alternative to the hypotheses analyzed in this study.

**An RNA-mediated Gene Conversion Model for the Loss of RNA Editing Sites**

We propose a model in which double-stranded breaks and gene conversion occur regularly in angiosperm mitochondrial genomes, perhaps playing a role in DNA repair. Occasional gene conversion with cDNA produced from edited mRNA (retroprocessing) would result in preferential C-to-T substitution at RNA editing sites. The length of DNA sequence typically affected by gene conversion would have to be short to explain why C-to-T substitutions at editing sites frequently occur without any change at nearby editing sites. Occasionally, however, gene conversion must affect much larger fragments, explaining the loss of large stretches of editing sites and intervening introns. Finally, in plant lineages with high mitochondrial mutation rates (e.g., *Silene noctiflora*), an elevated rate of DNA damage and double stranded breaks would accelerate the entire process.

While this model provides an attractive explanation of the data, the available evidence is still insufficient to fully support it or to exclude a role of selection in eliminating RNA editing sites in angiosperms. Separating the effects of selection and mutation bias is one of the most difficult and important challenges in the field of molecular evolution, and doing so will require additional studies that integrate phylogenetic, population genetic and mechanistic data.

**Reduced RNA Editing Content in *Silene* Species with Low Substitution Rates**

Although *S. latifolia* has a higher density of mitochondrial RNA editing sites than its rapidly evolving congeners, it maintains fewer sites than any other angiosperm analyzed to date (Table 1). The similarity between *S. latifolia* and other slowly evolving *Silene* species in the number and identity of mitochondrial RNA editing sites (see Results and Table 2) suggests that the *Silene* ancestor likely had a smaller complement of RNA editing sites than the typical angiosperm. Like the larger reductions in editing observed in *S. noctiflora* and *S. conica*, this pattern may be at least partially related to changes in mitochondrial substitution rate. Although dwarfed by recent accelerations in some *Silene* species (Figure 1), the ancestral mitochondrial substitution rate in the genus appears to be significantly higher than that of related angiosperms. Comparison of all mitochondrial protein gene sequences reveals that the *S. latifolia* lineages has exhibited a 56% higher synonymous substitution rate than the *Beta vulgaris* lineage (Figure S2), which itself has experienced a higher substitution rate than other fully sequenced angiosperm mitochondrial genomes (Alverson *et al*. 2010). Therefore, the relationship between RNA editing and evolutionary rates may extend to finer scales of rate variation (see also Alverson *et al*. 2010).

**Mutagenic Retroprocessing and Elevated Substitution Rates**

Parkinson *et al.* (2005) suggested that, because of the nearly complete loss of RNA editing sites in the few mitochondrial genes sequenced in *Pelargonium* and the high error rate of reverse transcriptase, an increase in retroprocessing itself might be at least part of the cause of mitochondrial mutation rate acceleration in *Pelargonium*. Based on a previous analysis that found little change in the frequency of RNA editing despite the

mutational acceleration in *Silene noctiflora*, it seemed unlikely that this mechanism would apply to *Silene* (Mower *et al*. 2007). However, given the evidence from the present study that *S. noctiflora* has, in fact, lost a substantial fraction of its RNA editing sites, it is worth reconsidering this possibility. If mutagenic retroprocessing has been a major cause of high mitochondrial mutation rates in *Silene*, we would expect to see a disproportionate increase in C-to-T substitution rate for sites that are edited in the mRNA sequence. We did not, however, observe such an effect in *S. noctiflora*, as the increase in substitution rate at RNA editing sites appears roughly proportional to the rate increase at non-edited sites. Therefore, an increase in the frequency of retroprocessing is unlikely to be the primary *cause* of the high mitochondrial mutation rates in *Silene*. Whether mitochondrial rate accelerations in other angiosperms (e.g., *Plantago* and *Pelargonium*) are the result of mutagenic retroprocessing remains an open question that is subject to testable predictions (Parkinson *et al*. 2005).

**Envoi**

We have found that increases in mitochondrial mutation rate in *Silene* are correlated with a rapid loss of RNA editing sites. However, rather than fundamentally altering the evolutionary forces that act on RNA editing, the high mutation rates in these lineages appear to have simply accelerated a pre-existing pattern in angiosperm evolution, a pattern that might depend more on a neutral model involving RNA-mediated gene conversion than on selection. Stephen Jay Gould (1990) famously asked what would happen if we were able to rewind and replay the tape of evolution. The mutational acceleration in some *Silene* mitochondrial genomes may instead allow us to effectively

fast-forward that tape and thereby provide a glimpse into the future of RNA editing in angiosperms.

ACKNOWLEDGEMENTS

We would like to thank Jeff Mower for his insightful comments on an earlier version of our manuscript and Michael Hood for providing *S. paradoxa* and *S. conica* seeds. This study was supported by NSF DEB-0808452 (to DBS and DRT) and NIH RO1-GM-70612 (to JDP). AJA was supported by an NIH Ruth L. Kirschstein NRSA Postdoctoral Fellowship (1F32GM080079-01A1).

REFERENCES

Alverson, A. J., X. Wei, D. W. Rice, D. B. Stern, K. Barry *et al*, 2010 Insights into the evolution of plant mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Mol. Biol. Evol. doi:10.1093/molbev/msq029

Barr, C. M., S. R. Keller, P. K. Ingvarsson, D. B. Sloan and D. R. Taylor, 2007 Variation in mutation rate and polymorphism among mitochondrial genes in *Silene vulgaris*. Mol. Biol. Evol. **24:** 1783-1791.

Bentolila, S., L. E. Elliott and M. R. Hanson, 2008 Genetic architecture of mitochondrial editing in *Arabidopsis thaliana* . Genetics **178:** 1693-1708.

Bernasconi, G., J. Antonovics, A. Biere, D. Charlesworth, L. F. Delph *et al*, 2009 *Silene* as a model system in ecology and evolution. Heredity **103:** 5-14.

Borner, G. V., S. Yokobori, M. Morl, M. Dorner and S. Paabo, 1997 RNA editing in metazoan mitochondria: Staying fit without sex. FEBS Lett. **409:** 320-324.

Burt, A., and R. Trivers, 2006 *Genes in Conflict: The Biology of Selfish Genetic Elements*. Belknap Press.

Chamary, J. V., J. L. Parmley and L. D. Hurst, 2006 Hearing silence: Non-neutral evolution at synonymous sites in mammals. Nat. Rev. Genet. **7:** 98-108.

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

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

Covello, P. S., and M. W. Gray, 1993 On the evolution of RNA editing. Trends Genet. **9:** 265-268.

Doyle, J. J., and J. L. Doyle, 1987 A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical bulletin **19:** 11-15.

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

Erixon, P., and B. Oxelman, 2008 Reticulate or tree-like chloroplast DNA evolution in *Sileneae* (Caryophyllaceae)? Mol. Phylogenet. Evol. **48:** 313-325.

Farajollahi, S., and S. Maas. 2010. Molecular diversity through RNA editing: a balancing act. Trends Genet. doi:10.1016/j.tig.2010.02.001.

Fiebig, A., S. Stegemann and R. Bock, 2004 Rapid evolution of RNA editing sites in a small non-essential plastid gene. Nucleic Acids Res. **32:** 3615-3622.

Geiss, K. T., G. M. Abbas and C. A. Makaroff, 1994 Intron loss from the NADH dehydrogenase subunit 4 gene of lettuce mitochondrial DNA: Evidence for homologous recombination of a cDNA intermediate. Mol. Gen. Genet. **243:** 97-105.

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

Gommans, W. M., S. P. Mullen and S. Maas, 2009 RNA editing: A driving force for adaptive evolution? BioEssays **31:** 1137-1145.

Gott, J. M., 2003 Expanding genome capacity via RNA editing. C. R. Biol. **326:** 901-908.

Gould, S. J., 1990 *Wonderful Life: The Burgess Shale and the Nature of History*. WW Norton & Company.

Gray, M. W., 2003 Diversity and evolution of mitochondrial RNA editing systems. IUBMB Life **55:** 227-233.

Gray, M. W., and P. S. Covello, 1993 RNA editing in plant mitochondria and chloroplasts. FASEB J. **7:** 64-71.

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

Hammani, K., K. Okuda, S. K. Tanz, A. L. Chateigner-Boutin, T. Shikanai *et al*, 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.

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

Hao, W., and J. D. Palmer, 2009 Fine-scale mergers of chloroplast and mitochondrial genes create functional, transcompartmentally chimeric mitochondrial genes. Proc. Natl. Acad. Sci. **106:** 16728-16733.

Hirose, T., T. Kusumegi, T. Tsudzuki and M. Sugiura, 1999 RNA editing sites in tobacco chloroplast transcripts: Editing as a possible regulator of chloroplast RNA polymerase activity. Mol. Gen. Genet. **262:** 462-467.

Horton, T. L., and L. F. Landweber, 2002 Rewriting the information in DNA: RNA editing in kinetoplastids and myxomycetes. Curr. Opin. Microbiol. **5**: 620-626.

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

Jobson, R. W., and Y. L. Qiu, 2008 Did RNA editing in plant organellar genomes originate under natural selection or through genetic drift? Biol. Direct **3:** 43.

Joly, S., L. Brouillet and A. Bruneau, 2001 Phylogenetic implications of the multiple losses of the mitochondrial *coxII.i3* intron in the angiosperms. Int. J. Plant Sci. **162:** 359-373.

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

Krishnasamy, S., R. A. Grant and C. A. Makaroff, 1994 Subunit 6 of the Fo-ATP synthase complex from cytoplasmic male-sterile radish: RNA editing and NH2-terminal protein sequencing. Plant Mol. Biol. **24:** 129-141.

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

Li, L., B. Wang, Y. Liu and Y. L. Qiu, 2009 The complete mitochondrial genome sequence of the hornwort *Megaceros aenigmaticus* shows a mixed mode of conservative yet dynamic evolution in early land plant mitochondrial genomes. J. Mol. Evol. **68:** 665-678.

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

Lu, M. Z., A. E. Szmidt and X. R. Wang, 1998 RNA editing in gymnosperms and its impact on the evolution of the mitochondrial coxI gene. Plant Mol. Biol. **37:** 225-234.

Lynch, M., 2007 *The Origins of Genome Architecture*. Sinauer Associates, Sunderland, MA.

Lynch, M., 2006 Streamlining and simplification of microbial genome architecture. Annu. Rev. Microbiol. **60:** 327-349.

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

Malek, O., K. Lattig, R. Hiesel, A. Brennicke and V. Knoop, 1996 RNA editing in bryophytes and a molecular phylogeny of land plants. EMBO J. **15:** 1403-1411.

Moenne, A., D. Begu and X. Jordana, 1996 A reverse transcriptase activity in potato mitochondria. Plant Mol. Biol. **31:** 365-372.

Mower, J., 2005 PREP-mt: Predictive RNA editor for plant mitochondrial genes. BMC Bioinformatics **6:** 96.

Mower, J. P., 2008 Modeling sites of RNA editing as a fifth nucleotide state reveals progressive loss of edited sites from angiosperm mitochondria. Mol. Biol. Evol. **25:** 52-61.

Mower, J. P., and J. D. Palmer, 2006 Patterns of partial RNA editing in mitochondrial genes of *Beta vulgaris*. Mol. Genet. Genomics **276:** 285-293.

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

Mulligan, R. M., K. L. C. Chang and C. C. Chou, 2007 Computational analysis of RNA editing sites in plant mitochondrial genomes reveals similar information content and a sporadic distribution of editing sites. Mol. Biol. Evol. **24:** 1971-1981.

Notsu, Y., S. Masood, T. Nishikawa, N. Kubo, G. Akiduki *et al*, 2002 The complete sequence of the rice (*Oryza sativa* L.) mitochondrial genome: Frequent DNA sequence acquisition and loss during the evolution of flowering plants. Mol. Genet. Genomics **268:** 434-445.

Palmer, J. D., and L. A. Herbon, 1988 Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J. Mol. Evol. **28:** 87-97.

Palmer, S., E. Schildkraut, R. Lazarin, J. Nguyen and J. A. Nickoloff, 2003 Gene conversion tracts in *Saccharomyces cerevisiae* can be extremely short and highly directional. Nucleic Acids Res. **31**: 1164-1173.

Parkinson, C. L., J. P. Mower, Y. L. Qiu, A. J. Shirk, K. Song *et al*, 2005 Multiple major increases and decreases in mitochondrial substitution rates in the plant family Geraniaceae. BMC Evol. Biol. **5:** 73.

Petersen, G., O. Seberg, J. I. Davis and D. W. Stevenson, 2006 RNA editing and phylogenetic reconstruction in two monocot mitochondrial genes. Taxon **55:** 871-886.

Picardi, E., T. M. Regina, A. Brennicke and C. Quagliariello, 2007 REDIdb: The RNA editing database. Nucleic Acids Res. **35:** D173-7.

Picardi, E., D. S. Horner, M. Chiara, R. Schiavon, G. Valle, and G. Pesole, 2010 Large-scale detection and analysis of RNA editing in grape mtDNA by RNA deep-sequencing. doi:10.1093/nar/gkq202.

Powell, L. M., S. C. Wallis, R. J. Pease, Y. H. Edwards, T. J. Knott *et al*, 1987 A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell **50:** 831-840.

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

Roy, S. W., and W. Gilbert, 2005 The pattern of intron loss. Proc. Natl. Acad. Sci. **102:** 713-718.

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

Rüdinger, M., H. T. Funk, S. A. Rensing, U. G. Maier and V. Knoop, 2009 RNA editing: Only eleven sites are present in the *Physcomitrella patens* mitochondrial transcriptome and a universal nomenclature proposal. Mol. Genet. Genomics **281:** 473-481.

Salone, V., M. Rüdinger, M. Polsakiewicz, B. Hoffmann, M. Groth-Malonek *et al*, 2007 A hypothesis on the identification of the editing enzyme in plant organelles. FEBS Lett. **581:** 4132-4138.

Semple, C., and K. H. Wolfe, 1999 Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. J. Mol. Evol. **48**: 555-564.

Shields, D. C., and K. H. Wolfe, 1997 Accelerated evolution of sites undergoing mRNA editing in plant mitochondria and chloroplasts. Mol. Biol. Evol. **14:** 344-349.

Sloan, D. B., and D. R. Taylor, 2010 Testing for selection on synonymous sites in plant mitochondrial DNA: The role of codon bias and RNA editing. J. Mol. Evol. doi: 10.1007/s00239-010-9346-y.

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

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

Steinhauser, S., S. Beckert, I. Capesius, O. Malek and V. Knoop, 1999 Plant mitochondrial RNA editing. J. Mol. Evol. **48:** 303-312.

Tillich, M., P. Lehwark, B. R. Morton and U. G. Maier, 2006 The evolution of chloroplast RNA editing. Mol. Biol. Evol. **23:** 1912-1921.

Tsudzuki, T., T. Wakasugi and M. Sugiura, 2001 Comparative analysis of RNA editing sites in higher plant chloroplasts. J. Mol. Evol. **53:** 327-332.

Turmel, M., C. Otis and C. Lemieux, 2003 The mitochondrial genome of *Chara vulgaris*: Insights into the mitochondrial DNA architecture of the last common ancestor of green algae and land plants. Plant Cell **15:** 1888-1903.

Wheeler, C. J., D. Maloney, S. Fogel and R. S. Goodenow, 1990 Microconversion between murine H-2 genes integrated into yeast. Nature **347:** 192-194.

Wikström, N., V. Savolainen and M. W. Chase, 2001 Evolution of the angiosperms: Calibrating the family tree. Proc. R. Soc. Lond. B. **268:** 2211.

Wolf, P. G., C. A. Rowe and M. Hasebe, 2004 High levels of RNA editing in a vascular plant chloroplast genome: Analysis of transcripts from the fern *Adiantum capillus-veneris* . Gene **339:** 89-97.

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

Yang, Z., 2007 PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24:** 1586-1591.

Yu, W., and W. Schuster, 1995 Evidence for a site-specific cytidine deamination reaction involved in C to U RNA editing of plant mitochondria. J. Biol. Chem. **270:** 18227-18233.

Yura, K., and M. Go, 2008 Correlation between amino acid residues converted by RNA editing and functional residues in protein three-dimensional structures in plant organelles. BMC Plant. Biol. **8:** 79.

Zehrmann, A., D. Verbitskiy, J. A. van der Merwe, A. Brennicke and M. Takenaka, 2009 A DYW domain-containing pentatricopeptide repeat protein is required for RNA editing at multiple sites in mitochondria of *Arabidopsis thaliana*. Plant Cell **21:** 558-567.

**Table 1**. RNA editing sites in the mitochondrial protein genes of four eudicots. The number of synonymous editing sites is in parentheses. NP indicates that a functional copy of the gene is not present in the mitochondrial genome of that species (although putative pseudogenes are still present in some cases).

| | Length (bp)[a] | *Arabidopsis* | *Beta* | *S. latifolia* | *S. noctiflora* |
|---|---|---|---|---|---|
| **Complex I** | | | | | |
| *nad1* | 978 | 24 (4) | 20 (2) | 19 (2) | 11 (1) |
| *nad2* | 1467 | 32 (8) | 24 (4) | 21 (3) | 18 (3) |
| *nad3* | 357 | 12 (2) | 12 (2) | 8 (1) | 5 (1) |
| *nad4* | 1488 | 32 (5) | 19 (2) | 16 (0) | 11 (0) |
| *nad4L* | 303 | 9 (0) | 10 (0) | 9 (0) | 6 (1) |
| *nad5* | 2019 | 27 (3) | 17 (0) | 18 (0) | 15 (0) |
| *nad6* | 570 | 10 (0) | 11 (2) | 10 (2) | 6 (0) |
| *nad7* | 1173 | 27 (7) | 20 (0) | 19 (1) | 9 (1) |
| *nad9* | 579 | 7 (0) | 5 (0) | 5 (0) | 1 (0) |
| **Complex III** | | | | | |
| *cob* | 1164 | 7 (0) | 13 (3) | 9 (1) | 6 (0) |
| **Complex IV** | | | | | |
| *cox1* | 1572 | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| *cox2* | 777 | 15 (4) | 9 (1) | 3 (0) | 2 (0) |
| *cox3* | 798 | 8 (1) | 4 (2) | 1 (0) | 1 (0) |
| **Complex V** | | | | | |
| *atp1* | 1509 | 5 (1) | 3 (0) | 3 (0) | 0 (0) |
| *atp4* | 555 | 8 (2) | 12 (2) | 11 (1) | 6 (0) |
| *atp6* | 720 | 1 (0) | 12 (0) | 11 (1) | 7 (0) |
| *atp8* | 492 | 0 (0) | 2 (1) | 2 (1) | 2 (1) |
| *atp9* | 225 | 4 (0) | 5 (1) | 4 (1) | 1 (0) |
| **Cytochrome C biogenesis** | | | | | |
| *ccmB* | 621 | 39 (7) | 30 (6) | 27 (4) | 19 (2) |
| *ccmC* | 666 | 28 (3) | 28 (6) | 23 (3) | 15 (1) |
| *ccmFc* | 1341 | 16 (2) | 13 (2) | 12 (1) | 8 (1) |
| *ccmFn* | 1743 | 34 (2)[b] | 23 (1) | 22 (0) | 16 (2) |
| **Ribosomal proteins** | | | | | |
| *rpl2* | 1050 | 1 (1) | NP | NP | NP |
| *rpl5* | 558 | 10 (1) | 5 (0) | 6 (0) | 4 (0) |
| *rpl16* | 405 | 5 (0)[c] | NP | NP | NP |
| *rps3* | 1050 | 13 (5) | 6 (0) | 4 (0)[d] | 3 (0)[d] |
| *rps4* | 1089 | 15 (1) | 11 (1) | NP | NP |
| *rps7* | 447 | 0 (0) | 3 (1) | NP | NP |
| *rps12* | 378 | 8 (2) | 6 (0) | NP | NP |
| *rps13* | 351 | NP | 2 (1) | 1 (0)[e] | 0 (0) |
| **Other protein genes** | | | | | |
| *matR* | 1986 | 9 (0) | 9 (1) | 8 (1) | 6 (1) |
| *mttB* | 753 | 24 (3) | 19 (0) | 15 (0) | 11 (1) |

| | | | | |
|---|---|---|---|---|
| **Total** | **430 (64)** | **353 (41)** | **287 (23)** | **189 (16)** |
| **Coding genes** | **31** | **30** | **26** | **27** |
| **Editing sites/100bp** | **1.49** | **1.27** | **1.13** | **0.73** |

[a]The reported length for each gene is based on a multiple sequence alignment and is therefore affected by both internal alignment gaps and trimming of non-homologous 5' and 3' ends. Editing sites in unalignable regions were included in species counts.

[b]The *ccmFn* gene in Arabidopsis is divided into two separate genes.

[c]The annotations for *rps3* and *rpl16* overlap in *Arabidopsis*. The overlapping region contains three editing sites, which were excluded from the *rpl16* count.

[d]Both *Silene* species lack a substantial 5' portion of *rps3* relative to other angiosperms. It is possible that *rps3* is a pseudogene in *Silene*.

[e]The *rps13* gene in *S. latifolia* contains an internal stop codon and is most likely a pseudogene.

**Table 2**. Mitochondrial and chloroplast RNA editing in five *Silene* species.

| | Length analyzed (bp) | *Silene conica* | *Silene noctiflora* | *Silene latifolia* | *Silene vulgaris* | *Silene paradoxa* |
|---|---|---|---|---|---|---|
| **Mitochondrial** | | | | | | |
| *ccmFn* | 221 | 5 | 8 | 7 | 7 | 8 |
| *cob* | 508 | 2 | 3 | 6 | 6 | 6 |
| *nad2* | 462 | 7 | 7 | 10 | 9 | 9 |
| *nad5* | 520 | 5 | 7 | 10 | 9 | 10 |
| *nad6* | 470 | 5 | 5 | 9 | 9 | 9 |
| *nad7* | 419 | 3 | 3 | 6 | 6 | 6 |
| *nad9* | 434 | 1 | 1 | 5 | 5 | 5 |
| **Total (mt)** | **3034** | **28** | **34** | **53** | **51** | **53** |
| | | | | | | |
| **Chloroplast** | | | | | | |
| *ndhB* | 996 | 9 | 9 | 9 | 9 | 9 |
| *psbL* | 77 | 0 | 1 | 1 | 1 | 1 |
| *rpoB* | 382 | 3 | 2 | 3 | 3 | 3 |
| *rpoC1* | 260 | 0 | 0 | 0 | 0 | 0 |
| *rps2* | 272 | 1 | 1 | 1 | 1 | 1 |
| **Total (cp)** | **1987** | **13** | **13** | **14** | **14** | **14** |

# FIGURES

**FIGURE 1.** Mitochondrial rate accelerations in *Silene*. Branch lengths correspond to the number of synonymous substitutions per site for mitochondrial DNA (A) and chloroplast DNA (B). Analyses were performed in PAML v4.1 (Yang. 2007) on concatenated datasets (see Table 2) using a codon-based model of evolution and constrained topology as described previously (Sloan *et al*. 2009).

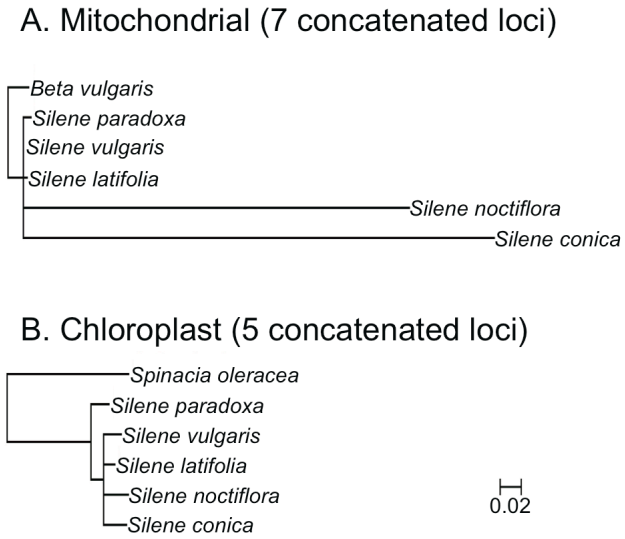


**FIGURE 2.** The distribution of nucleotide states at 106 positions where a C-to-U RNA editing site was inferred to be lost in *S. noctiflora*. Grey indicates sites of amino acid conservation between *S. latifolia* and *S. noctiflora*, while black indicates amino acid difference.

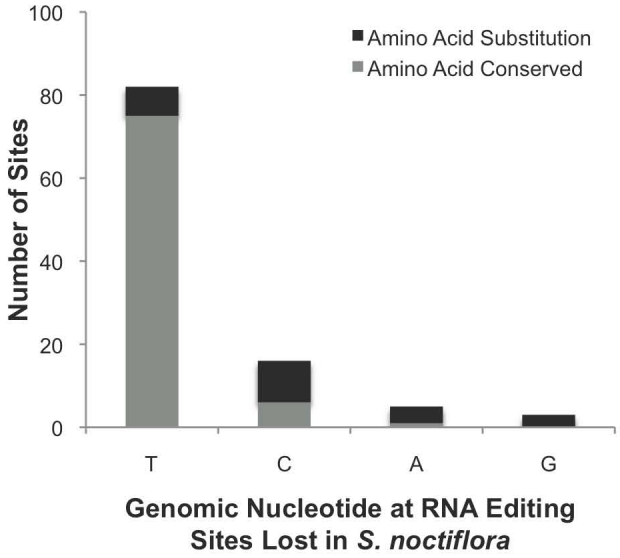


**FIGURE 3.** Loss of RNA editing sites in rapidly-evolving mitochondrial genes of *S. noctiflora*. Each point represents a single gene. Genes falling below the 1:1 line have fewer editing sites in *S. noctiflora* than in *S. latifolia*.

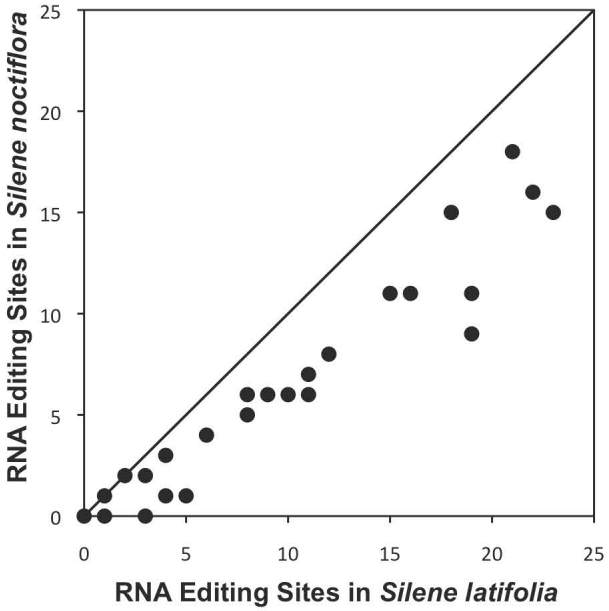


**FIGURE 4.** Distribution of lost and retained RNA editing sites in *Silene noctiflora*. Vertical lines indicate ancestral RNA editing sites that have been retained in *S. noctiflora*. Open triangles indicate ancestral editing sites that have been lost in *S. noctiflora* (but not in *S. latifolia*) by C-to-T substitution at the genomic level. Editing sites lost by other mechanisms and putative editing site gains are not shown. Filled black and grey triangles indicate the presence of *cis*- and *trans*-splicing introns, respectively. Position values exclude intron sequences.

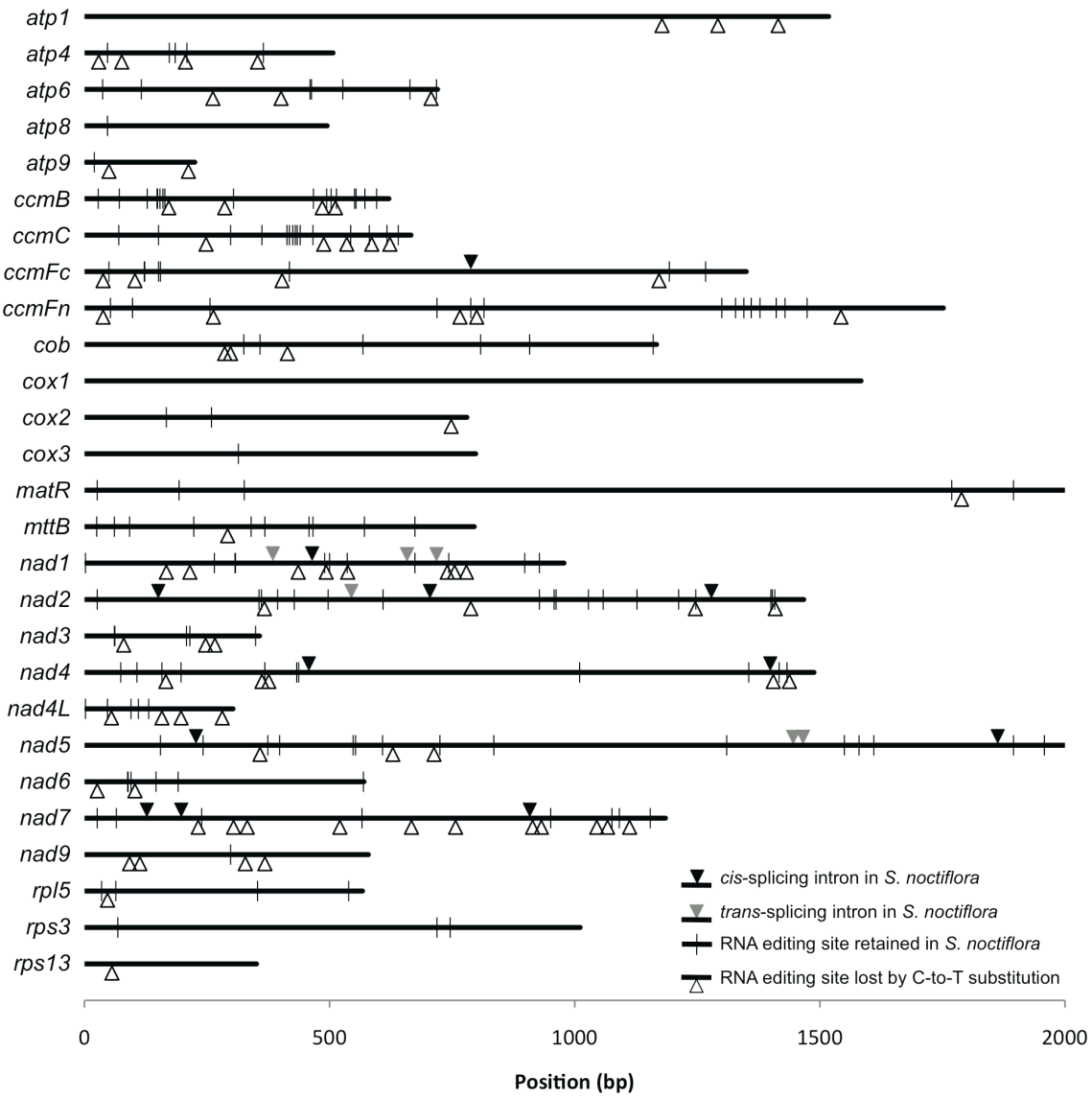


**FIGURE 5.** Loss of introns in *cox2* (A) and *nad7* (B) is associated with loss of adjacent RNA editing sites in *Silene*. Black triangles and vertical lines indicate introns and editing sites, respectively. Grey shading covers regions surrounding lost introns in which RNA editing sites have also been lost by C-to-T substitution. A more detailed phylogenetic distribution of *cox2* RNA editing sites and introns is provided in Figure S1. RNA editing data for outgroup species were obtained from REDIdb (Picardi *et al*. 2007). Position values exclude intron sequences.

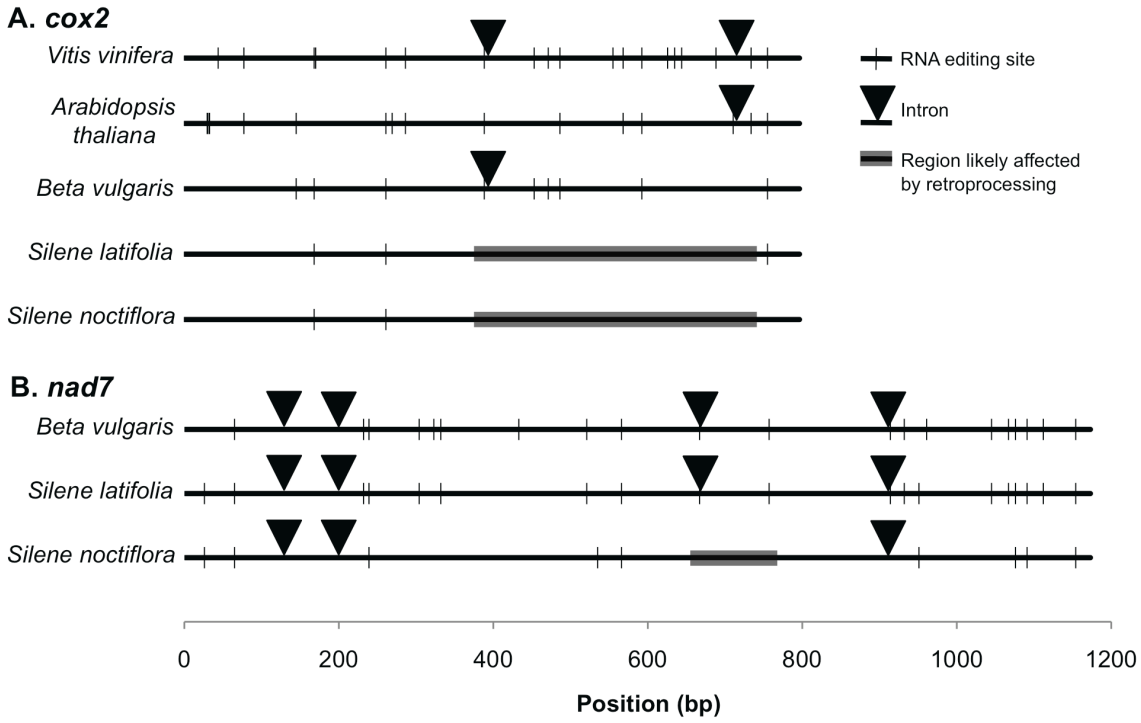
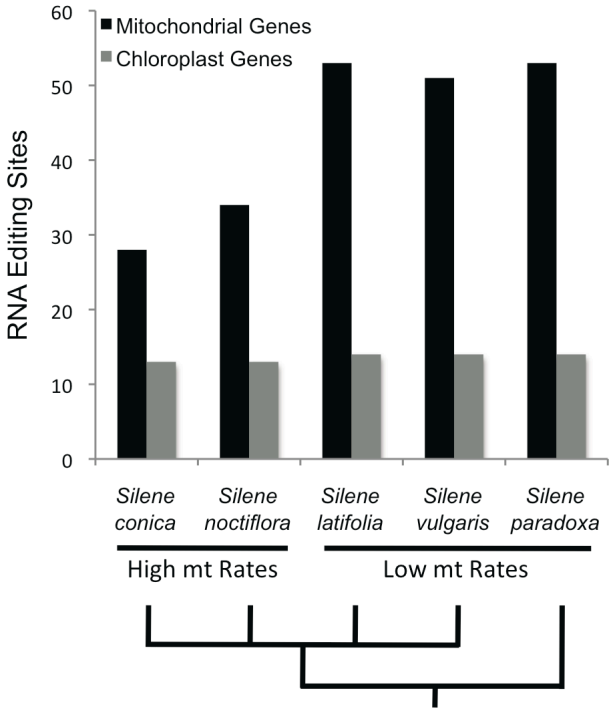


**FIGURE 6.** Number of RNA editing sites in portions of selected (see Table 2) mitochondrial genes (black bars) and chloroplast genes (grey bars) in five *Silene* species. The current understanding of the phylogenetic relationships among these species is depicted below (Erixon and Oxelman. 2008; Sloan *et al*. 2009; A. Rautenberg, D.B. Sloan, V. Aldén and B. Oxelman, unpublished data).



# Chapter 7.

# Rapid evolution of genomic obesity in 'mutator' mitochondria of flowering plants[1]

[1]Formatted as a co-authored manuscript (Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR) tentatively planned for submission to *Science*.

ABSTRACT

Comparisons across diverse eukaryotes have suggested that increased mutational pressure selects for reduced genome size and complexity. To test this hypothesis, we sequenced mitochondrial genomes from two species with recently and dramatically accelerated mitochondrial mutation rates in the angiosperm genus *Silene*. Surprisingly, both genomes exhibit greatly expanded non-coding content. At 6.7 and 11.3 Mb, they are the largest known mitochondrial genomes, whereas slowly evolving *Silene* mitochondrial genomes are smaller than average for angiosperms. Consequently, this genus contains ~98% of known variation in organelle genome size. The expanded genomes exhibit numerous architectural changes, including novel multi-circular structures and qualitative differences in recombinational activity. The evolution of mutation, recombination, and genome structure can therefore be extremely rapid and interrelated in ways not predicted by current evolutionary theories.

MAIN TEXT

Mitochondrial genomes display striking diversity in size, structure and functional organization *(1)*, mirroring broader patterns of variation in genome architecture *(2, 3)*. For example, in contrast to the small and streamlined genomes found in the mitochondria of bilaterian animals (typically 14-20 kb) *(4)*, seed plant mitochondrial genomes (200-2900 kb) contain abundant intergenic and intronic sequences *(5)*. Plant mitochondrial genomes are also characterized by extremely low point mutation rates, further distinguishing them from their fast-evolving animal counterparts *(6)*. This observation motivated the hypothesis that differences in mutation rate are a major determinant of variation in organelle genome architecture *(7)*. This argument represents an extension of the mutational burden hypothesis *(3, 8)*, which posits that non-coding elements generate a mutational liability and proliferate primarily by non-adaptive means, as genetic drift overwhelms the weak selective costs associated with the increased chance of mutational disruption of genome function (e.g., elimination of conserved sequence required for intron splicing or generation of improper transcription factor binding sites in intergenic regions). Because the magnitude of such costs should be directly proportional to the mutation rate, fast-evolving genomes are predicted to experience more intense selection for genomic reduction. The discovery that some angiosperms have dramatically accelerated mitochondrial mutation rates, sometimes orders of magnitude higher than closely related species *(9)*, presents a unique opportunity to test the predicted association between high mutation rate environments and the evolution of streamlined genomes.

Within the last 5-10 Myr, several species in the genus *Silene* (Caryophyllaceae) have experienced dramatic increases in mitochondrial mutation rates *(9, 10)*. We

sequenced purified mitochondrial DNA (mtDNA) from three *Silene* species, yielding complete genome assemblies for *S. noctiflora* and *S. vulgaris* and a high quality draft assembly for *S. conica*. We also included the previously published mitochondrial genome of *S. latifolia* in our analyses *(11)*. The genomic data confirm that *S. noctiflora* and *S. conica* have experienced massive accelerations in nucleotide substitution rates (Fig. 1, Fig. S1) and correlated increases in the frequency of insertions and deletions (indels) (Figs. S2-S3).

Contrary to predictions of genomic reduction, the fast-evolving mitochondrial genomes of *S. noctiflora* and *S. conica* have experienced unprecedented expansions, resulting in sizes of 6.7 Mb and 11.3 Mb, respectively (Fig. 2). Remarkably, they are larger than most bacterial genomes and even some nuclear genomes. In contrast, the more slowly evolving mitochondrial genomes of *S. latifolia* (0.25 Mb) and *S. vulgaris* (0.43 Mb) are typical, if not below average in size, for angiosperms. *Silene* mitochondrial genome sizes have thus diverged by more than 40-fold in just a few million years.

The genomic expansion in *S. noctiflora* and *S. conica* cannot be explained by detectable increases in gene or intron content. Although these genomes contain duplicate copies of some genes (particularly rRNA genes; Table S1), they possess fewer unique genes than other angiosperm mitochondrial genomes (Figs. 1, S4-S5). Notably, the *S. conica* and *S. noctiflora* mitochondrial genomes contain only two or three tRNA genes, which is far fewer than most angiosperms and even below the already reduced tRNA gene content found in other *Silene* species (Fig. 1, Fig. S5) *(11)*.The four *Silene* genomes have nearly identical sets of introns (Table 1) with only one observed intron loss (the third *nad4* intron in *S. noctiflora*), and average intron lengths in the expanded *S.*

*noctiflora* and *S. conica* genomes are actually ~10% shorter than in *S. latifolia* and *S. vulgaris* (Fig. S6). The contraction of introns in these otherwise expanded genomes suggests that, unlike in many eukaryotic *(12)* and prokaryotic systems *(13)*, the variation in abundance of non-coding content in *Silene* mitochondrial genomes is not dictated by the relative frequency and size of small insertion and deletion mutations (Fig. S3).

Intergenic sequences account for 99% of the bloated mitochondrial genomes in *S. noctiflora* and *S. conica*. As in other vascular plants *(5, 14)*, the intergenic regions of all four *Silene* mitochondrial genomes contain sequences of both nuclear and plastid (chloroplast) origin. Although the expanded mitochondrial genomes of *S. noctiflora* and *S. conica* contain more of this "promiscuous" DNA than their smaller *Silene* counterparts (Table 1), contributions from these sources do not scale proportionately with the increases in genome size and constitute less than 1% of the intergenic content in both species (Table 1). More substantial components (13% and 7%) of the intergenic regions in each of these two genomes exhibit similarity to sequences from other plant mitochondrial genomes (Table 1). Most of this conserved sequence (>650 kb) is only shared between *S. noctiflora* and *S. conica*, potentially supporting a common origin of the large genomes and high mutation rates in these species, an issue which is currently unresolved by molecular phylogeny (Supplementary Text). Nonetheless, >85% of the voluminous intergenic sequence in these two species lacks detectable homology with any of the nuclear, plastid, or mitochondrial sequences available in the GenBank nr/nt database.

Repeated sequences constitute a variable and often large component of seed plant mitochondrial genomes *(15)*, and *Silene* species are exceptional in both respects (Table 1,

Figs. S7-S9). The *S. conica* mitochondrial genome contains a remarkable 4.6 Mb (40.8%) of dispersed repeats, which is more than any other sequenced plant mitochondrial genome in both absolute and percentage terms *(15)*. The largest repeats are >80 kb, but the bulk of the repetitive content consists of an enormous number of small, imperfect, and often overlapping repeats (Fig. S7-S9). In contrast, repeat sequences make up no more than 19% of any of the other three *Silene* mitochondrial genomes. Notably, proportional repeat coverage in the relatively small *S. vulgaris* genome (18.8%; 80 kb of repeats) exceeds that of the much larger *S. noctiflora* genome (10.9%; 735 kb), indicating that there is no absolute relationship between repetitive content and genome size in this group.

*Silene noctiflora* and *S. conica* also have evolved novel mitochondrial genome structures. Although the relationship between genome maps and *in vivo* physical structure remains uncertain for angiosperm mtDNAs *(16)*, they usually map as an equilibrium mixture consisting of a "master circle", comprising the entire sequence content of the genome, and a collection of two or more "subgenomic circles" that arise via high-frequency recombination between large direct repeats *(17, 18)*. This model applies to *S. latifolia (11)* and to the vast majority of *S. vulgaris* mitochondrial genome content (see Supplementary Materials and Methods and Table S2). In contrast, the *S. noctiflora* and *S. conica* mitochondrial genomes each assemble into dozens of largely autonomous and relatively small, circular-mapping chromosomes. The *S. noctiflora* mitochondrial genome consists of 59 distinct circular chromosomes ranging from 66 to 192 kb (Table S2). Many of these do not share any large repeats (>1 kb) with other chromosomes. Even in cases where two or more *S. noctiflora* chromosomes do share large repeats (up to 6.3 kb in size), the clear majority of paired-end sequencing reads (>90% in all cases) support the

conformation involving two smaller circles rather than a single combined circle. Although the extremely repetitive nature of the *S. conica* mitochondrial genome precluded complete genome assembly, its structural organization is similar to that of *S. noctiflora*. The vast majority of sequence content assembled into 128 circular-mapping chromosomes ranging from 44 to 163 kb (Table S2). Most of these chromosomes share only short repeats with other parts of the genome. Within both genomes, the relative abundance of the numerous circles appears to be fairly similar with average read depths in single-copy regions spanning 1.7- and 3.1- fold ranges among chromosomes in *S. noctiflora* and *S. conica*, respectively. We did not detect a single intact gene in many chromosomes (20 in *S. noctiflora* and 86 in *S. conica*), raising important questions about the evolutionary forces that maintain their presence and abundance within the mitochondrion. Although these genomic structures are novel for plant mitochondria, various forms of multi-circular organelle genomes have evolved independently in diverse eukaryotic lineages, *(e.g., 19-21)*.

A previous Southern blot analysis showed that the six-copy 1.4 kb repeat in the *S. latifolia* mitochondrial genome is highly recombinationally active, such that the many alternative conformations of the mitochondrial genome in this species occur in roughly equivalent frequencies *(11)*. Paired-end 454 sequence data suggest a comparably high level of repeat-mediated recombinational activity in the *S. vulgaris* mitochondrial genome (Fig. 3A). The relative frequency of recombinant genome conformations increases with repeat size, and all surveyed repeats longer than 100 bp exhibit evidence of recombination. The two largest surveyed repeats pairs (0.9 and 3.0 kb) appear to be at or near a 50:50 equilibrium in the *S. vulgaris* genome (Fig. 3A).

The fast-evolving mitochondrial genomes of *S. noctiflora* and *S. conica* exhibit reduced frequencies of recombinant genome conformations compared to other *Silene* genomes (Fig. 3B) and all other examined angiosperm mitochondrial genomes. Even the largest repeats in the *S. noctiflora* genome (up to 6.3 kb) generate only a small minority of recombinant products. The largest repeats in the *S. conica* genome (up to 87 kb) far exceed our paired-end library span, but analysis of shorter repeats suggests that the genome has experienced a similar shift in the relationship between repeat length and the frequency of recombinant products (Fig. 3B).

In summary, the mitochondrial genomes of several closely related *Silene* species vary by ~100-fold in nucleotide substitution rate and have experienced extraordinarily rapid divergence in genome size, structure, and complexity. The unprecedented mitochondrial genome expansions in *S. noctiflora* and *S. conica* cannot be readily explained in the context of increased mutational "burden" (Supplementary Text) or the directional pressure of small indel bias. Instead, the expanded intergenic regions in these species add to a long-standing mystery regarding the origins of intergenic sequences in plant mitochondrial genomes *(5)*. It is likely that much of this intergenic content is derived from duplicated mitochondrial sequence and promiscuous DNA of nuclear and plastid origin. That only a small fraction of the intergenic sequences in *S. noctiflora* and *S. conica* can be traced to such sources may reflect the rapid rates of sequence divergence in these mitochondrial genomes and the limited availability of sequence data from *Silene* nuclear genomes.

The molecular mechanisms responsible for the increased mutation rates and the expansion of genome size in *S. noctiflora* and *S. conica* mitochondria are unknown, but

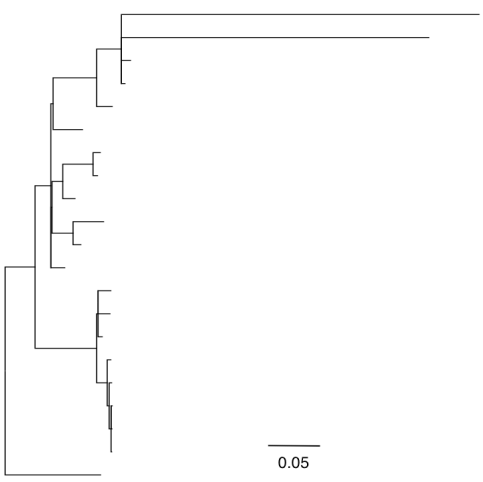
the observed differences in the frequency of recombinant genome conformations between these species and their more slowly evolving congeners suggest that recombination could be a key underlying factor. The *S. noctiflora* and *S. conica* mitochondrial genomes differ in a suite of architectural properties, including rates of point mutations and indels, presence of duplicated and divergent gene copies, frequency of RNA editing *(22)*, genome size, and structural organization (Table 1). Many, perhaps all, of these traits are probably affected by intragenomic recombination and gene conversion. Recombinational processes are key components of DNA repair, and the regulation of these processes has been shown to be important for the maintenance of genome stability in plant mitochondria *(18, 23)*. Our findings highlight the need to characterize *Silene* nuclear content, both as a source for potential donor sequence and to analyze candidate gene families known to be involved in recombination and other aspects of organelle genome maintenance. Unraveling the process of sequence gain and turnover in these rapidly evolving mitochondrial genomes should provide insight into the evolutionary forces underlying the tremendous variation in size, organization, and complexity of eukaryotic genomes.

ACKNOWLEDGEMENTS

We thank the WUSTL Genome Center and UVA Biomolecular Research Facility for DNA sequencing. This research was supported by NSF (MCB-1022128, DEB-0808452, and DEB-0621867), NIH (RO1-GM-70612, 1F32GM080079), IU's METACyt Initiative (Lilly Endowment), and the Jefferson Scholars Foundation.

MAIN TEXT FIGURES AND TABLES

**Fig. 1.** Sequence divergence, genome size, and gene content in seed plant mitochondria. Branch lengths are scaled to the number of synonymous nucleotide substitution per site ($d_S$) based on an analysis of all shared protein genes. Genome size ranges are reported for species with multiple sequences available. Gene counts exclude duplicates and putative pseudogenes.




| Species | Genome Size (kb) | Protein Genes | tRNA Genes |
|---|---|---|---|
| Silene conica | 11,318 | 25 | 2 |
| Silene noctiflora | 6728 | 26 | 3 |
| Silene vulgaris | 427 | 25 | 6 |
| Silene latifolia | 253 | 25 | 9 |
| Beta vulgaris | 365-501 | 30 | 21 |
| Nicotiana tabacum | 431 | 37 | 21 |
| Arabidopsis thaliana | 367 | 31 | 17 |
| Brassica napus | 222 | 32 | 18 |
| Carica papaya | 477 | 39 | 20 |
| Cucurbita pepo | 983 | 38 | 26 |
| Citrullus lanatus | 379 | 38 | 21 |
| Vitis vinifera | 773 | 39 | 23 |
| Oryza sativa | 435-559 | 35 | 19 |
| Triticum aestivum | 453 | 33 | 16 |
| Bambusa oldhamii | 510 | 35 | 18 |
| Sorghum bicolor | 469 | 32 | 19 |
| Tripsacum dactyloides | 704 | 32 | 17 |
| Zea mays | 536-740 | 32 | 18 |
| Zea luxurians | 539 | 32 | 18 |
| Zea perennis | 570 | 32 | 18 |
| Cycas taitungensis | 415 | 41 | 22 |

**Fig. 2.** *Silene* mitochondrial genome sizes relative to all sequenced mitochondrial and eubacterial genomes.

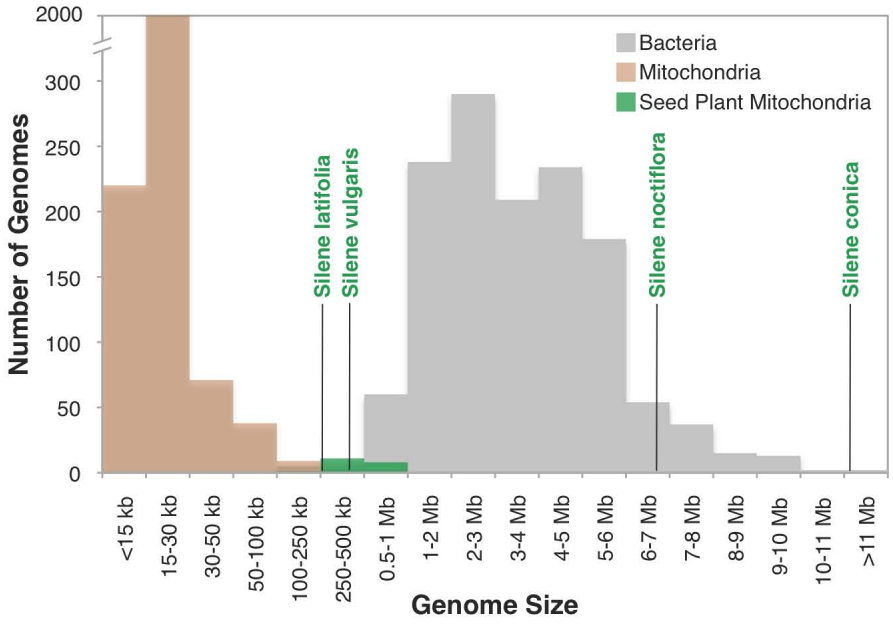


**Fig. 3.** Repeat-mediated recombinational activity in the low mutation rate *S. latifolia* and *S. vulgaris* mitochondrial genomes (A) and the fast-evolving *S. noctiflora* and *S. conica* mitochondrial genomes (B). Each point represents a pair of repeats, and its position on the Y-axis denotes the proportion of recombinant genome conformations detected with paired-end 454 reads. The dashed lines indicate the level at which equal frequencies of read pairs support recombinant and non-recombinant conformations. The *S. latifolia* mitochondrial genome was not sequenced with 454 paired-end reads, but Southern blot hybridizations indicated that alternative genome conformations associated with its six-copy 1.4 kb repeat exist at roughly equivalent frequencies *(11)*, as indicated by the large, black X.

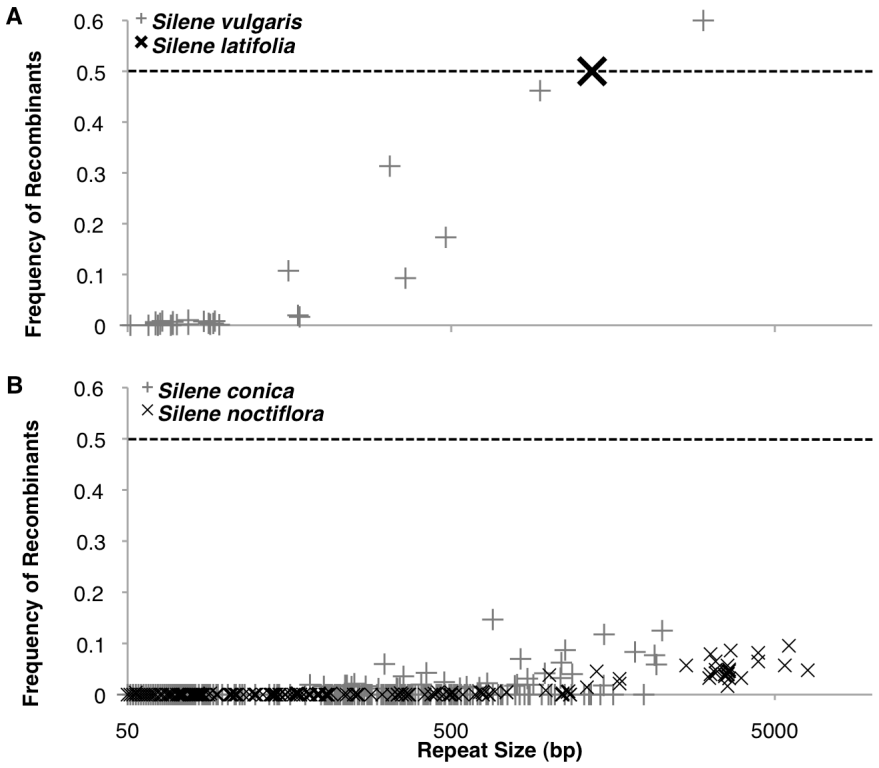


**Table 1.** Summary of four *Silene* mitochondrial genomes.

| | ***Silene latifolia*** | ***Silene vulgaris*** | ***Silene noctiflora*** | ***Silene conica*** |
|---|---|---|---|---|
| **Genome Size in kb** | 253 | 427 | 6728 | 11,318 |
| **Circular Chromosomes** | 1 | 4 | 59 | 129+ |
| **% G+C Content** | 42.6 | 41.8 | 40.8 | 43.1 |
| **Protein Genes*** | 25 | 25 | 26 | 25 |
| **tRNA Genes*** | 9 | 6 | 3 | 2 |
| Native | 6 | 3 | 3 | 2 |
| Plastid-derived | 3 | 3[†] | 0 | 0 |
| **rRNA Genes*** | 3 | 3 | 3 | 3 |
| **Introns*** | 19 | 19 | 18 | 19 |
| *cis*-spliced | 13 | 13 | 12 | 13 |
| *trans*-spliced | 6 | 6 | 6 | 6 |
| **Genic Content in kb (% coverage)** | 51 (20.3) | 48 (11.2) | 72 (1.1) | 77 (0.7) |
| Exonic | 34 (13.6) | 31 (7.2) | 58 (0.9) | 57 (0.5) |
| Intronic[‡] | 17 (6.7) | 17 (4.0) | 14 (0.2) | 20 (0.2) |
| **Intergenic Content in kb (% coverage)** | 202 (79.7) | 379 (88.8) | 6656 (98.9) | 11,241 (99.3) |
| Plastid-derived | 2 (1.0) | 10 (2.3) | 17 (0.3) | 35 (0.3) |
| Conserved with other plant mtDNA[§] | 95 (37.7) | 73 (17.0) | 843 (12.5) | 834 (7.4) |
| Conserved with GenBank nr/nt[§‖] | 5 (2.0) | 3 (0.7) | 20 (0.3) | 16 (0.1) |
| Uncharacterized | 99 (39.0) | 294 (68.9) | 5776 (85.9) | 10,356 (91.5) |
| **Repetitive Content in kb (% coverage)** | 17 (6.7) | 80 (18.8) | 735 (10.9) | 4621 (40.8) |
| Large repeats: >1 kb | 12 (4.9) | 57 (13.3) | 110 (1.6) | 1121 (9.9) |
| Small repeats: ≤1 kb | 5 (1.8) | 23 (5.5) | 625 (9.3) | 3500 (30.9) |
| **RNA Editing Sites** | 287 | 271[¶] | 189 | 182[¶] |
| **Non-Syn. Substitution Rate (x10$^{-9}$/yr)** | 0.08 | 0.35 | 8.90 | 9.98 |
| **Syn. Substitution Rate (x10$^{-9}$/yr)** | 0.70 | 1.60 | 58.17 | 68.22 |
| **$d_N/d_S$** | 0.12 | 0.22 | 0.15 | 0.15 |

*Counts exclude duplicate genes/introns
[†]Two of the *S. vulgaris* plastid-derived tRNA genes may not be functional (Fig. S5).
[‡]Intron lengths only include *cis*-splicing introns.
[§]Excludes regions of plastid-origin.
[‖]Excludes regions conserved in other plant mitochondrial genomes.
[¶]Predicted

SUPPLEMENTARY TEXT

**The Mutational Burden Hypothesis**

The mutational burden hypothesis (MBH) proposes that deleterious mutations are an important force selecting against large and complex genomes *(3, 7, 8, 24)*. This hypothesis has potentially sweeping explanatory power, but some of its tenets are controversial *(25-28)*, and a handful of studies seeking to test it have produced mixed results *(29-35)*. Our study was designed to take advantage of the recent and dramatic changes in mitochondrial mutation rate within the genus *Silene* to test the predicted effects of mutational pressure on the evolution of genome size and architecture.

The co-occurrence of mutational acceleration and genome expansion in the mitochondria of *S. noctiflora* and *S. conica* runs counter to patterns in other eukaryotic mitochondrial genomes (e.g., plants vs. animals) and the predictions of the MBH *(7)*. However, it is important to note that the predicted effects of mutational pressure depend not only on the *intensity* of selection generated by mutation (which is proportional to the mutation rate, $\mu$), but also on the *efficacy* of selection against those mutations (which is proportional to the effective population size, $N_e$). Therefore, an important parameter for the MBH is the product of effective population size and mutation rate ($N_e\mu$).

The levels of intraspecific nucleotide diversity ($\pi$) at neutral sites can be used to estimate $N_e\mu$, although such estimates are restricted to the recent period over which the polymorphism was generated (i.e., the coalescent time), which can be considerably shorter than the entire history of a species or lineage. We compared levels of intraspecific polymorphism in each of the four *Silene* species for a sample of mitochondrial loci as well as a single locus in both the nuclear and plastid genomes (Table S3). The level of

polymorphism within each species does not show any clear relationship with mutation rate or genome size, as the fast-evolving and expanded mitochondrial genomes in *S. conica* and *S. noctiflora* exhibit the highest and lowest levels of polymorphism, respectively. Although mitochondrial loci in *S. conica* do exhibit higher synonymous nucleotide diversity ($\pi_S$) than in more slowly evolving *Silene* species (especially if the highly polymorphic *S. vulgaris atp1* locus *(36-38)* is excluded), they do not show the proportional increases in polymorphism that would be expected based on the accelerated mitochondrial mutation rates in *S. conica* (even after accounting for the approximately two-fold differences in generation times across these *Silene* species *(39)*). Furthermore, the fast-evolving *S. noctiflora* mitochondrial genome is completely devoid of polymorphism at all surveyed loci (see also *(40)*).

The levels of mitochondrial polymorphism in *S. noctiflora* and *S. conica* suggest two alternative (but non-exclusive) interpretations: 1) that either or both of these species have a recent history of lower $N_e$ than their congeners and/or 2) that either or both of these species has experienced a recent reversion to lower mitochondrial mutation rates as observed in other angiosperm lineages with histories of dramatic mitochondrial rate acceleration *(41, 42)*. There is some evidence for the latter interpretation in *S. noctiflora*, which shows little mitochondrial sequence divergence relative to its sister lineage, *S. turkestanica (10)*. The effects of reduced $N_e$ could apply to all three genomic compartments (e.g., as a result of a demographic bottleneck) or be restricted to a subset of loci because of selective sweeps or increased background selection at linked sites. The complete absence of polymorphism across loci from all three genomes in *S. noctiflora* raises the possibility of a recent bottleneck in this species.

The arguments originally outlined by Lynch et al. *(7)* that relate mutation rate to the evolution of organelle genome architecture treat the mutation rate as an independent variable that drives evolution, paying little attention to how the mutation rate itself evolves *(43, 44)*. More recently, Lynch *(45)* has proposed that lineages with low $N_e$ are less capable of selecting against deleterious mutator alleles and, therefore, tend to evolve higher mutation rates. Based on this argument, the MBH could be reformulated to predict a positive correlation between genome size and mutation rate (the very opposite of the pattern described in *(7)*) in cases of extreme variation in $N_e$. Such a prediction would be based two separate theoretical relationships, both driven by differences in $N_e$: 1) the classic negative association between $N_e$ and genome size predicted by the MBH *(3, 24)* and 2) the more recent argument for a negative relationship between $N_e$ and $\mu$ *(45)*.

Although this interpretation of the MBH could potentially be consistent with the associated increases in mitochondrial mutation rate and genome size in *S. noctiflora* and *S. conica*, it still appears, at best, to provide an incomplete explanation of the *Silene* data. In particular, if reduced $N_e$ is the primary force driving the evolution of mitochondrial mutation rates and genome architecture, according to the MBH, we would expect additional changes in genome architecture including a proliferation of intronic sequence and RNA editing sites. However, we found the opposite patterns. Specifically, introns are on average slightly shorter in *S. noctiflora* and *S. conica* (Fig. S6). This result is actually consistent with the predicted effects of increased mutation rates *(7)*. Although the reduction in the frequency of RNA editing in *S. noctiflora* and *S. conica* (Table 1) is also superficially consistent with *a priori* predictions, a detailed analysis of the pattern of

editing site losses failed to support an increase in the intensity of selection against RNA editing in these high-rate lineages *(22)*.

The inability of existing theory to explain the extreme patterns of divergence in *Silene* mitochondrial genomes points to a valuable opportunity to expand our understanding of the evolutionary forces that shape genomic complexity. Although this study was restricted to a small number of species from a single genus, it captured enormous variation in genome architecture (e.g., approximately 98% of the known range of organelle genome sizes), indicating that profound evolutionary mechanisms acting on the genome are still to be identified.

**Single vs. Multiple Origins of Mitochondrial Rate Acceleration and Genome Expansion**

The striking similarities in mitochondrial genome architecture and substitution rate between *S. noctiflora* and *S. conica* raise the question of whether these shared characteristics reflect a single set of evolutionary events or parallel changes in independent lineages. The relationships among the four *Silene* species analyzed in this study have not been well resolved by previous molecular phylogenetic studies, as these lineages appear to have radiated along with the other major groups in *Silene* subgenus *Behenantha* over a narrow window of evolutionary time *(10, 46)*. Therefore, if observed changes in the rates of nucleotide substitution and structural evolution did occur in a common ancestor of *S. noctiflora* and *S. conica*, they must have done so very shortly before the divergence of these two lineages. Because of the extreme mitochondrial substitution rate variation among species, the present study does not provide the

necessary data for further resolving these relationships by traditional molecular phylogeny. Nevertheless, the sequence content of the expanded *S. noctiflora* and *S. conica* mitochondrial genomes may be informative. Notably, these genomes share a large amount of intergenic sequences that are not found in any other seed plant mitochondrial genome, including those of *S. latifolia* and *S. vulgaris*. Such sequences total 659 kb and 760 kb in *S. noctiflora* and *S. conica*, respectively, and show little or no homology with any available sequences in the GenBank nr/nt database. Dispersed repeats cover 11.1% of this shared sequence in *S. noctiflora* and 19.8% in *S. conica*. These values are on par with or less than the levels of overall repeat coverage in the corresponding genomes (Table 1), indicating that the shared sequences are not caused by a proliferation of repetitive elements. Instead, the shared intergenic sequences may be the remnants of an ancestral genomic expansion that preceded the divergence of *S. noctiflora* and *S. conica*, suggesting a possible sister relationship between these two lineages. However, we cannot rule out the possibility that the shared sequences are the result of parallel acquisitions from similar sources, such as the nuclear genomes in each species or the largely uncharacterized population of substoichiometric sequences within the mitochondria. Given the persistent uncertainty regarding the phylogenetic history of the four *Silene* species in this study, these relationships were left unresolved in all substitution rate analyses (see Materials and Methods). Generating sequence data from other genomic compartments, particularly from a large number of unlinked nuclear loci, may provide more insight into the history of these *Silene* species.

**Materials and Methods**

**Study System.** The genus *Silene* (Caryophyllaceae) consists of approximately 700 predominantly herbaceous species of flowering plants *(47)*, many of which are used as models in ecology and evolution *(48)*. The four species used in the present study are all members of the subgenus *Behenantha (10)* and are native to and widely distributed across Europe and, in some cases, parts of Asia *(49)*. They also have been widely introduced outside of their native range *(39)*. *Silene noctiflora* L. and *S. conica* L. both have annual life histories *(50)*, and they are largely selfing hermaphrodites but produce a low frequency of pistillate (female) flowers and can therefore be characterized as gynomonoecious *(51-53;* DBS, pers. obs.*)*. *Silene latifolia* Poir. and *S. vulgaris* (Moench) Garcke are short-lived perennials with an average generation time of approximately two years *(39)* that maintain dioecious and gynodioecious breeding systems, respectively *(50, 51)*.

**Source Material and mtDNA Extraction.** Details of the *Silene latifolia* mitochondrial genome project were described previously *(11)*. For each of the other three species, approximately 200 g of tissue was collected from multiple individuals of a single maternal family. The maternal lineages were derived from seeds originally collected in Abruzzo, Italy (*S. conica*); Eggleston, VA, USA (*S. noctiflora*); or Stuarts Draft, VA, USA (*S. vulgaris*). All aboveground tissue was used for *S. vulgaris*, including leaves, stems, and flowers, while only leaf tissue was collected for *S. noctiflora* and *S. conica*. Mitochondrial DNA was purified from harvested tissue using established protocols based on differential centrifugation, treatment with DNase I, and then either CsCl gradients or phenol:chloroform extraction *(54, 55)*.

**454 and Illumina Sequencing.** For each of the species, 3 kb paired-end libraries were prepared following standard protocols for sequencing on a Roche 454 GS-FLX platform with Titanium reagents. Additional libraries were prepared (also following standard Roche protocols) for the larger *S. noctiflora* and *S. conica* mitochondrial genomes, including shotgun libraries for both species and a 12 kb paired-end library for *S. noctiflora*. The latter was constructed following the standard 8 kb protocol, but the larger 12 kb average fragment size range was selected based on the size distribution of the DNA sample after shearing. Each library was run on a single quarter-plate region except for the *S. conica* shotgun library and the *S. noctiflora* 12 kb paired-end library, which were each run on two quarter-plate regions. The shotgun library for *S. noctiflora* was constructed and sequenced by the Genome Center at Washington University in St. Louis. All other 454 library construction and sequencing was performed at the Genomics Core Facility in the University of Virginia's Department of Biology.

To generate sufficient starting material for Illumina library construction, mtDNA samples were amplified with GenomiPhi V2 (GE Healthcare, Piscataway, NJ). Paired-end sequencing libraries were generated and tagged with multiplex barcodes using the NEBNext DNA Sample Prep Reagent Set 1 (New England Biolabs, Ipswich, MA) in accordance with protocols developed by the University of California Davis Genome Center. In brief, DNA samples were sonicated to a peak fragment size of between 300 and 600 bp. DNA fragments were then end polished and ligated to adaptors carrying a unique 6 bp barcode. The resulting samples were gel-purified and amplified with 14 PCR cycles using paired-end library primers. The three libraries were included in a larger sample pool and sequenced in a single lane of a 2 x 85 bp paired-end run on an Illumina

GAII. Sequencing was performed at the Biomolecular Research Facility in the University of Virginia's School of Medicine.

**Genome Assembly.** Each quarter-plate 454 run produced between 32 and 104 Mb of sequence. The total sequencing yield was 270, 210, and 51 Mb for the *S. noctiflora*, *S. conica*, and *S. vulgaris* mtDNA samples, respectively. However, not all sequence data were used in primary genome assembly. For *S. noctiflora*, only the shotgun and 3 kb paired-end data were analyzed in the initial assembly process. The 12 kb paired-end data were only used to resolve structures associated with large (>3 kb) repeats and to quantify the frequency of alternative genome conformations resulting from recombination among repeat copies (see below). For the smaller, *S. vulgaris* mitochondrial genome, a single quarter-plate run produced very high coverage (>80x). Preliminary analyses suggested use of the entire dataset increased fragmentation in the assembly. Therefore, a random set of sequence reads totaling 25 Mb was selected for initial assembly. The full *S. vulgaris* dataset was used only for subsequent quantification of alternative genome conformations.

For each genome, the 454 sequence reads were assembled with Roche's GS *de novo* Assembler v2.3 ("Newbler") using default settings. The resulting assemblies produced average read depths of 20x, 25x, and 42x for the *S. conica*, *S. noctiflora*, and *S. vulgaris* mitochondrial genomes, respectively. Although the assemblies contained few, if any, gaps or low-coverage regions, they were highly fragmented because of the repetitive and recombinational nature of these genomes (Fig. 3, Figs. S7-S10). The assemblies also contained contigs from contaminating nuclear and plastid DNA. True mitochondrial contigs were distinguished based on read depth and connectivity to other contigs in the assembly. Contigs were sorted manually based on two types of data describing the

connectivity between contigs: 1) paired-end reads that mapped to two different contigs and 2) single reads that were split by the assembler and assigned to the ends of two different contigs. Based on these data, contigs were organized into "subgenomes", each of which represented either a closed circular assembly or a single-copy assembly flanked on either side by recombinationally active repeats. Each of these subgenomic contig groups was then reassembled using a custom set of Perl and BASH scripts that identified all sequencing reads uniquely associated with the corresponding contigs and ran a new assembly based solely on those reads. The resulting subgenomic assemblies were then manually edited and combined as necessary with the aid of Consed v17.0 *(56)*.

The largest repeats in both the *S. conica* and *S. vulgaris* mitochondrial genomes exceeded the 3 kb span size of their respective paired-end libraries. Therefore, the relationships between the single-copy regions flanking these large repeats are ambiguous. These ambiguities were tentatively resolved based on the pattern observed in smaller repeats within each genome (Fig. 3).

Based on the high level of recombinational activity among smaller repeats in *S. vulgaris*, we assumed that large repeats also have high recombinational activity. Therefore, we assembled the majority of the *S. vulgaris* genome content into a single chromosome, analogous to the "master circle" typically reported for plant mitochondrial genomes. As discussed previously *(11)*, the arrangement of repeats and single-copy regions within this circle should be considered only one of many possible alternative representations. We also identified three small circles that were not included in the main assembly. One of these circles (chromosome 4) shows no evidence of recombinational activity with the rest of the genome, while the other two do share repeats that recombine

with the main chromosome. However, in both of these cases, the repeats are small (<500 bp), and the clear majority of reads support the closed circle conformations over a single combined circle. For convenience, we refer to these three circles as chromosomes, but their small size and (in some cases) substantial degree of recombinational activity with the rest of the genome distinguish them from the chromosomal structure that characterizes the *S. noctiflora* and *S. conica* mitochondrial genomes.

In contrast to *S. vulgaris*, the bulk of the *S. noctiflora* and *S. conica* mitochondrial genomes map to discrete circular chromosomes that exhibit little or no recombinational activity with the rest of the genome. In both species, repeats show much less evidence of recombination than repeats of similar size in *S. latifolia* and *S. vulgaris* (Fig. 3). Moreover, in cases of recombinationally active repeats, the clear majority of paired-end reads (>90% in all cases in *S. noctiflora* and the vast majority of cases in *S. conica*; Fig. 3) support minimally sized circular conformations rather than larger combined circles. Therefore, for assembly ambiguities associated with repeats exceeding the 3 kb paired-end library span in *S. conica*, it was assumed that minimally sized circles predominate over larger combined conformations.

It should be noted that the maps generated from the assembly of DNA sequence data do not necessarily reflect the *in vivo* molecular structure of the genome. In particular, linear concatamers and overlapping linear fragments can assemble as circular maps *(57)*. Efforts to directly observe the molecular structure of angiosperm mitochondrial genomes have identified a complex mixture of linear, circular, and branched molecules *(58, 59)*, indicating that the circular maps produced by genome projects may be abstractions or over-simplifications.

**Mapping Illumina Sequence Data.** To correct base calling errors including insertion and deletion errors known to be associated with long single-nucleotide repeats (i.e., homopolymers) in 454 sequence data, we mapped Illumina sequence data onto the completed mitochondrial genome assemblies for each species. After removal of multiplex barcodes and quality trimming, Illumina sequencing yielded average read lengths between 53 and 69 bp with a total of 398, 326, and 168 Mb of sequence data for *S. noctiflora*, *S. conica,* and *S. vulgaris*, respectively. Paired-end read mapping was performed with SOAP v2.20 *(60)* with the following parameters: m 100, x 900, g 3, r 2. A set of custom Perl scripts were used to call SOAP, parse the resulting output, and modify the genome sequence based on well-supported sequence conflicts. These scripts were run recursively until additional iterations did not produce any further improvement to the sequence.

For both *S. vulgaris* and *S. noctiflora*, Illumina mapping provided high-depth (>10x) coverage for essentially the entire genome (>99.9%). This process identified 55 sequence corrections in *S. vulgaris* and 1734 corrections in *S. noctiflora*, the vast majority of which were associated with homopolymer runs. In contrast, because of the larger size and repetitive complexity of the *S. conica* mitochondrial genome, more than 10% of the sequence had coverage levels below 10x. Furthermore, the recursive mapping approach described above failed to converge for numerous regions in the genome, indicating low confidence in many of the sequence corrections indicated by the Illumina data. To avoid incorporating false sequence changes, we did not use the Illumina data to perform genome-wide corrections in *S. conica*. Consequently, the reported genome sequence likely contains some errors associated with homopolymer runs. We did, however, use the

Illumina data to verify basecalls in *S. conica* coding genes and introns, including cases of frameshift mutations.

**Gene Annotation and Characterization of Intergenic Content.** The annotation of protein, rRNA, and tRNA genes was performed using a combination of local BLAST *(61)* and tRNA-scan *(62)* as described previously *(35)*. Annotated genomes sequences were deposited in GenBank (Table S2).

To identify sequence of plastid origin in the *Silene* mitochondrial genomes, each genome was searched against a database of seed plant plastid genomes, using NCBI-BLASTN (v2.2.24+) with the following parameter settings: dust no, gapopen 8, gapextend 6, penalty -4, reward 5, word_size 7. Only hits with a raw score of at least 250 were considered. These hits were subsequently filtered to exclude matches involving mitochondrial protein and rRNA genes known to have ancient plastid homologs (e.g. mitochondrial *atp1* and plastid *atpA (63)*). We also excluded hits with very high AT contents (>72%), because we found these to be almost exclusively false positives resulting from the use of sensitive BLAST parameters.

To identify intergenic sequence conserved in other plant mitochondrial genomes, all intergenic regions (excluding those of plastid origin) were searched against a database of all sequenced seed plant mitochondrial genomes using NCBI-BLASTN (v2.2.24+) and the following search parameters: task blastn, dust no, gapopen 5, gapextend 2, reward 2, penalty -3, word_size 9. All hits with a raw score of at least 70 were considered homologous.

To identify additional conserved sequences (particularly ones of nuclear origin), the remaining intergenic regions (i.e., excluding annotated genes, plastid-derived

sequence, and regions conserved with other plant mitochondrial genomes) were searched against the GenBank nr and nt databases (release date 12/15/2010) using NCBI-BLASTX and BLASTN (v2.2.24+). Default settings were used for BLASTX, whereas the BLASTN search parameters were as follows: dust yes, gapopen 5, gapextend 2, reward 2, penalty -3, word_size 9. All BLASTX hits with a raw score of at least140 and all BLASTN hits with a raw score of 70 or above were considered homologous.

**Characterization of Repetitive Content.** Tandem repeats in each *Silene* mitochondrial genome were identified with Tandem Repeat Finder v4.04 *(64)*, but these represented a negligible fraction of total repeat content in each genome and are not reported separately. Dispersed repeats were identified by searching each genome against itself with NCBI-BLASTN (v2.2.24+) using default parameter settings. All hits with a raw score of at least 30 (which corresponds to a perfect repeat of at least 30 bp under these search parameters) were considered repeats.

**Analysis of Recombinational Activity.** We used paired-end reads from 454 sequencing to quantify the relative abundance of alternative genome conformations associated with repeat-mediated recombination. In the absence of any recombination or alternative genome conformations, 454 read pairs should map to positions in the genome that are consistent with the size span of the sequencing library (ca. 3 or 12 kb in this case). However, the presence of genomic rearrangements will result in read pairs that are inconsistent with the reported genome conformation. Therefore, for each pair of repeated sequences in a genome, we quantified the number of 454 read pairs that are inconsistent with the reported genome assembly but are consistent with either of the predicted products of recombination between the repeats. This number was then compared against

the total number of consistent read pairs in the genome that span one of the two repeat copies to determine the relative abundance of the recombinant products.

To perform this analysis, 454 paired-end reads were mapped on the corresponding genome sequence using Roche's GS Reference Mapper v2.3 software with default parameters. For *S. noctiflora*, only reads from the 12 kb paired-end library were used. The resulting output was filtered to exclude duplicate read pairs with identical start positions for both the left and right sequences, as these were assumed to have been generated by the PCR amplification step in paired-end library, making them non-independent data points. Inspection of the mapping output suggested that the analysis was too stringent in identifying consistent read pairs. Therefore, any "inconsistent" read pairs that mapped in a proper orientation within a distance of 4 to 16 kb for a 12 kb library or 1 to 6 kb for a 3 kb library were reclassified as consistent. These size ranges were determined based on manual inspection of the distribution of mapping spans.

Identified repeats within each genome (see above) were filtered based on multiple criteria prior to inclusion in this analysis. First, only repeats of at least 50 bp in length and at least 95% sequence identity were considered. Additional repeat pairs were excluded because their proximity to each other or to other repeats would have led to ambiguity in the interpretation of paired-end mapping results. Specifically, repeats were excluded if the two copies were separated by less than the maximum library span or if there was a "correlated" pair of larger repeats within the maximum library span of each repeat copy. Finally, for *S. conica* and *S. vulgaris* (for which only 3 kb paired-end libraries were available), repeat pairs were excluded if one of the repeat copies was within 100 bp of the start of any other repeat >500 bp in size. These cases were excluded because the presence

of adjoining repeats would preclude unambiguous mapping of reads to the flanking sequence. Because of the limited physical coverage and short (3 kb) span length in the *S. conica* paired-end data, there are many repeat pairs (particularly large repeats) in this genome that passed the aforementioned criteria, but have an insufficient number of read pairs to precisely measure the relative frequency of alternative genome conformations. Therefore, frequencies are only reported for repeat pairs that have at least five consistent read pairs spanning each copy. Finally, because of the enormous number of small repeats in the *S. conica* mitochondrial genome, only a random sample of 5% of repeat pairs shorter than 200 bp was included.

To validate our methodological approach, we ran a set of control analyses that used the same set of repeats except that we reversed the coordinates for one of the copies. Therefore, these analyses assessed rearrangements associated with the same genomic regions but would only detect alternative genome conformations if recombination occurred between two homologous sequences lined up in opposite orientations. The frequency of alternative genome conformations was at or near zero for every one of these control analyses (Fig. S10). This suggests that baseline level of genome rearrangement and chimeric artifacts is very low in our dataset and that the alternate genome conformations detected by these methods are the genuine result of repeat-mediated recombination.

**Estimates of Nucleotide Substitution Rate.** Previous analyses based on individual genes have identified massive variation in mitochondrial substitution rates among genes and species within the genus *Silene (9, 10, 37, 38)*. To assess these patterns at a genome-wide scale, all protein genes were aligned with MUSCLE v3.7 *(65)* and levels of

synonymous ($d_S$) and non-synonymous ($d_N$) divergence were estimated using PAML v4.4 *(66)* as described previously *(10)*. Analyses were run both on individual genes and on a concatenated dataset of all shared protein genes. Most analyses included six species (*Arabidopsis thaliana*, *Beta vulgaris*, and all four *Silene* species), but a larger dataset of sequenced seed plant mitochondrial genomes was also analyzed. In all cases, the phylogenetic relationships among the four *Silene* species were left unresolved (i.e., as a four-way polytomy). Because substitutions at RNA editing sites can artificially inflate estimates of $d_N$ *(67)*, we excluded all codons that were found to be edited based on genome-wide datasets from four species *(22, 68, 69)*. To estimate absolute rates of nucleotide substitution in these genomes, $d_N$ and $d_S$ values were divided by an approximate divergence time of 6 Myr *(9, 10, 70)*. However, these estimates should be considered only rough approximations because of the uncertainty in divergence time *(10)* and the potential bias associated with recent polymorphisms *(71, 72)*.

**Indel Analysis.** To determine the frequency and size distribution of indels, all protein genes (including introns) from the four *Silene* species and the outgroup *Beta vulgaris* were aligned with MUSCLE v3.7 and adjusted manually. Unalignable regions at the 5' and 3' ends of genes were excluded. The resulting alignments were first analyzed to identify all indels that were unique to a single species and did not overlap with any other indels. These indel events could be unambiguously assigned to an individual species, and the resulting data show that the overwhelming majority of structural changes have occurred in the *S. noctiflora* and *S. conica* lineages (Fig. S2). Given the extremely small number of indels that have occurred in the *S. vulgaris* and *S. latifolia* lineages, the sequences from these species represent a good approximation of the ancestral state for all

four *Silene* species. Therefore, the *S. latifolia* gene sequences were used in pairwise comparisons with both *S. noctiflora* and *S. conica* to infer the polarity of all indel events, including those that overlap other indels in the aforementioned five-species alignment.

**Prediction of RNA Editing Sites**

A genome-wide analysis of C-to-U RNA editing sites by cDNA sequencing has been reported previously for *S. latifolia* and *S. noctiflora (22)*. To estimate the frequency of RNA editing in *S. vulgaris* and *S. conica*, protein gene sequences were analyzed with a predictive algorithm (PREP-mt *(73)*). Control analyses using *Silene* sequences with known editing sites suggested that different stringency settings (C-values) are appropriate for species with different rates of sequence evolution. Specifically, the *S. conica* data were analyzed with C = 0.8 and the *S. vulgaris* data were analyzed with C = 0.7. PREP-mt does not identify synonymous editing sites, so the reported totals were increased by 10% to approximate the contribution of synonymous edits based on observed rates in other *Silene* genomes *(22)*. All intact protein genes were analyzed as well as the following putative pseudogenes: *rps13* (*S. latifolia*), *rps3* (*S. conica*, *S. latifolia*, and *S. noctiflora*), and *ccmFc* (*S. conica*). For genes with duplicates within the genome, only a single gene copy was analyzed.

**Estimating Nucleotide Polymorphism**

To estimate levels of sequence variation within each of the four *Silene* species in this study, we PCR amplified and Sanger sequenced a sample of five mitochondrial loci as well as a single plastid and nuclear locus for multiple, geographically dispersed populations. Sequencing methods, source populations, and polymorphism data for *S. vulgaris* and *S. latifolia* were reported previously *(37, 38)*. Source populations for *S.*

*noctiflora* and *S. conica* are summarized in Table S4. Sequence data from each species were analyzed with DnaSP v5 *(74)* to calculate nucleotide diversity and the number of segregating sites for each locus. For the nuclear *X4*/*XY4* locus, a single haplotype was randomly selected from each individual for calculation of polymorphism data. Haplotypes were inferred from diploid sequence data using the program PHASE v2.1 *(75)*. Novel sequences were deposited in GenBank (accessions JF722621-JF722652 ).

SUPPLEMENTARY FIGURES AND TABLES

**Fig. S1.** Levels of synonymous ($d_S$) and non-synonymous ($d_N$) sequence divergence in terms of substitutions per site for protein genes in *Silene* mitochondrial genomes.

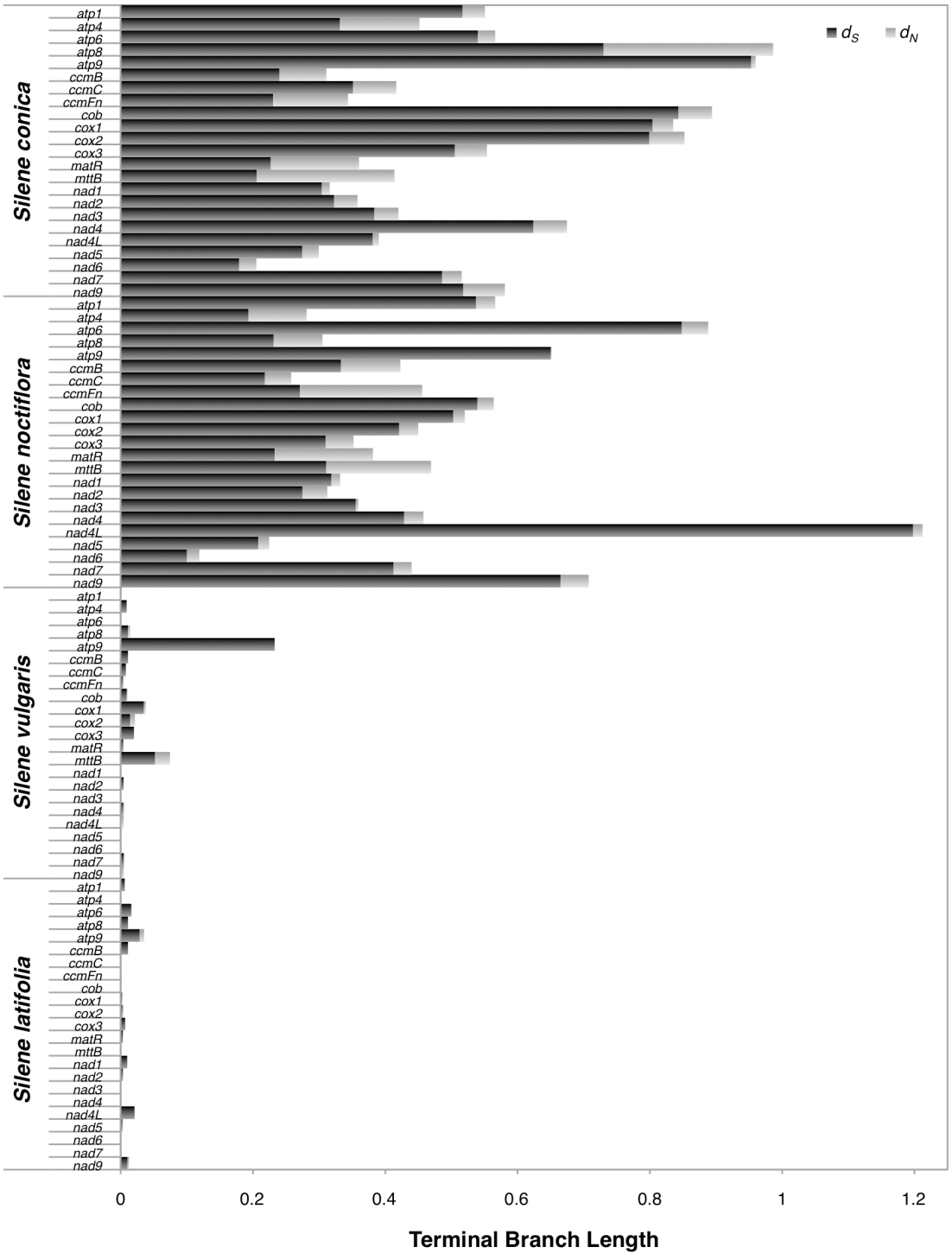


**Fig. S2.** Number of indels in mitochondrial protein genes and introns that are unique to each of the four *Silene* species.

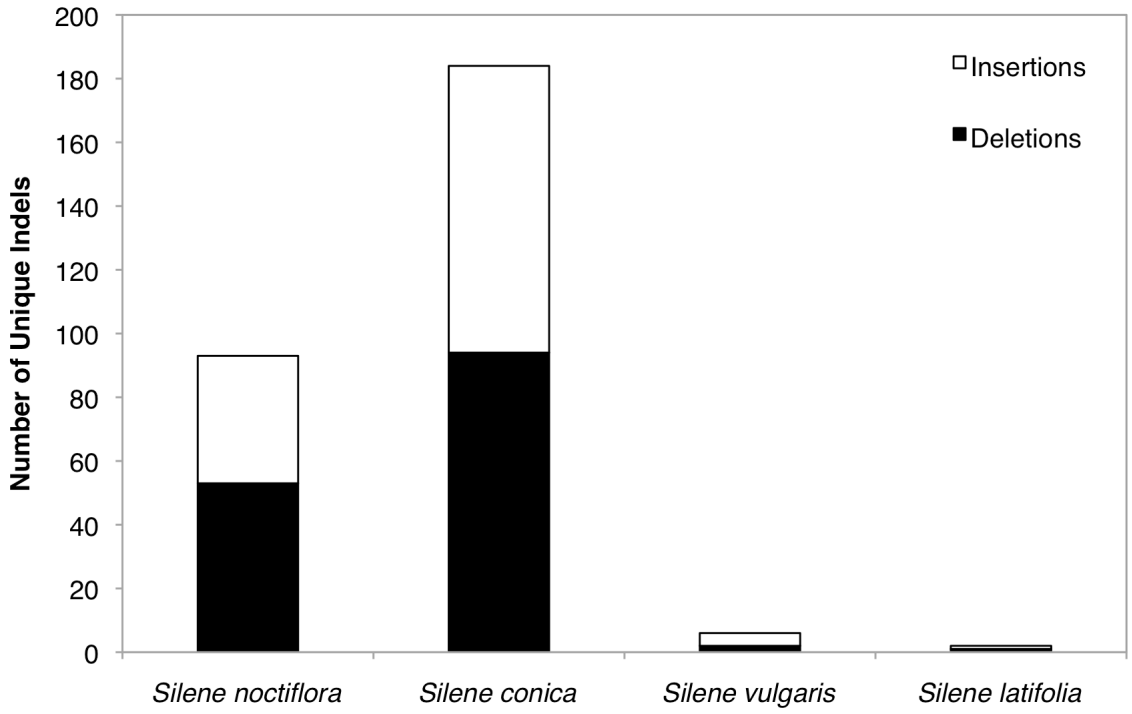
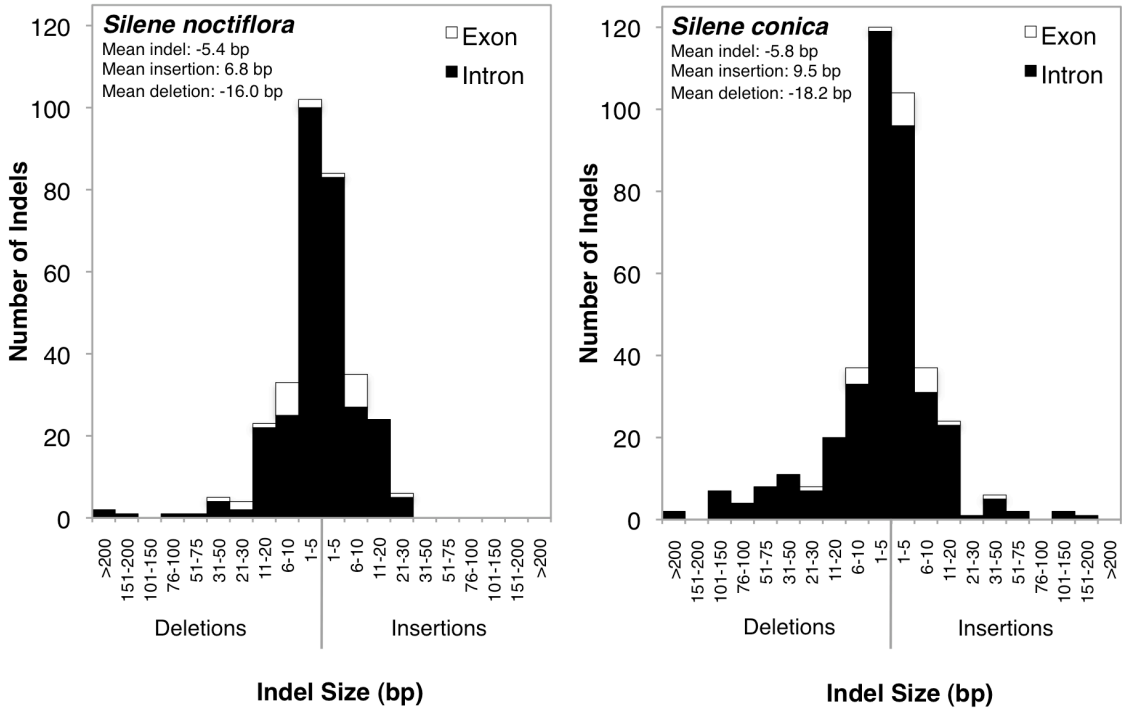


**Fig. S3.** Size distribution of indels in *S. noctiflora* and *S. conica* mitochondrial protein genes inferred using *S. latifolia* genes sequences as an approximation of the ancestral state.



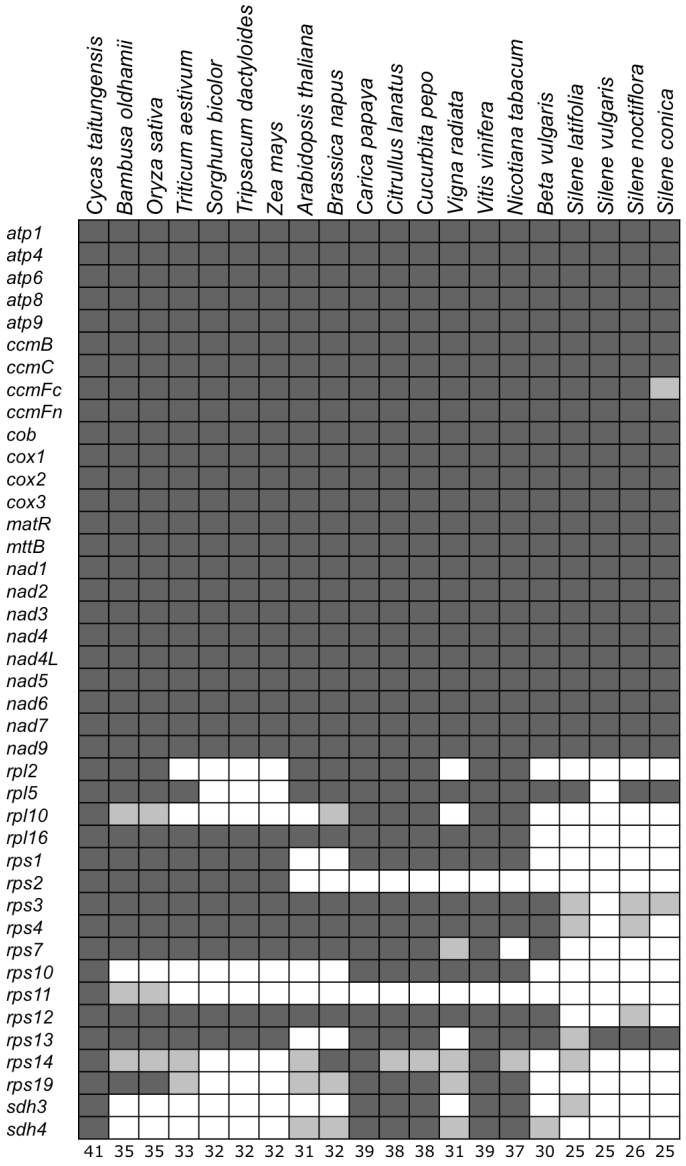



**Fig. S4.** Protein gene content in sequenced seed plant mitochondrial genomes. Dark shading indicates the presence of an intact reading frame, whereas light shading indicates the presence of only a putative pseudogene. The numbers at the bottom of each group indicate the total number of intact genes for that species. Note that the *ccmFc* gene, which is universally present in all other seed plants surveyed to date *(76)*, is classified as a pseudogene in *S. conica*. It has experienced numerous structural mutations in this lineage including multiple frame shifts in the second exon that introduce premature stop codons. However, cDNA sequencing confirms that this gene is transcribed, spliced, and RNA edited in *S. conica*, so it is possible that the gene is still functional in its truncated form.

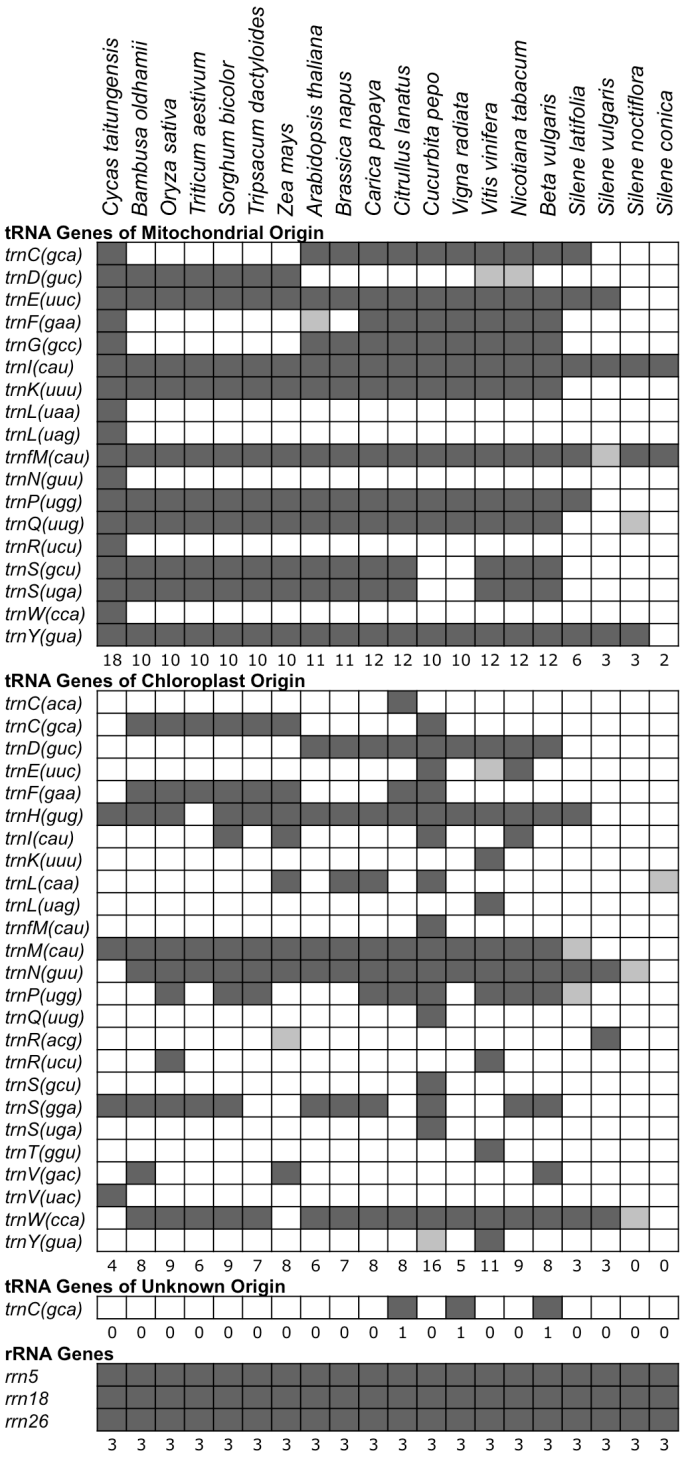



**Fig. S5.** RNA gene content in sequenced seed plant mitochondrial genomes. Dark shading indicates the presence of an intact folding structure, whereas light shading indicates the presence of only a putative pseudogene. The values at the bottom of each group indicate the total number of intact genes for that species. In some cases, the presence of an intact gene may not actually indicate functionality. This is particularly true for tRNA genes embedded within recently transferred regions of plastid DNA *(35, 77)*. For example, the *trnN(guu)* and *trnR(acg)* genes in *S. vulgaris* may not be functional, as they are within a 2.6 kb region that appears to have been recently transferred from the plastid genome (based on its perfect sequence identity with the exception of a single 18 bp deletion). They are not orthologous to the plastid-derived

copies of *trnN(guu)* and *trnR(acg)* in other seed plant mitochondria. In *Cycas*, the *trnL(uaa)*, *trnQ(uug)*, and *trnR(ucu)* genes are classified based on sequence homology to other land plant tRNAs even though their genomically encoded anticodons differ (CAA, CUG, and CCU, respectively). It is possible that these three anticodons undergo C-to-U RNA editing to restore the ancestral codon as observed in other vascular plants *(78, 79)*.

**Fig. S6.** Lengths of *cis*-spliced introns in *Silene* mitochondrial genomes.

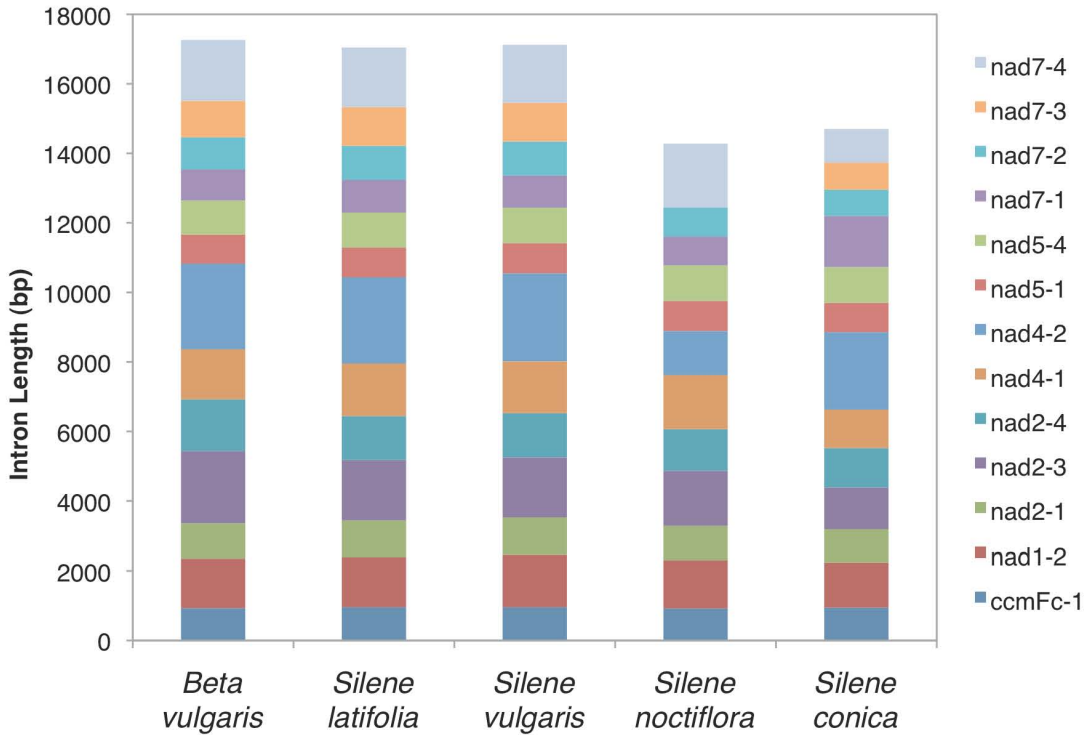


**Fig. S7.** Size distribution of repetitive content by the number of repeat pairs (left column) and total repeat length (right column). Both datasets are based on all repeat pairs identified with BLAST by searching each genome against itself. Note that this method is different than counting individual repeat copies, which cannot be unambiguously identified when repeats exist in numerous partially overlapping copies, as they do in these genomes. For example, a repeat with four *copies* would be associated with six unique repeat *pairs*. Because of the enormous number of multi-copy, overlapping repeats in *S. conica*, the total length of repeat pairs exceeds the size of the genome even though more than half of it is single copy. For these same reasons, the distribution of repeat lengths in this figure differs from the repeat coverage statistics reported in Table 1, which consider what fraction of the genome is covered by repeats but not the total number of repeat pairs.

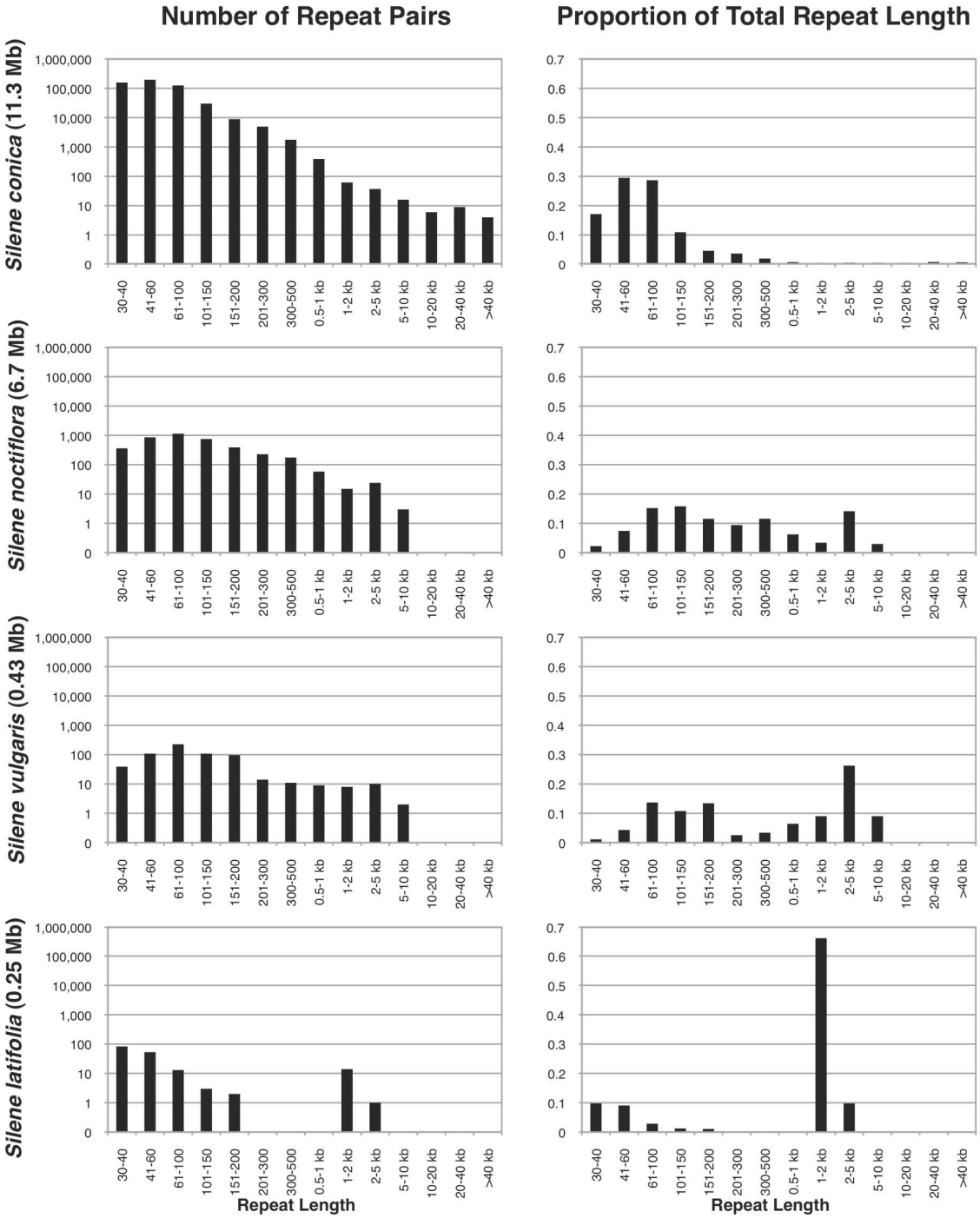

Number of Repeat Pairs
Proportion of Total Repeat Length
Silene conica (11.3 Mb)
Silene noctiflora (6.7 Mb)
Silene vulgaris (0.43 Mb)
Silene latifolia (0.25 Mb)
1,000,000
100,000
10,000
1,000
100
10
1
0
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
30-40
41-60
61-100
101-150
151-200
201-300
300-500
0.5-1 kb
1-2 kb
2-5 kb
5-10 kb
10-20 kb
20-40 kb
>40 kb
Repeat Length

**Fig. S8.** Repeat coverage depth in *Silene* mitochondrial genomes. For each curve, the y-intercept indicates the proportion of the mitochondrial genome that is single-copy in that species. Other points along the curve indicate the cumulative genomic coverage up to a certain repeat depth. For example, the height of the curve at a value of 10 on the x-axis indicates the fraction of the genome represented by all nucleotide positions that match nine or fewer repeats elsewhere in the genome.

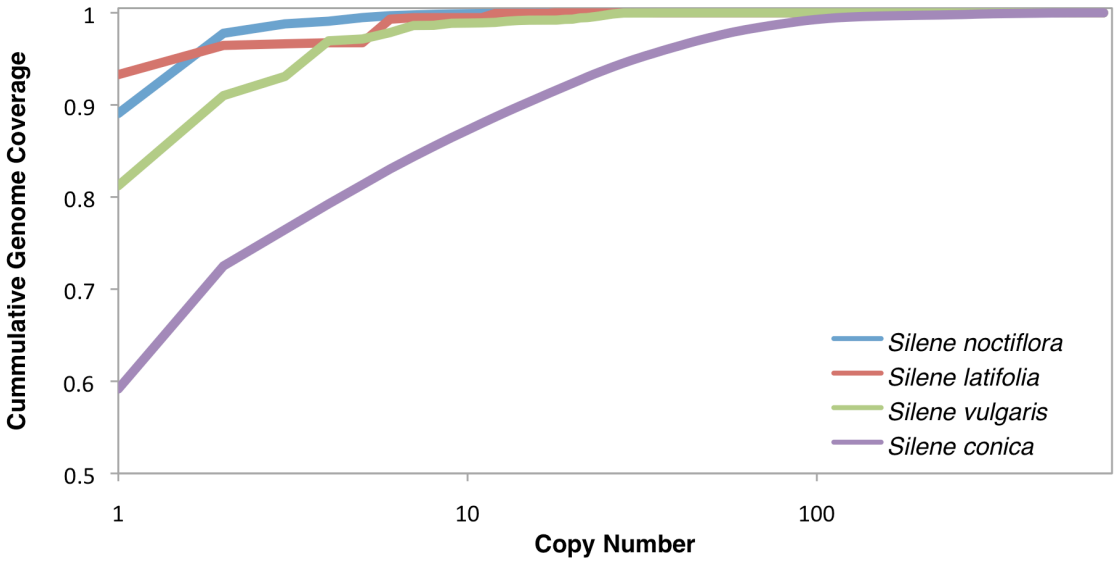


**Fig. S9.** Distribution of percent sequence identity between pairs of repeats detected by BLAST. Only repeat pairs greater than 300 bp in length were used to calculate these distributions. Whereas all repeat pairs of this length are completely identical or nearly so in *S. latifolia* and *S. vulgaris*, the fast-evolving *S. noctiflora* and *S. conica* mitochondrial genomes contain a large proportion of divergent sequence pairs. This pattern is consistent with a reduction in recombinational activity (including gene conversion) in the *S. noctiflora* and *S. conica* genomes allowing divergence between duplicated sequences.

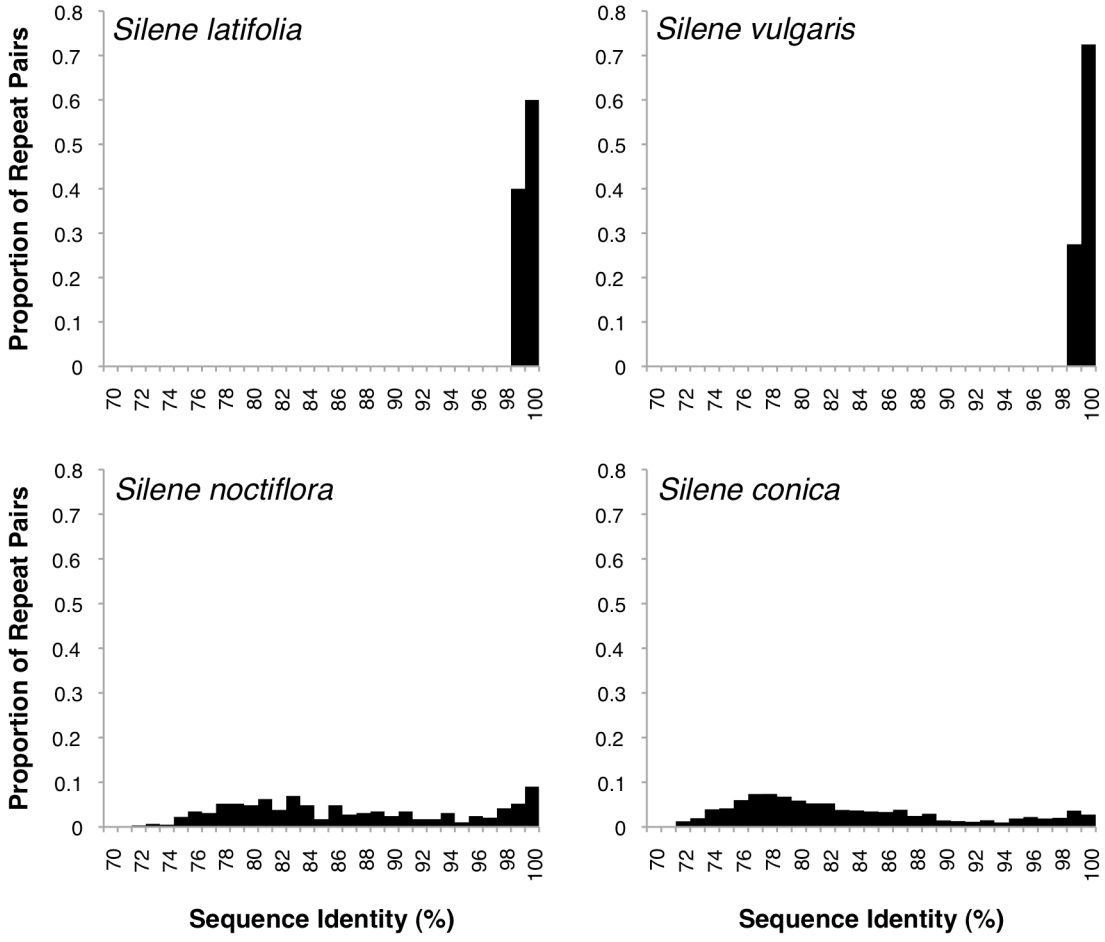


**Fig. S10.** Assays of repeat-mediated recombinational activity in *Silene* mitochondrial genomes. (A) The left column shows the data presented in Fig. 3 individually for each species (note the change in scale for each species). (B) The right column reports the analysis of the exact same repeats except in reversed orientation as a measure of the baseline level of alternative genome conformations and/or library construction artifacts in each species. Note that not all repeat pairs are shown (see Materials and Methods for filtering criteria).

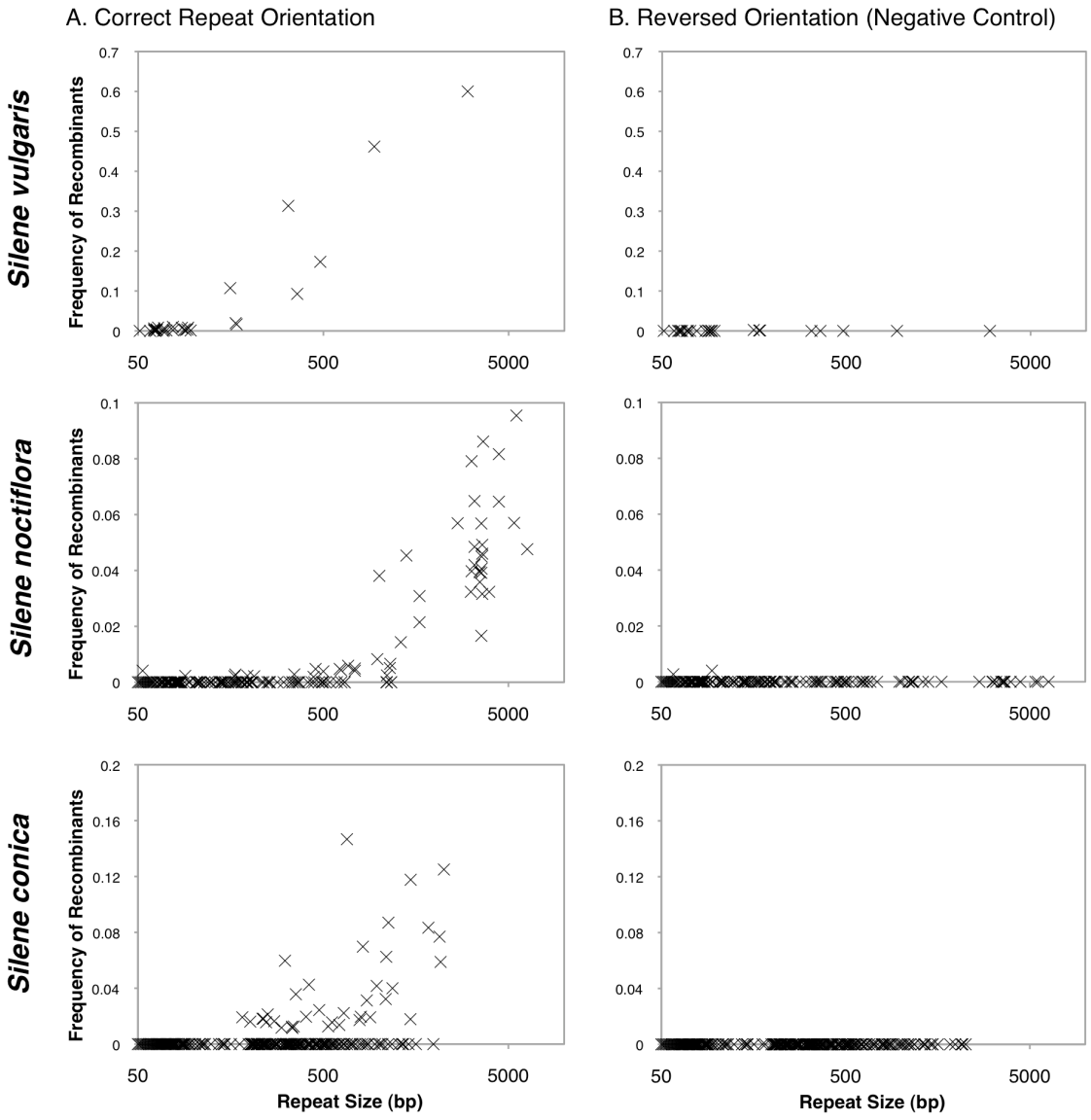


**Table S1.** Duplicate genes in *Silene* mitochondrial genomes. Values indicate cases where more than one full-length gene or exon copy exists within the corresponding genome. Bold values indicate that the co-existing copies differ in sequence. For cases in which a mixture of identical and divergent copies exist, the total number of copies is shown in plain text and the number of unique sequences is shown parenthetically in bold. Numerous cases of small duplicated gene fragments are not reported.

| Gene | *S. latifolia* | *S. vulgaris* | *S. noctiflora* | *S. conica* |
|---|---|---|---|---|
| *atp4* | . | . | 2 | . |
| *atp6* | . | . | . | **2** |
| *atp8* | . | . | . | 3 |
| *ccmB* | . | . | 3 | 2 |
| *ccmFc* | . | . | . | 2 |
| *cox3* | . | . | . | 3 |
| *mttB* | . | . | **2** | . |
| *nad1-exons2-3* | . | . | . | 2 |
| *nad1-exon5* | . | . | . | **2** |
| *nad2-exon1* | . | . | **4** | **.** |
| *nad2-exon2* | . | . | **2** | **.** |
| *nad3* | . | . | 3 | 4 |
| *nad4* | . | . | . | 2 |
| *nad4L* | . | . | 2 | . |
| *rpl5* | . | . | **2** | . |
| *rps13* | . | . | **2** | 3(**2**) |
| *rps3* | . | . | . | 2 |
| *rrn26* | . | . | 5 | 3 |
| *rrn18* | . | . | 5 | 4 |
| *rrn5* | **2** | . | 5 | **5** |
| *trnfM* | . | . | 5(**3**) | **3** |
| *trnI* | . | . | 5(**3**) | 3(**2**) |

**Table S2.** Summary of length, GC content, and gene content of circular chromosomes and (partially assembled genomic fragments in *S. conica*).

| Species | Chromosome No. | Length (bp) | GC Content | Intact Genes | GenBank |
|---|---|---|---|---|---|
| *Silene latifolia* | 1 | 253,413 | 42.56% | 37 | HM562727 |
| *Silene vulgaris* | 1 | 394,403 | 41.63% | 31 | JF750427 |
| *Silene vulgaris* | 2 | 14,341 | 41.06% | 1 | JF750429 |
| *Silene vulgaris* | 3 | 12,697 | 45.86% | 4 | JF750430 |
| *Silene vulgaris* | 4 | 5,697 | 46.31% | 0 | JF750428 |
| *Silene noctiflora* | 1 | 191,963 | 40.70% | 3 | JF750481 |
| *Silene noctiflora* | 2 | 161,299 | 40.66% | 1 | JF750484 |
| *Silene noctiflora* | 3 | 153,996 | 41.17% | 1 | JF750436 |
| *Silene noctiflora* | 4 | 152,707 | 41.25% | 1 | JF750485 |
| *Silene noctiflora* | 5 | 148,612 | 41.22% | 3 | JF750460 |
| *Silene noctiflora* | 6 | 147,958 | 40.80% | 2 | JF750443 |
| *Silene noctiflora* | 7 | 146,587 | 41.03% | 2 | JF750469 |
| *Silene noctiflora* | 8 | 145,239 | 41.10% | 1 | JF750478 |
| *Silene noctiflora* | 9 | 142,006 | 40.70% | 2 | JF750470 |
| *Silene noctiflora* | 10 | 140,341 | 40.78% | 1 | JF750437 |
| *Silene noctiflora* | 11 | 138,824 | 41.08% | 1 | JF750465 |
| *Silene noctiflora* | 12 | 134,806 | 41.14% | 1 | JF750433 |
| *Silene noctiflora* | 13 | 130,152 | 40.81% | 0 | JF750451 |
| *Silene noctiflora* | 14 | 130,008 | 40.25% | 0 | JF750431 |
| *Silene noctiflora* | 15 | 129,012 | 40.81% | 0 | JF750472 |
| *Silene noctiflora* | 16 | 128,843 | 40.33% | 1 | JF750453 |
| *Silene noctiflora* | 17 | 127,930 | 40.03% | 0 | JF750434 |
| *Silene noctiflora* | 18 | 127,245 | 40.40% | 0 | JF750445 |
| *Silene noctiflora* | 19 | 127,193 | 42.28% | 1 | JF750486 |
| *Silene noctiflora* | 20 | 126,609 | 40.64% | 1 | JF750456 |
| *Silene noctiflora* | 21 | 126,452 | 40.58% | 0 | JF750477 |
| *Silene noctiflora* | 22 | 122,890 | 40.35% | 1 | JF750440 |
| *Silene noctiflora* | 23 | 122,485 | 40.60% | 1 | JF750452 |
| *Silene noctiflora* | 24 | 119,408 | 40.69% | 0 | JF750448 |
| *Silene noctiflora* | 25 | 115,341 | 40.98% | 1 | JF750447 |
| *Silene noctiflora* | 26 | 114,914 | 40.81% | 0 | JF750471 |
| *Silene noctiflora* | 27 | 113,308 | 42.15% | 2 | JF750479 |
| *Silene noctiflora* | 28 | 108,290 | 41.02% | 2 | JF750462 |
| *Silene noctiflora* | 29 | 108,152 | 41.25% | 1 | JF750450 |
| *Silene noctiflora* | 30 | 108,040 | 40.58% | 0 | JF750442 |
| *Silene noctiflora* | 31 | 107,738 | 40.08% | 0 | JF750449 |
| *Silene noctiflora* | 32 | 106,477 | 40.57% | 1 | JF750466 |
| *Silene noctiflora* | 33 | 104,288 | 40.96% | 1 | JF750454 |
| *Silene noctiflora* | 34 | 103,926 | 40.70% | 1 | JF750489 |
| *Silene noctiflora* | 35 | 103,557 | 40.05% | 0 | JF750432 |
| *Silene noctiflora* | 36 | 103,548 | 40.22% | 1 | JF750435 |
| *Silene noctiflora* | 37 | 103,320 | 40.70% | 0 | JF750458 |

| | | | | | |
|---|---|---|---|---|---|
| *Silene noctiflora* | 38 | 102,347 | 40.14% | 2 | JF750444 |
| *Silene noctiflora* | 39 | 100,876 | 40.46% | 1 | JF750467 |
| *Silene noctiflora* | 40 | 100,579 | 41.45% | 3 | JF750459 |
| *Silene noctiflora* | 41 | 100,078 | 40.47% | 0 | JF750438 |
| *Silene noctiflora* | 42 | 98,550 | 40.59% | 1 | JF750464 |
| *Silene noctiflora* | 43 | 97,627 | 40.50% | 0 | JF750439 |
| *Silene noctiflora* | 44 | 96,564 | 41.12% | 1 | JF750475 |
| *Silene noctiflora* | 45 | 96,233 | 40.71% | 0 | JF750480 |
| *Silene noctiflora* | 46 | 95,084 | 40.92% | 0 | JF750441 |
| *Silene noctiflora* | 47 | 94,621 | 41.15% | 2 | JF750446 |
| *Silene noctiflora* | 48 | 94,201 | 39.91% | 0 | JF750487 |
| *Silene noctiflora* | 49 | 92,946 | 41.03% | 1 | JF750457 |
| *Silene noctiflora* | 50 | 92,480 | 40.75% | 3 | JF750463 |
| *Silene noctiflora* | 51 | 92,366 | 40.92% | 0 | JF750476 |
| *Silene noctiflora* | 52 | 91,804 | 41.15% | 1 | JF750455 |
| *Silene noctiflora* | 53 | 91,595 | 40.79% | 2 | JF750474 |
| *Silene noctiflora* | 54 | 89,951 | 40.93% | 0 | JF750488 |
| *Silene noctiflora* | 55 | 86,782 | 39.77% | 1 | JF750473 |
| *Silene noctiflora* | 56 | 81,416 | 40.97% | 3 | JF750483 |
| *Silene noctiflora* | 57 | 74,922 | 40.57% | 2 | JF750461 |
| *Silene noctiflora* | 58 | 67,018 | 40.71% | 0 | JF750468 |
| *Silene noctiflora* | 59 | 66,365 | 40.66% | 3 | JF750482 |
| *Silene conica* | 1 | 163,071 | 43.08% | 0 | JF750534 |
| *Silene conica* | 2 | 155,921 | 42.47% | 1 | JF750515 |
| *Silene conica* | 3 | 151,014 | 43.56% | 1 | JF750506 |
| *Silene conica* | 4 | 149,089 | 43.15% | 1 | JF750505 |
| *Silene conica* | 5 | 147,648 | 42.60% | 0 | JF750497 |
| *Silene conica* | 6 | 142,373 | 43.31% | 0 | JF750508 |
| *Silene conica* | 7 | 137,495 | 43.66% | 1 | JF750520 |
| *Silene conica* | 8 | 130,956 | 43.51% | 0 | JF750557 |
| *Silene conica* | 9 | 127,523 | 43.15% | 1 | JF750503 |
| *Silene conica* | 10 | 125,751 | 42.93% | 0 | JF750563 |
| *Silene conica* | 11 | 125,117 | 43.06% | 1 | JF750512 |
| *Silene conica* | 12 | 120,801 | 43.56% | 3 | JF750511 |
| *Silene conica* | 13 | 119,487 | 42.72% | 0 | JF750517 |
| *Silene conica* | 14 | 119,001 | 43.51% | 3 | JF750513 |
| *Silene conica* | 15 | 118,929 | 42.89% | 0 | JF750558 |
| *Silene conica* | 16 | 117,607 | 43.37% | 1 | JF750516 |
| *Silene conica* | 17 | 115,763 | 42.93% | 0 | JF750593 |
| *Silene conica* | 18 | 114,589 | 43.12% | 1 | JF750518 |
| *Silene conica* | 19 | 110,652 | 43.59% | 3 | JF750578 |
| *Silene conica* | 20 | 110,427 | 42.60% | 1 | JF750519 |
| *Silene conica* | 21 | 108,941 | 42.98% | 0 | JF750555 |
| *Silene conica* | 22 | 108,940 | 42.66% | 0 | JF750536 |
| *Silene conica* | 23 | 107,216 | 43.61% | 3 | JF750510 |
| *Silene conica* | 24 | 106,709 | 42.67% | 0 | JF750526 |
| *Silene conica* | 25 | 105,730 | 42.95% | 0 | JF750532 |
| *Silene conica* | 26 | 105,655 | 43.24% | 2 | JF750531 |

| | | | | | |
|---|---|---|---|---|---|
| *Silene conica* | 27 | 105,520 | 43.78% | 0 | JF750504 |
| *Silene conica* | 28 | 103,419 | 43.08% | 1 | JF750501 |
| *Silene conica* | 29 | 103,231 | 43.42% | 1 | JF750599 |
| *Silene conica* | 30 | 103,102 | 42.75% | 0 | JF750527 |
| *Silene conica* | 31 | 103,003 | 42.35% | 0 | JF750494 |
| *Silene conica* | 32 | 102,316 | 43.32% | 2 | JF750587 |
| *Silene conica* | 33 | 102,020 | 43.69% | 0 | JF750588 |
| *Silene conica* | 34 | 101,608 | 43.27% | 0 | JF750535 |
| *Silene conica* | 35 | 101,353 | 42.93% | 1 | JF750547 |
| *Silene conica* | 36 | 100,148 | 43.41% | 1 | JF750528 |
| *Silene conica* | 37 | 99,951 | 43.88% | 0 | JF750493 |
| *Silene conica* | 38 | 99,512 | 43.73% | 0 | JF750533 |
| *Silene conica* | 39 | 98,308 | 42.73% | 2 | JF750529 |
| *Silene conica* | 40 | 98,068 | 42.61% | 0 | JF750545 |
| *Silene conica* | 41 | 97,429 | 42.93% | 2 | JF750546 |
| *Silene conica* | 42 | 97,283 | 42.49% | 0 | JF750548 |
| *Silene conica* | 43 | 97,269 | 43.10% | 0 | JF750583 |
| *Silene conica* | 44 | 97,111 | 43.83% | 0 | JF750522 |
| *Silene conica* | 45 | 96,762 | 42.38% | 0 | JF750596 |
| *Silene conica* | 46[a] | 96,395 | 43.30% | 0 | JF750626 |
| *Silene conica* | 47 | 93,489 | 43.75% | 1 | JF750614 |
| *Silene conica* | 48 | 93,235 | 42.79% | 0 | JF750530 |
| *Silene conica* | 49 | 92,438 | 43.65% | 2 | JF750624 |
| *Silene conica* | 50 | 91,619 | 43.22% | 0 | JF750538 |
| *Silene conica* | 51 | 91,401 | 43.21% | 1 | JF750597 |
| *Silene conica* | 52 | 88,472 | 43.12% | 1 | JF750564 |
| *Silene conica* | 53 | 88,260 | 42.84% | 1 | JF750543 |
| *Silene conica* | 54 | 87,616 | 43.50% | 0 | JF750542 |
| *Silene conica* | 55 | 86,501 | 43.40% | 0 | JF750521 |
| *Silene conica* | 56 | 85,606 | 42.75% | 1 | JF750523 |
| *Silene conica* | 57 | 85,117 | 43.44% | 1 | JF750514 |
| *Silene conica* | 58[a] | 85,095 | 43.21% | 1 | JF750627 |
| *Silene conica* | 59 | 84,942 | 42.64% | 1 | JF750495 |
| *Silene conica* | 60 | 84,201 | 43.09% | 1 | JF750539 |
| *Silene conica* | 61 | 83,288 | 42.77% | 2 | JF750610 |
| *Silene conica* | 62 | 82,551 | 43.30% | 0 | JF750585 |
| *Silene conica* | 63 | 81,689 | 43.12% | 1 | JF750544 |
| *Silene conica* | 64 | 81,671 | 43.07% | 1 | JF750507 |
| *Silene conica* | 65 | 81,294 | 42.56% | 0 | JF750569 |
| *Silene conica* | 66 | 81,243 | 43.19% | 0 | JF750540 |
| *Silene conica* | 67 | 80,675 | 42.57% | 0 | JF750541 |
| *Silene conica* | 68 | 79,841 | 43.68% | 0 | JF750574 |
| *Silene conica* | 69 | 79,368 | 43.98% | 0 | JF750572 |
| *Silene conica* | 70 | 79,279 | 43.09% | 0 | JF750525 |
| *Silene conica* | 71 | 78,978 | 42.85% | 0 | JF750576 |
| *Silene conica* | 72 | 78,953 | 43.18% | 1 | JF750591 |
| *Silene conica* | 73 | 78,367 | 43.70% | 0 | JF750549 |
| *Silene conica* | 74 | 77,721 | 42.70% | 0 | JF750561 |

| | | | | | |
|---|---|---|---|---|---|
| *Silene conica* | 75[a] | 77,256 | 42.85% | 0 | JF750628 |
| *Silene conica* | 76 | 77,149 | 43.58% | 0 | JF750589 |
| *Silene conica* | 77 | 76,904 | 42.63% | 0 | JF750550 |
| *Silene conica* | 78 | 76,619 | 43.14% | 0 | JF750556 |
| *Silene conica* | 79 | 76,441 | 43.50% | 0 | JF750559 |
| *Silene conica* | 80 | 76,039 | 42.82% | 0 | JF750551 |
| *Silene conica* | 81 | 75,310 | 44.06% | 0 | JF750605 |
| *Silene conica* | 82 | 74,810 | 43.34% | 0 | JF750552 |
| *Silene conica* | 83 | 74,506 | 43.02% | 0 | JF750562 |
| *Silene conica* | 84 | 73,936 | 43.40% | 0 | JF750553 |
| *Silene conica* | 85 | 73,670 | 43.66% | 0 | JF750566 |
| *Silene conica* | 86 | 73,557 | 42.77% | 1 | JF750554 |
| *Silene conica* | 87 | 73,226 | 42.98% | 0 | JF750490 |
| *Silene conica* | 88 | 71,828 | 42.74% | 0 | JF750567 |
| *Silene conica* | 89 | 70,246 | 42.63% | 0 | JF750601 |
| *Silene conica* | 90 | 69,453 | 43.58% | 0 | JF750598 |
| *Silene conica* | 91 | 69,443 | 43.85% | 2 | JF750602 |
| *Silene conica* | 92 | 68,046 | 43.32% | 0 | JF750586 |
| *Silene conica* | 93 | 67,617 | 42.47% | 0 | JF750491 |
| *Silene conica* | 94 | 67,525 | 43.40% | 2 | JF750524 |
| *Silene conica* | 95 | 67,127 | 43.57% | 0 | JF750560 |
| *Silene conica* | 96 | 66,830 | 43.66% | 3 | JF750612 |
| *Silene conica* | 97 | 66,402 | 43.30% | 0 | JF750608 |
| *Silene conica* | 98 | 65,772 | 43.15% | 0 | JF750565 |
| *Silene conica* | 99 | 65,543 | 42.98% | 1 | JF750590 |
| *Silene conica* | 100 | 65,424 | 42.94% | 0 | JF750604 |
| *Silene conica* | 101 | 65,272 | 43.14% | 2 | JF750606 |
| *Silene conica* | 102 | 65,085 | 42.82% | 0 | JF750499 |
| *Silene conica* | 103 | 64,608 | 43.08% | 0 | JF750571 |
| *Silene conica* | 104 | 64,281 | 42.56% | 0 | JF750575 |
| *Silene conica* | 105 | 63,852 | 42.89% | 0 | JF750603 |
| *Silene conica* | 106 | 63,308 | 43.03% | 0 | JF750502 |
| *Silene conica* | 107 | 63,002 | 42.39% | 0 | JF750496 |
| *Silene conica* | 108 | 62,749 | 43.19% | 2 | JF750595 |
| *Silene conica* | 109 | 62,571 | 42.59% | 0 | JF750568 |
| *Silene conica* | 110 | 62,405 | 42.50% | 0 | JF750492 |
| *Silene conica* | 111 | 60,812 | 42.58% | 0 | JF750498 |
| *Silene conica* | 112 | 60,476 | 43.99% | 0 | JF750537 |
| *Silene conica* | 113 | 60,079 | 42.92% | 0 | JF750580 |
| *Silene conica* | 114 | 59,915 | 43.06% | 0 | JF750570 |
| *Silene conica* | 115 | 59,463 | 42.63% | 0 | JF750509 |
| *Silene conica* | 116 | 58,995 | 43.34% | 0 | JF750615 |
| *Silene conica* | 117 | 58,600 | 43.17% | 0 | JF750592 |
| *Silene conica* | 118 | 58,548 | 43.27% | 0 | JF750609 |
| *Silene conica* | 119 | 57,218 | 43.19% | 0 | JF750579 |
| *Silene conica* | 120 | 55,005 | 43.15% | 0 | JF750594 |
| *Silene conica* | 121 | 54,881 | 43.55% | 0 | JF750611 |
| *Silene conica* | 122 | 54,719 | 43.14% | 0 | JF750581 |

| | | | | | |
|---|---|---|---|---|---|
| *Silene conica* | 123 | 54,696 | 43.46% | 2 | JF750582 |
| *Silene conica* | 124 | 54,650 | 42.92% | 0 | JF750577 |
| *Silene conica* | 125 | 52,701 | 43.29% | 0 | JF750573 |
| *Silene conica* | 126 | 52,589 | 42.33% | 0 | JF750607 |
| *Silene conica* | 127 | 46,047 | 42.39% | 0 | JF750584 |
| *Silene conica* | 128 | 43,958 | 42.46% | 0 | JF750600 |
| *Silene conica* | Fragment_01[b] | 52,090 | 43.02% | 0 | JF750629 |
| *Silene conica* | Fragment_02[b] | 51,272 | 42.85% | 0 | JF750613 |
| *Silene conica* | Fragment_03[b] | 44,768 | 43.08% | 0 | JF750500 |
| *Silene conica* | Fragment_04[b] | 40,160 | 42.70% | 0 | JF750616 |
| *Silene conica* | Fragment_05[b] | 4,346 | 44.91% | 0 | JF750620 |
| *Silene conica* | Fragment_06[b] | 4,039 | 46.50% | 0 | JF750621 |
| *Silene conica* | Fragment_07[c] | 3,765 | 48.76% | 0 | JF750625 |
| *Silene conica* | Fragment_08[b] | 1,959 | 46.66% | 0 | JF750623 |
| *Silene conica* | Fragment_09[b] | 576 | 54.17% | 0 | JF750622 |
| *Silene conica* | Fragment_10[b] | 503 | 39.96% | 0 | JF750619 |
| *Silene conica* | Fragment_11[b] | 353 | 44.19% | 0 | JF750618 |
| *Silene conica* | Fragment_12[b] | 288 | 54.17% | 0 | JF750617 |

[a]Circle broken into two pieces because of assembly gaps.

[b]Unassembled fragment with alternative connections to other parts of the genome. Not circular mapping.

[c]Uncharacterized high-copy element.

**Table S3.** Nucleotide polymorphism within *Silene* species

| | | *Silene latifolia* (n = 28) | | | *Silene vulgaris* (n = 40) | | | *Silene conica* (n = 5) | | | *Silene noctiflora* (n = 9) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Mitochondrial** | **Length (nt)** | ***S*** | $\pi_S$ | $\pi_N$ | ***S*** | $\pi_S$ | $\pi_N$ | ***S*** | $\pi_S$ | $\pi_N$ | ***S*** | $\pi_S$ | $\pi_N$ |
| *atp1* | 951 | 4 | 0.003 | 0.001 | 24 | 0.023 | 0.001 | 5 | 0.008 | 0.001 | 0 | 0.000 | 0.000 |
| *atp4* | 246-282 | 1 | 0.000 | 0.002 | 0 | 0.000 | 0.000 | 7 | 0.038 | 0.004 | 0 | 0.000 | 0.000 |
| *atp6* | 282 | 1 | 0.002 | 0.000 | 2 | 0.005 | 0.000 | 7 | 0.038 | 0.002 | 0 | 0.000 | 0.000 |
| *cox3* | 597 | 4 | 0.001 | 0.002 | 5 | 0.000 | 0.002 | 2 | 0.003 | 0.001 | 0 | 0.000 | 0.000 |
| *nad9* | 393 | 4 | 0.006 | 0.001 | 1 | 0.000 | 0.001 | 4 | 0.016 | 0.002 | 0 | 0.000 | 0.000 |
| *nad4L-atp4* (intergenic) | 136-513 | 1 | 0.001 | -- | 4 | 0.004 | -- | 2 | 0.009 | -- | 0 | 0.000 | -- |
| Total (coding only) | 2469-2505 | 14 | 0.002 | 0.001 | 32 | 0.010 | 0.001 | 25 | 0.015 | 0.002 | 0 | 0.000 | 0.000 |
| **Plastid** | | | | | | | | | | | | | |
| *trnL* (intronic) | 384-509 | 7 | 0.001 | -- | 7 | 0.001 | -- | 0 | 0.000 | -- | 0 | 0.000 | -- |
| **Nuclear** | | | | | | | | | | | | | |
| fructose-2,6-bisphosphatase (X4/XY4) | 528 | 41 | 0.086 | 0.010 | 44 | 0.042 | 0.004 | 14 | 0.047 | 0.001 | 0 | 0.000 | 0.000 |

*S* = Number of segregating sites
$\pi_S$ = Synonymous (or non-coding in the case of introns and intergenic sequences) nucleotide diversity
$\pi_N$ = Non-synonymous nucleotide diversity


| | | *Silene latifolia* (n = 28) | | | *Silene vulgaris* (n = 40) | | | *Silene conica* (n = 5) | | | *Silene noctiflora* (n = 9) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Mitochondrial** | **Length (nt)** | ***S*** | $\pi_S$ | $\pi_N$ | ***S*** | $\pi_S$ | $\pi_N$ | ***S*** | $\pi_S$ | $\pi_N$ | ***S*** | $\pi_S$ | $\pi_N$ |
| *atp1* | 951 | 4 | 0.003 | 0.001 | 24 | 0.023 | 0.001 | 5 | 0.008 | 0.001 | 0 | 0.000 | 0.000 |
| *atp4* | 246-282 | 1 | 0.000 | 0.002 | 0 | 0.000 | 0.000 | 7 | 0.038 | 0.004 | 0 | 0.000 | 0.000 |
| *atp6* | 282 | 1 | 0.002 | 0.000 | 2 | 0.005 | 0.000 | 7 | 0.038 | 0.002 | 0 | 0.000 | 0.000 |
| *cox3* | 597 | 4 | 0.001 | 0.002 | 5 | 0.000 | 0.002 | 2 | 0.003 | 0.001 | 0 | 0.000 | 0.000 |
| *nad9* | 393 | 4 | 0.006 | 0.001 | 1 | 0.000 | 0.001 | 4 | 0.016 | 0.002 | 0 | 0.000 | 0.000 |
| *nad4L-atp4* (intergenic) | 136-513 | 1 | 0.001 | -- | 4 | 0.004 | -- | 1 | 0.004 | -- | 0 | 0.000 | -- |
| Total (coding only) | 2469-2505 | 14 | 0.002 | 0.001 | 32 | 0.010 | 0.001 | 25 | 0.015 | 0.002 | 0 | 0.000 | 0.000 |
| **Plastid** | | | | | | | | | | | | | |
| *trnL* (intronic) | 384-509 | 7 | 0.001 | -- | 7 | 0.001 | -- | 0 | 0.000 | -- | 0 | 0.000 | -- |
| **Nuclear** | | | | | | | | | | | | | |
| fructose-2,6-bisphosphatase (X4/XY4) | 528 | 41 | 0.086 | 0.010 | 44 | 0.042 | 0.004 | 14 | 0.047 | 0.001 | 0 | 0.000 | 0.000 |

*S* = Number of segregating sites
$\pi_S$ = Synonymous (or non-coding in the case of introns and intergenic sequences) nucleotide diversity
$\pi_N$ = Non-synonymous nucleotide diversity

**Table S4.** Source of *S. noctiflora* and *S. conica* populations for polymorphism studies

| | Population Code | Seed Collection Location/Source |
|---|---|---|
| ***Silene noctiflora*** | OSR (genome) | Giles County, VA, USA |
| | BDA | Budapest, Hungary |
| | BRP | Nelson County, VA, USA |
| | BWT | Tübingen, Germany |
| | OPL | Opole, Poland |
| | PKC | Přední Kopanina, Czech Republic |
| | SGH | Albemarle County, VA, USA (tissue sample only) |
| | TTP | South River, ON, Canada |
| | UMN | Minneapolis, MN, USA (tissue sample only) |
| ***Silene conica*** | ABR (genome) | Abruzzo, Italy (provided by M. Hood; collected by F. Conti) |
| | BOX | DNA provided by B. Oxelman (voucher: P. Erixon 70 UPS) |
| | FBG | Frankfurt Botanical Garden, Germany (provided by L. Gimenez) |
| | KGA | Wroclaw, Poland (provided by Kew Gardens) |
| | KGB | Norfolk, England (provided by Kew Gardens) |

REFERENCES

1. B. F. Lang, M. W. Gray, G. Burger, *Annu. Rev. Genet.* **33**, 351 (1999).

2. T. R. Gregory, *The Evolution of the Genome* (Elsevier, Amsterdam, 2005).

3. M. Lynch, *The Origins of Genome Architecture* (Sinauer Associates, Sunderland, MA, 2007), pp. 494.

4. J. L. Boore, *Nucleic Acids Res.* **27**, 1767 (1999).

5. J. P. Mower, D. B. Sloan, A. J. Alverson, in Plant Genome Diversity, J. F. Wendel, Ed. (Springer, In press), vol. 1.

6. K. H. Wolfe, W. H. Li, P. M. Sharp, *Proc. Natl. Acad. Sci.* **84**, 9054 (1987).

7. M. Lynch, B. Koskella, S. Schaack, *Science* **311**, 1727 (2006).

8. M. Lynch, *Annu. Rev. Microbiol.* **60**, 327 (2006).

9. J. P. Mower, P. Touzet, J. S. Gummow, L. F. Delph, J. D. Palmer, *BMC Evol. Biol.* **7**, 135 (2007).

10. D. B. Sloan, B. Oxelman, A. Rautenberg, D. R. Taylor, *BMC Evol. Biol.* **9**, 260 (2009).

11. D. B. Sloan, A. J. Alverson, H. Storchova, J. D. Palmer, D. R. Taylor, *BMC Evol. Biol.* **10**, 274 (2010).

12. D. A. Petrov, *Theor. Popul. Biol.* **61**, 531 (2002).

13. C. H. Kuo, H. Ochman, *Genome Biol. Evol.* **1**, 145 (2009).

14. V. Knoop, U. Volkmar, J. Hecht, F. Grewe, in Plant Mitochondria, F. Kempken, Ed. (Springer, 2011), pp. 3-29.

15. A. J. Alverson, S. Zhuo, D. W. Rice, D. B. Sloan, J. D. Palmer, *PLoS One* **6**, e16404 (2011).

16. A. J. Bendich, *Curr. Genet.* **24**, 279 (1993).

17. J. D. Palmer, C. R. Shields, *Nature* **307**, 437 (1984).

18. A. Marechal, N. Brisson, *New Phytol.* **186**, 299 (2010).

19. Z. Zhang, B. R. Green, T. Cavalier-Smith, *Nature* **400**, 155 (1999).

20. J. Lukes, H. Hashimi, A. Zikova, *Curr. Genet.* **48**, 277 (2005).

21. R. Shao, E. F. Kirkness, S. C. Barker, *Genome Res.* **19**, 904 (2009).

22. D. B. Sloan, A. H. MacQueen, A. J. Alverson, J. D. Palmer, D. R. Taylor, *Genetics* **185**, 1369 (2010).

23. M. P. Arrieta-Montiel, V. Shedge, J. Davila, A. C. Christensen, S. A. Mackenzie, *Genetics* **183**, 1261 (2009).

24. M. Lynch, J. S. Conery, *Science* **302**, 1401 (2003).

25. V. Daubin, N. A. Moran, *Science* **306**, 978 (2004).

26. B. Charlesworth, N. Barton, *Curr. Biol.* **14**, R233 (2004).

27. C. H. Kuo, N. A. Moran, H. Ochman, *Genome Res.* **19**, 1450 (2009).

28. K. D. Whitney, T. Garland Jr, *PLoS Genet.* **6**, e1001080 (2010).

29. A. E. Vinogradov, *Science* **304**, 389 (2004).

30. S. Yi, J. T. Streelman, *Trends Genet.* **21**, 643 (2005).

31. T. R. Gregory, J. D. Witt, *Genome* **51**, 309 (2008).

32. D. R. Smith, R. W. Lee, *BMC Evol. Biol.* **8**, 156 (2008).

33. D. R. Smith, R. W. Lee, *BMC Evol. Biol.* **9**, 120 (2009).

34. D. R. Smith, R. W. Lee, *Mol. Biol. Evol.* **27**, 2244 (2010).

35. A. J. Alverson *et al.*, *Mol. Biol. Evol.* **27**, 1436 (2010).

36. G. J. Houliston, M. S. Olson, *Genetics* **174**, 1983 (2006).

37. C. M. Barr, S. R. Keller, P. K. Ingvarsson, D. B. Sloan, D. R. Taylor, *Mol. Biol. Evol.* **24**, 1783 (2007).

38. D. B. Sloan, C. M. Barr, M. S. Olson, S. R. Keller, D. R. Taylor, *Mol. Biol. Evol.* **25**, 243 (2008).

39. D. R. Taylor, S. R. Keller, *Evolution* **61**, 334 (2007).

40. P. Touzet, L. F. Delph, *Genetics* **181**, 631 (2009).

41. Y. Cho, J. P. Mower, Y. L. Qiu, J. D. Palmer, *Proc Natl Acad Sci* **101**, 17741 (2004).

42. C. L. Parkinson *et al.*, *BMC Evol. Biol.* **5**, 73 (2005).

43. A. H. Sturtevant, *Q. Rev. Biol.* **12**, 464 (1937).

44. P. D. Sniegowski, P. J. Gerrish, T. Johnson, A. Shaver, *Bioessays* **22**, 1057 (2000).

45. M. Lynch, *Trends Genet.* **26**, 345 (2010).

46. P. Erixon, B. Oxelman, *Mol. Phylogenet. Evol.* **48**, 313 (2008).

47. A. R. Brach, H. Song, *Taxon* **55**, 188 (2006).

48. G. Bernasconi *et al.*, *Heredity* **103**, 5 (2009).

49. J. Jalas, J. Suominen, Eds., *Atlas florae Europaeae. Distribution of vascular plants in Europe. III. Caryophyllaceae* (Cambridge University Press, Cambridge, 1987).

50. A. R. Clapham, T. G. Tutin, E. F. Warburg, *Flora of the British Isles* (Cambridge University Press, Cambridge, 1952).

51. C. Desfeux, S. Maurice, J. P. Henry, B. Lejeune, P. H. Gouyon, *Proc. R. Soc. Lond. B* **263**, 409 (1996).

52. S. H. Folke, L. F. Delph, *Int. J. Plant Sci.* **158**, 501 (1997).

53. S. L. Davis, L. F. Delph, *Int. J. Plant Sci.* **166**, 475 (2005).

54. R. Kolodner, K. K. Tewari, *Proc. Natl. Acad. Sci.* **69**, 1830 (1972).

55. J. D. Palmer, *Nucleic Acids Res.* **10**, 1593 (1982).

56. D. Gordon, C. Abajian, P. Green, *Genome Res.* **8**, 195 (1998).

57. A. J. Bendich, *Mol. Cell* **39**, 831 (2010).

58. A. J. Bendich, *J. Mol. Biol.* **255**, 564 (1996).

59. S. Backert, T. Borner, *Curr. Genet.* **37**, 304 (2000).

60. R. Li *et al.*, *Bioinformatics* **25**, 1966 (2009).

61. C. Camacho *et al.*, *BMC Bioinformatics* **10**, 421 (2009).

62. T. M. Lowe, S. R. Eddy, *Nucleic Acids Res.* **25**, 955 (1997).

63. W. Hao, J. D. Palmer, *Proc. Natl. Acad. Sci.* **106**, 16728 (2009).

64. G. Benson, *Nucleic Acids Res.* **27**, 573 (1999).

65. R. C. Edgar, *Nucleic Acids Res.* **32**, 1792 (2004).

66. Z. Yang, *Mol. Biol. Evol.* **24**, 1586 (2007).

67. M. Z. Lu, A. E. Szmidt, X. R. Wang, *Plant Mol. Biol.* **37**, 225 (1998).

68. P. Giege, A. Brennicke, *Proc. Natl. Acad. Sci.* **96**, 15324 (1999).

69. J. P. Mower, J. D. Palmer, *Mol. Genet. Genomics* **276**, 285 (2006).

70. B. Frajman, F. Eggens, B. Oxelman, *Systematic Biology* **58**, 328 (2009).

71. G. I. Peterson, J. Masel, *Mol. Biol. Evol.* **26**, 2595 (2009).

72. D. Charlesworth, *Heredity* **105**, 509 (2010).

73. J. P. Mower, *Nucleic Acids Res.* **37**, W253 (2009).

74. P. Librado, J. Rozas, *Bioinformatics* **25**, 1451 (2009).

75. M. Stephens, N. J. Smith, P. Donnelly, *Am. J. Hum. Genet.* **68**, 978 (2001).

76. K. L. Adams, Y. L. Qiu, M. Stoutemyer, J. D. Palmer, *Proc Natl Acad Sci* **99**, 9905 (2002).

77. P. Leon, V. Walbot, P. Bedinger, *Nucleic Acids Res.* **17**, 4089 (1989).

78. F. Grewe, P. Viehoever, B. Weisshaar, V. Knoop, *Nucleic Acids Res.* **37**, 5093 (2009).

79. F. Grewe *et al.*, *Nucleic Acids Res.* (In Press).

# Chapter 8.

# Recent acceleration of plastid sequence and structural evolution coincides with extreme mitochondrial divergence within the angiosperm genus *Silene* [1]

[1]Formatted as a co-authored manuscript (Sloan DB, Alverson AJ, Wu M, Palmer JD, Taylor DR) tentatively planned for submission to *Genome Biology and Evolution*.

ABSTRACT

The angiosperm genus *Silene* exhibits some of the most extreme and rapid divergence ever identified in mitochondrial genome architecture and substitution rates. Although organelle genomes share a high degree of functional interdependence and many components of their genetic machinery, the patterns of divergence within *Silene* have been considered mitochondrial-specific based on the absence of correlated changes in the small number of available nuclear and plastid (chloroplast) gene sequences. To better assess the potential relationship between mitochondrial and plastid evolution, we sequenced the plastid genomes from four *Silene* species with fully sequenced mitochondrial genomes. We found that two species with fast-evolving mitochondrial genomes, *S. noctiflora* and *S. conica*, also exhibit accelerated rates of sequence and structural evolution in their plastid genomes, but the specific nature of these changes is markedly different from those in the mitochondrial genome. For example, in contrast to the pattern observed in mitochondrial DNA, which applies to all genes and appears to be mutationally driven, the plastid substitution rate accelerations are restricted to a subset of genes and preferentially affect non-synonymous sites, indicating altered selection pressures are acting on specific plastid-encoded functions in these species. Indeed, some *S. noctiflora* and *S. conica* plastid genes show strong evidence of positive selection with $d_N/d_S$ ratios significantly greater than one. In contrast, two species with more slowly evolving mitochondrial genomes, *S. latifolia* and *S. vulgaris*, also show low rates of nucleotide substitution in plastid genes and have a plastid genome structure that has remained essentially unchanged since the origin of angiosperms. These results raise the

possibility that common evolutionary forces could be shaping the extreme but distinct patterns of divergence in both organelle genomes within this genus.

## INTRODUCTION

Plants and other photosynthetic eukaryotes share the distinction of having a second endosymbiotically derived organelle, the plastid, that coexists with mitochondria in the cytoplasm (Gould et al. 2008; Kim and Archibald 2009). There are clear parallels in the long-term evolution of mitochondrial and plastid genomes. For example, both have experienced massive gene loss (Adams and Palmer 2003; Timmis et al. 2004), which appears to be a universal pattern in obligately intracellular symbionts (Andersson and Kurland 1998; Moran and Wernegreen 2000). Because organelle gene loss has generally been associated with transfer of genetic control to the nucleus, most of the genes required for organelle function are now found in the nuclear genome, including essentially all those responsible for the maintenance of organellar DNA (Day and Madesis 2007; Chapter 1). Interestingly, many of the plant genes involved in DNA replication and repair in one organelle genome have related paralogs that function in the other (Zaegel et al. 2006; Shedge et al. 2007; Cappadocia et al. 2010). Furthermore, many nuclear genes are targeted to both organelles, including a disproportionate fraction of genes associated with DNA synthesis and processing (Carrie et al. 2009). Therefore, the evolution of DNA replication and repair machinery in organelles involves a complex history of transferring, co-opting, duplicating, re-targeting, and replacing genes (reviewed in Chapter 1).

Despite sharing many components of their DNA replication and repair machinery, mitochondrial and plastid genomes differ greatly in their structural organization and evolution. For example, seed plant plastid genomes are gene-dense and exhibit a high degree of syntenic conservation (Raubeson and Jansen 2005). In contrast, seed plant mitochondrial genomes contain an abundance of non-coding sequence and experience

rapid rates of rearrangement among and even within species (Mower et al. In press). Mitochondrial and plastid genomes also exhibit different rates of nucleotide substitution, which are believed to reflect underlying differences in mutation rate. Rates of synonymous substitutions are typically two to four times faster in plastid DNA than mitochondrial DNA in seed plants (Wolfe et al. 1987; Palmer and Herbon 1988; Drouin et al. 2008). However, a handful of seed plant lineages exhibit dramatic increases in mitochondrial substitution rates, even reaching levels that are more typical of fast-evolving animal mitochondrial genomes (Cho et al. 2004; Parkinson et al. 2005; Bakker et al. 2006; Mower et al. 2007; Sloan et al. 2008; Sloan et al. 2009; Ran et al. 2010).

Observed cases of rate acceleration in plant mitochondrial DNA have often been described as mitochondrial-specific phenomena because sequenced nuclear and plastid genes have shown little or no correlated increase in substitution rate (e.g., Cho et al. 2004; Mower et al. 2007; Sloan et al. 2008). Nevertheless, there is limited evidence to suggest that these changes in mitochondrial rate may not be entirely independent of evolution in the plastid genome. For example, the Geraniaceae, which has experienced a series of extreme changes in mitochondrial substitution rate (Parkinson et al. 2005), also exhibits abnormally high rates of structural evolution in the plastid genome and accelerated substitution rates in a subset of plastid genes (Chumley et al. 2006; Guisinger et al. 2008; Guisinger et al. 2011; Blazier et al. In Press). Likewise, gnetophytes exhibit elevated rates of evolution in both plastid and mitochondrial genomes (Mower et al. 2007; McCoy et al. 2008; Wu et al. 2009). However, comparisons based on complete genomes from both the mitochondria and plastids are not available in these cases to assess the potential evolutionary parallels between the organelle genomes.

We recently reported the complete mitochondrial genome sequence of four *Silene* species with highly divergent mitochondrial substitution rates (Chapter 7). Two of these species (*S. noctiflora* and *S. conica*) exhibit nearly 100-fold increases in synonymous substitution rates, whereas the other two (*S. latifolia* and *S. vulgaris*) have maintained low rates that are more typical of other angiosperm mitochondrial genomes. The accelerated rates in *S. noctiflora* and *S. conica* are associated with unprecedented expansions in mitochondrial genome size (from a few hundred kb to many Mb) as well as numerous other changes in mitochondrial gene and genome architecture.

To assess the relationship, if any, between mitochondrial and plastid genome evolution in *Silene*, we sequenced the complete plastid genomes from the same four species. The fast-evolving mitochondrial lineages (*S. noctiflora* and *S. conica*) do not show evidence of comparable genome-wide increases in plastid synonymous substitution rates. However, they do exhibit substantial rate accelerations in a subset of plastid genes, particularly at non-synonymous sites, suggesting that altered selection pressures are acting on specific plastid pathways in these species. In addition, the *S. noctiflora* and *S. conica* plastid genomes have experienced rapid structural evolution. In contrast, the *S. latifolia* and *S. vulgaris* plastid genomes are highly conserved relative to other angiosperms. These results provide an example of recent and correlated accelerations in mitochondrial and plastid genome evolution among closely related species, but the specific patterns of sequence and structural change differ between the two organelle genomes in many fundamental respects. We discuss the possibility that shared forces are acting, either directly or indirectly, on both mitochondrial and plastid genomes in *Silene*.

RESULTS

**Gene content in *Silene* plastid genomes**

The complete plastid genomes of four *Silene* species (*S. latifolia*, *S. vulgaris*, *S. noctiflora*, and *S. conica*) were sequenced using a combination of Roche 454 and Illumina technology (Fig. 1, Fig. S1). The *Silene* plastid genomes are typical in size relative to other angiosperms and exhibit a classic circular genome map with a pair of large inverted repeats (IRs) separating two single-copy regions (Table 1). The four genomes share a gene complement that encodes 77 proteins, 30 tRNAs, and 4 rRNAs. Genes coding for the translation initiation factor A (*infA*) and the ribosomal protein subunit L23 (*rpl23*) appear to be present only as pseudogenes in the genomes of all four species and are not included in the totals above. These two genes have been lost independently in multiple angiosperm lineages, including other species within the Caryophyllales (Zurawski and Clegg 1987; Millen et al. 2001; Funk et al. 2007; Logacheva et al. 2008). The *infA* gene has been subject to repeated functional transfers to the nucleus (Millen et al. 2001), whereas there is evidence that *rpl23* has been functionally replaced by its cytosolic counterpart in other species (Bubunenko et al. 1994). In addition to the functional loss of *infA* and *rpl23*, it is possible that other annotated genes in the *Silene* plastid genomes are actually pseudogenes. For example, as reported previously (Sloan et al. 2009), the intron-encoded open reading frame (ORF) *matK* contains an internal frameshift in *S. conica*. The gene encoding the RNA polymerase α subunit (*rpoA*) also contains an internal frameshift approximately 200 bp upstream of the normal stop codon position in three of the four species (*S. latifolia*, *S. vulgaris*, and *S. noctiflora*). In both *matK* and *rpoA*, frameshifts have occurred in

homopolymer (i.e., single-nucleotide repeat) regions and have introduced premature stop codons. Finally, multiple genes are highly divergent in sequence and/or structure in *S. conica* and *S. noctiflora* (see below) and could be pseudogenes. In particular, the *accD* gene in *S. noctiflora* has experienced a major deletion at its 3' end and, therefore, codes for a product that is only about one-third of the typical length.

**Rapid structural evolution in *Silene noctiflora* and *Silene conica* plastid genomes**

The *S. latifolia* and *S. vulgaris* plastid genomes show nearly perfect syntenic conservation with *Spinacia* (Fig. 2) and other angiosperms (data not shown), suggesting that these two species have maintained the ancestral angiosperm genomic structure (Raubeson and Jansen 2005) for well over 100 million years (Bell et al. 2010). In contrast, the two species with fast-evolving mitochondrial genomes (*S. noctiflora* and *S. conica*) have experienced numerous changes in plastid genome structure in just the few million years since the divergence of these four *Silene* species, including multiple inversions, intron losses, large indels, and shifts in the IR boundaries.

The gene order in the *S. noctiflora* plastid genome suggests that it has experienced four inversions involving six breakpoints found in or between the following gene pairs: *psbM-trnD*, *accD-psaI*, *psbB-clpP*, *petL-psbE, psbD-trnT, trnT-trnE* (Fig. 2). The *S. conica* genome appears to have experienced a single inversion with a pair of breakpoints (*psaA-ycf3* and *psaI-ycf4*) that are distinct from any of those involved in the *S. noctiflora* rearrangements (Fig. 2). At least some of the observed inversions are likely the result of recombination between short inverted repeats (Knox et al. 1993). All four *Silene* species have a pair of divergent inverted repeats (approximately 170 bp and 80% sequence

identity) that coincide with the breakpoints for the *S. conica* inversion. Likewise, *S. noctiflora* has a unique pair of inverted repeats (154 bp, 99% sequence identity) corresponding to the *petL-psbE* and *psbD-trnT* breakpoints. However, repeats associated with other inversion events in *S. noctiflora* are not readily identifiable. Interestingly, the *S. conica* inversion interrupts a genomic fragment that was previously sequenced and analyzed in a number of species within the *Sileneae* including *S. conica* (Erixon and Oxelman 2008a). This earlier study did not detect the inversion found in our analysis of the *S. conica* plastid genome, which could indicate that it is polymorphic within the species. However, the 2008 study was based on PCR and Sanger sequencing of individual fragments, so artifacts involving PCR-mediated recombination are also possible explanations for the discrepancy (Alverson et al. 2011).

The *S. latifolia* and *S. vulgaris* plastid genomes are predicted to share an identical complement of 19 group II introns (including the *trans*-splicing first intron of *rps12* (Koller et al. 1987)) as well as a single group I intron in the *trnL-UAA* gene. The *S. noctiflora* and *S. conica* genomes each lack four of these introns. Both species have lost the *rpoC1* intron as well as both introns in the fast-evolving *clpP* gene (Erixon and Oxelman 2008b). In addition, *S. noctiflora* and *S. conica* have uniquely lost the *rpl16* and *atpF* introns, respectively. All five of these introns have been lost independently multiple times in other angiosperms (Downie et al. 1996; Campagna and Downie 1998; Jansen et al. 2007). Like other members of the core Caryophyllales, all four *Silene* species lack the *rpl2* intron found in other plant lineages (Downie et al. 1991; Logacheva et al. 2008). In every case, missing introns have been precisely excised at their normal splicing boundaries.

In addition to the deletions associated with intron loss, the *S. noctiflora* and *S. conica* plastid genomes have also experienced a total of 10 and 11 large indels of >100 bp in size, respectively. Ten of these large indel events are found in coding sequence (all in the highly divergent *accD*, *ycf1*, and *ycf2* genes), while the remainder are in non-coding sequences (10 in intergenic regions and one in the *trnL-UAA* intron). In contrast, no indels of this size have occurred in the *S. latifolia* or *S. vulgaris* lineages. The *S. noctiflora* and *S. conica* plastid genomes also appear to have experienced a higher frequency of small indels. Alignments of all four *Silene* species with the outgroup *Spinacia oleracea* identified a total of 18, 46, 107, and 151 unique, non-overlapping indels of >100 bp in size for *S. latifolia*, *S. vulgaris*, *S. conica*, and *S. noctiflora*, respectively

The structures of the *S. noctiflora* and *S. conica* plastid genomes have also been altered by changes in the boundaries between their IRs and single-copy regions. Although the precise boundaries of the IR in angiosperms are subject to frequent changes (Goulding et al. 1996), they are generally found within the coding sequences of *rps19* and *ycf1*, which appears to be the ancestral state for most angiosperm lineages and is true for both *S. latifolia* and *S. vulgaris*. These two species share identical boundary positions between the IR and large single-copy region and differ only slightly (<100 bp) in the positions of their boundaries between the IR and small single-copy region. In contrast, the IR in *S. noctiflora* and *S. conica* has contracted at the boundary with the large single-copy region and expanded at the boundary with the small single-copy region (Fig. 1). As a result, the IR in *S. noctiflora* and *S. conica* does not contain any portion of *rps19* and lacks a substantial fraction of *rpl2*. In addition, the IR now includes the entirety of *ycf1* in

both species, as well as *rps15* and a portion of *ndhH* in *S. noctiflora*. Although the directions of these apparent boundary shifts are the same in both *S. noctiflora* and *S. conica*, the magnitudes of the expansions and contractions differ (Fig. 1). Differences in IR boundary positions account for 0.7 kb of the observed 3 kb difference in IR length between *S. noctiflora* and *S. conica* (Table 1). The remaining 2.3 kb difference in IR length between these species is the result of indels within the IR, particularly in the coding sequences of *ycf1* and *ycf2*.

Although it appears likely that the gross differences in IR boundary positions among *Silene* species are the result of changes in *S. noctiflora* and *S. conica*, it is difficult to confidently infer the ancestral state for this group. The IR boundary positions in *S. latifolia* and *S. vulgaris* are similar but not identical to a large number of other angiosperms, including *Spinacia oleracea*, the closest outgroup with a sequenced plastid genome (Schmitz-Linneweber et al. 2001). Interestingly, the only other sequenced plastid genome within the Caryophyllales (from *Fagopyrum esculentum*) also has an expanded IR that contains a full copy of *ycf1* (Logacheva et al. 2008). Therefore, an alternative possibility is that the most recent common ancestor of the Caryophyllales had an expanded plastid IR that included the entirety of *ycf1* and that the positions of the IR boundaries within *ycf1* in *S. latifolia*, *S. vulgaris*, and *Spinacia* are the result of multiple reversions.

**Elevated and variable substitution rates in the plastid genomes of *Silene noctiflora* and *Silene conica***

*Silene noctiflora* and *S. conica* exhibit increased substitution rates in plastid genes. However, the observed rate accelerations differ in two important respects relative to the dramatic rate increases in the mitochondrial genomes of these species. First, the elevated plastid rates are primarily driven by a disproportionate increase in the frequency of non-synonymous substitutions ($d_N$) with only a modest two- or three-fold change in synonymous divergence ($d_S$). In contrast, mitochondrial $d_N$ and $d_S$ values have each increased by nearly two orders of magnitude in these species, resulting in virtually no change in the $d_N/d_S$ ratio (Fig. 3). Second, whereas the mitochondrial rate accelerations in *S. noctiflora* and *S. conica* appear to be fairly uniform and genome-wide phenomena (Chapter 7), plastid rates differ dramatically among genes (Fig. 4, Fig. S2).

Plastid genes in five major complexes associated with photosynthesis show little or no rate increase in *S. noctiflora* and *S. conica* (Fig. 4A). In contrast, informational protein genes including RNA polymerase subunits and particularly ribosomal proteins show more substantial increases (Fig. 4B). Some additional plastid genes have experienced even greater rate changes, including the large ORFs *ycf1* and *ycf2* (which are known to be essential for cell survival but are otherwise uncharacterized; Drescher et al. 2000) as well as the protease subunit *clpP* (Fig. 4C.), which was previously found to have highly accelerated substitution rates in multiple lineages within the tribe *Sileneae*, including *S. conica* (Erixon and Oxelman 2008b). The *accD* gene, which is required for fatty acid biosynthesis, shows some evidence of substitution rate acceleration (Fig. S2) and has also undergone rapid structural evolution including large deletions in both *S. noctiflora* and *S. conica*.

**Phylogenetic analysis of *Silene* plastid DNA**

The results from analysis of multiple concatenated datasets do not provide a clear consensus on the phylogenetic history of the four *Silene* species investigated in this study. A concatenated dataset of all plastid protein genes except *accD, clpP, ycf1*, and *ycf2* supports a sister relationship between the fast-evolving *S. noctiflora* and *S. conica* lineages (Fig. 5A). However, this support disappears when the analysis is restricted to photosynthesis-related genes (Fig. 5B), which do not exhibit major rate accelerations in *S. noctiflora* and *S. conica* (Fig. 4A). Instead, when analyzed separately, these genes support a sister relationship between *S. latifolia* and *S. conica* (Fig. 5B). Analysis of shared intron sequences provides weak support for yet another topology with *S. latifolia* sister to *S. noctiflora* (Fig. 5C). In all three analyses, internal branch lengths are very short, indicating a rapid radiation of all four *Silene* lineages.

## DISCUSSION

**Recent and correlated changes in mitochondrial and plastid genome evolution**

Recent sequencing of the *S. noctiflora* and *S. conica* mitochondrial genomes revealed that they are exceptional even compared to the already complex mitochondrial genomes of most flowering plants, exhibiting extreme changes in genomic architecture and rate of sequence evolution (Chapter 7). In this study, we have shown that the plastid genomes in these species have also experienced recent and rapid divergence that distinguishes them from "typical" plastid genomes of angiosperms, including other members of the same genus. Although comparisons of complete mitochondrial and plastid genome sequences have not been performed in other angiosperm species with accelerated mitochondrial

substitution rates, there is some evidence to suggest that similar correlated increases in the rate of sequence and/or structural evolution in both organelle genomes have occurred in lineages such as the Geraniaceae and gnetophytes (Parkinson et al. 2005; Chumley et al. 2006; Mower et al. 2007; Guisinger et al. 2008; McCoy et al. 2008; Wu et al. 2009; Guisinger et al. 2011; Blazier et al. In Press).

These cases constitute a scant number of data points, but they raise the possibility that a shared mechanism may be affecting both organelle genomes. The mapping and sequencing of angiosperm plastid genomes has far outpaced the progress on mitochondrial genomes. As a result, there are numerous angiosperm lineages that have been identified as having accelerated and/or rearranged plastid genomes, but for which we have little or no mitochondrial data, such as the Campanulaceae (including the Lobeliaceae), Fabaceae, Goodeniaceae, Oleaceae, Passifloraceae, and Ranunculaceae (Jansen et al. 2007; Jansen et al. 2008 and references therein). Many of these lineages contain plastid genomes that are far more divergent and rearranged than those found in *Silene* and, therefore, represent a natural starting place for generating additional mitochondrial genome sequences. It is unlikely that there is any simple or absolute relationship between the organelle genomes. For example, note that some of the most divergent plastid genomes in the Geraniaceae (Guisinger et al. 2011; Blazier et al. In Press) occur in genera with only moderately accelerated mitochondrial substitution rates (Parkinson et al. 2005). Nevertheless, more comprehensive comparisons of organelle genomes across angiosperms may help identify mechanisms that jointly affect mitochondrial and plastid genome evolution.

The idea that rates of sequence evolution might be correlated between mitochondrial and plastid genomes is not new. In fact, there are many factors expected to affect rates and patterns of evolution at an organismal level (Ohta 1992; Whittle and Johnston 2002; Smith and Donoghue 2008). Therefore, one of the intriguing elements of this study is not necessarily that the mitochondrial and plastid genomes are both highly divergent in *S. noctiflora* and *S. conica*, but that they are divergent in such different ways (Table 2). Our findings raise the question of what evolutionary mechanisms could generate these correlated, yet distinct, patterns of divergence between the mitochondrial and plastid genomes. There are many potential answers to this question (including plain and simple coincidence), but one intriguing possibility involves modification of nuclear genes coding for dual targeted protein products. For example, homologs of the bacterial *recA* gene are known to play an important role in plant organelle genome stability, and the *Arabidopsis* genome contains three characterized *recA* homologs with one targeted to the plastids, one targeted to the mitochondria, and one targeted to both organelles (*RECA2*) (Shedge et al. 2007; Rowan et al. 2010). Modification of the dual targeted (*RECA2*) gene could affect the evolution of both genomes but in potentially different ways given the possibility that the gene product serves different functional roles in the two organelles or maintains different levels of redundancy with other members of the gene family.

The discovery and history of the bacterial *mutS* homolog *MSH1* may also be informative with respect to correlated patterns of evolution between mitochondrial and plastid genomes. This nuclear locus was originally named *CHM* (for chloroplast mutator), because mutants exhibited a variegated leaf phenotype and modifications in

plastid morphology that could subsequently be inherited maternally (Redei 1973). Therefore, it was predicted that disruptions of this nuclear gene destabilize the plastid genome. Subsequent work, however, has shown that the *MSH1* gene product is predominantly, perhaps solely, targeted to the mitochondria where it regulates recombinational activity and genome reorganization (Martinez-Zapater et al. 1992; Abdelnoor et al. 2003; Shedge et al. 2007; Arrieta-Montiel et al. 2009). Therefore, the documented effects of *MSH1/CHM* on plastids may be mediated indirectly through physiological pathways linking these two organelles. Because mitochondria and plastids maintain a high degree of functional interdependence, (Roussell et al. 1991; Woodson and Chory 2008; Yoshida and Noguchi 2011), it is possible that perturbation of one organelle genome will have direct evolutionary consequences for the other. *MSH1, RECA*, and other gene families known to be involved in plant organelle genome stability (e.g., Zaegel et al. 2006; Cappadocia et al. 2010) represent important candidates for further investigation in *Silene*.

**Causes of substitution rate variation among plastid genes**

The pattern of mitochondrial substitution rate acceleration in *S. noctiflora* and *S. conica* has been attributed to genome-wide increases in the mutation rate (Mower et al. 2007; Sloan et al. 2009; Chapter 7). However, a similar explanation appears to be inconsistent with the observed substitution patterns in the plastid genomes of these same species. The magnitude of rate accelerations in *S. noctiflora* and *S. conica* vary markedly across plastid genes. Some of this variation might be explained by "localized hypermutation" as have been proposed in cases of gene-specific rate accelerations in both plastid (Magee et

al. 2010) and mitochondrial (Sloan et al. 2009) genomes. However, even a model with a diverse range of localized, gene-specific mutation rates could not explain the disproportional increases in $d_N$ found in many genes. Instead, the observed increases in $d_N/d_S$ suggest a history of relaxed purifying selection and/or increased positive selection acting on plastid genes in *S. noctiflora* and *S. conica*. Some loci exhibit $d_N/d_S$ ratios that are significantly greater than one when averaged across the entire length of the gene (Table 3), strongly suggesting at least some role for positive selection in the rate accelerations observed in these species.

The differences in substitution rate and $d_N/d_S$ across functional classes of plastid genes (Fig. 4, Fig. S2) suggest that changes in selection pressure may be associated with specific biochemical pathways rather than the entire genome. Interestingly, the patterns of rate variation among genes in *S. noctiflora* and *S. conica* exhibit some clear parallels with the evolution of plastid genomes within the Geraniaceae, which have experienced a longer and more extreme history of genome rearrangement (Chumley et al. 2006; Guisinger et al. 2008; Guisinger et al. 2011; Blazier et al. In Press). For example, both lineages show a high degree of sequence conservation in genes directly involved in photosynthesis and greater levels of divergence in other genes such as ribosomal proteins. Furthermore, the most divergent genes in *S. noctiflora* and *S. conica*, including *accD, clpP, ycf1* and *ycf2*, have been lost completely within multiple lineages in the Geraniaceae (Guisinger et al. 2008; Guisinger et al. 2011). The parallels are not perfect, however. Some of the highest levels of divergence in the Geraniaceae are found in genes coding for RNA polymerase subunits (Guisinger et al. 2008), which show only modest accelerations in *S. noctiflora* and *S. conica* (Fig. 4). In addition, one clade within the

Geraniaceae appears to have lost all functional copies of its *ndh* genes (Blazier et al. In Press), which are highly conserved in *Silene*.

The evolution of plastid genomes in non-photosynthetic angiosperms may also provide insight into the patterns of selection acting on *Silene* plastid genes. As expected, evolution of a non-photosynthetic lifestyle is generally associated with plastid genome reduction and gene loss (Wolfe et al. 1992; Delannoy et al. In Press). Nevertheless, non-photosynthetic angiosperms still retain a plastid genome, demonstrating that the functional importance of plastids extends beyond photosynthetic pathways to include more general biosynthetic roles within the cell (Barbrook et al. 2006; Benning et al. 2006). Many of the genes retained in the plastid genomes of non-photosynthetic plants are often required for plastid gene expression. For example, of the 42 functional genes identified in the highly reduced plastid genome of the parasitic eudicot *Epifagus virginiana*, only four (*accD, clpP*, *ycf1* and *ycf2*) are not involved in plastid gene expression (Wolfe et al. 1992). The non-photosynthetic orchid *Rizanthella gardneri*, which has the smallest sequenced plastid genome of any land plant, has independently converged on a remarkably similar set of genes (including *accD, clpP*, *ycf1* and *ycf2*) (Delannoy et al. In Press).

Strikingly, these same four genes exhibit the greatest accelerations in the rate of sequence and/or structural evolution in *S. noctiflora* and *S. conica*, suggesting that there have been significant changes in selection pressures acting on non-photosynthetic pathways in plastids in both *Silene* species. Although all four of these genes are widely conserved in lands plants (Delannoy et al. In Press), each has been lost from the plastid genome of some lineages, including multiple angiosperms (Katayama and Ogihara 1996;

Knox and Palmer 1999; Chumley et al. 2006; Haberle et al. 2008; Guisinger et al. 2011). There is also evidence for positive selection acting on these genes in other independent lineages of land plants (Erixon and Oxelman 2008b; Greiner et al. 2008). Knockout experiments in tobacco have shown that all four are essential (Drescher et al. 2000; Shikanai et al. 2001; Kuroda and Maliga 2003; Kode et al. 2005). The protein encoded by *clpP* is a component of a complex multimeric protease with broad substrate specificity within the plastid (Peltier et al. 2004; Stanne et al. 2009), whereas *accD* codes for a subunit of the acetyl-CoA carboxylase, which is involved in fatty acid biosynthesis (Kode et al. 2005). Despite the essential nature of *ycf1* and *ycf2* (Drescher et al. 2000), the specific functions of these genes have not yet been characterized. *Silene* and other lineages with a history of extreme divergence in *accD*, *clpP*, *ycf1*, and *ycf2* may provide an opportunity to better understand the more general role of these genes and their related pathways in plants.

**Single or multiple origins of accelerated organelle genome evolution in *Silene*?**

The clear similarities between *S. noctiflora* and *S. conica* in the evolution of both mitochondrial and plastid genomes raise the obvious question of whether these lineages form a monophyletic group that experienced shared ancestral changes associated with the organelle genomes. Although it is tempting to assume that commonalities between these species (e.g., shared intron losses in *clpP* and *rpoC1*) reflect common ancestry, phylogenetic analyses in other angiosperms have shown that such patterns can and do occur in parallel across independent evolutionary lineages (e.g., Guisinger et al. 2011). Therefore, an independent phylogenetic estimate of the relationships between these

*Silene* lineages would be desirable. However, such analyses have generally failed to resolve relationships among the major lineages of *Silene* subgenus *Behenantha*, including the species used in this study (Erixon and Oxelman 2008a; Sloan et al. 2009). Our analyses based on data from complete plastid genomes yielded similar ambiguities (Fig. 5). Likewise, an analysis of a small number of nuclear genes across this genus found that some loci support monophyly between the *S. noctiflora* and *S. conica* lineages, while others do not (A. Rautenberg, D.B. Sloan, V. Aldén, and B. Oxelman, unpublished results). Therefore, the question of single vs. multiple origins of accelerated organelle genome evolution in *Silene* remains unresolved. Efforts are underway to produce deep transcriptome sequencing coverage of multiple *Silene* species. The resulting dataset should help disentangle the phylogenetic relationships within *Silene* as well as elucidate the cyto-nuclear interactions that have shaped the extreme patterns of organelle genome evolution in this genus.

## METHODS

### Source material and plastid DNA extraction

For each of four *Silene* species (*S. latifolia* Poir., *S. vulgaris* (Moench) Garcke, *S. noctiflora* L., and *S. conica* L.), approximately 200g of fresh tissue was collected from multiple individuals from a single maternal family. The maternal families and collection methods correspond to those previously described for mitochondrial genome sequencing (Sloan et al. 2010b; Chapter 7). Intact chloroplasts were isolated using a combination of differential centrifugation and separation on a sucrose step gradient (Palmer 1986; Jansen

et al. 2005). Chloroplasts were then lysed, and DNA was purified by phenol:chloroform extraction. These preparations yielded between 4 and 20 μg of DNA per species. The purity of plastid DNA was confirmed by restriction digestion.

**Roche 454 and Illumina sequencing**

For each plastid DNA sample, shotgun libraries were constructed with multiplex identifier (MID) tags following standard protocols for sequencing on a Roche 454 GS-FLX platform with Titanium reagents. MID-tagged libraries were sequenced as part of a larger pooled sample with each of the four species constituting the equivalent of 2.5% of a full 454 plate. All 454 library construction and sequencing was performed at the Genomics Core Facility in the University of Virginia's Department of Biology.

Multiplex barcoded libraries were also prepared for paired-end sequencing on an Illumina GAII sequencing platform as described previously (Chapter 7). For *S. noctiflora*, plastid DNA was amplified with GenomiPhi V2 (GE Healthcare, Piscataway, NJ) to produce sufficient starting material for Illumina library construction. All other libraries were generated without whole genome amplification. The barcoded libraries were sequenced as part of a larger pooled sample in a single Illumina lane on a 2 x 85 bp paired-end run with each species representing 8% of the pool. Illumina sequencing was performed at the Biomolecular Research Facility in the University of Virginia's School of Medicine.

**Genome assembly and annotation**

Shotgun 454 sequencing produced between 3.7 and 7.1 Mb of sequence data for each species. These reads were assembled with Roche's GS *de novo* Assembler v2.3 ("Newbler") using default settings. Initial assembly produced complete or nearly complete plastid genome sequences. Sequencing coverage in single-copy regions for each of the four species ranged from 21 to 46x, and, as expected, roughly twice those coverage levels were obtained for the IR. The assemblies for each species contained as many as three gaps, but these generally reflected uncertainty regarding the length of long homopolymer regions. These regions were combined and then corrected with Illumina data (see below) to produce finished genomes.

454 data are known to have high insertion and deletion error rates associated with long homopolymer regions. To correct errors in the 454 assembly, paired-end Illumina reads were mapped to the genome using SOAP v2.20 (Li et al. 2009) as described previously (Chapter 7). After quality trimming and removal of multiplex barcode sequences, the Illumina run produced between 40 and 259 Mb of sequence with an average read length between 60 and 65 bp for each species. This dataset provided deep coverage for the entirety of all four genomes with an average read depth between 297 and 1400x. The Illumina mapping results were used to identify and correct between 50 and 96 sequencing errors per genome, the vast majority of which were associated with homopolymer lengths.

Protein, tRNA, and rRNA gene content in each of the finished genomes was annotated using DOGMA (Wyman et al. 2004). The resulting annotated genome sequences were deposited to GenBank (accessions JF715054-JF715057).

**Analysis of genomic inversions and indels**

To reconstruct the history of large inversions in *Silene* plastid genomes, gene order and orientation in each genome was compared to the inferred ancestral state for angiosperms (Raubeson and Jansen 2005) using GRIMM v2.0.1 (Tesler 2002). In addition, all four *Silene* plastid genomes were aligned with the outgroup *Spinacia oleracea* using MAUVE v2.3.1 (Darling et al. 2010).

To identify and quantify the number of indels in each plastid genome, syntenic blocks of sequence for all four *Silene* species and the outgroup *Spinacia oleracea* were aligned using MUSCLE v3.7 (Edgar 2004). Intergenic regions containing inversion breakpoints were not included in this analysis. Large indels (>100 bp) were identified by manual inspection of the sequence alignments. In many cases, the size, number, and polarity of smaller indel events were ambiguous because multiple indels often overlap in structurally variable regions. Therefore, to estimate the relative frequency of smaller indels (<100 bp) in each species, we restricted our focus to the subset of events that are unique to a single species within the aligned dataset and show no overlap with structural variants in the other four species. A custom Perl script was used to identify all indels meeting these criteria.

**Phylogenetic analysis and substitution rate variation**

To assess the phylogenetic relationships among the four *Silene* species, nucleotide sequences from all *Silene* protein genes and introns were aligned with the corresponding sequences from the closest available outgroup, *Spinacia oleracea*, as well as *Arabidopsis thaliana* (for protein-coding sequences only). Alignments were performed using

MUSCLE v3.7 (Edgar 2004) and adjusted manually. Phylogenetic analyses were performed with RAxML v7.0.4 on three different concatenated datasets: 1) all protein genes except *accD, clpP, ycf1,* and *ycf2* (which were excluded because of extreme sequence and/or structural divergence in *S. noctiflora* and *S. conica*), 2) all protein genes in the photosynthesis related *atp, pet, ndh, psa,* and *psb* complexes, and 3) all introns. RAxML analyses were performed with the following parameters: -f d, -b 1, -p 1, -#1000, and -m GTRGAMMA.

The relative rates of sequence divergence in the four *Silene* genomes (and the outgroups *Spinacia* and *Arabidopsis*) were analyzed using both codon- and nucleotide-based models of evolution in PAML v 4.4 (Yang 2007) as described previously (Sloan et al. 2009; Sloan et al. 2010b). Because the phylogenetic relationships among the four *Silene* species are not confidently resolved (Fig. 5), all PAML analyses implemented a constrained topology with the four *Silene* species radiating from a single polytomy. Protein coding sequences were analyzed with codon-based models to separately quantify the rate of synonymous and non-synonymous substitution, whereas RNA genes and intronic sequences were analyzed with nucleotide-based models. Analyses were performed on the following concatenated and individual gene datasets: 1) a concatenation of all protein genes except *accD* (see below); 2) separate concatenations of each of the following protein gene sets: *atp, pet, ndh, psa, psb, rpl, rpo,* and *rps*; 3) each of the following individual protein genes: *ccsA, cemA, clpP, matK, rbcL, ycf1, ycf2, ycf3,* and *ycf4*; 4) a concatenation of all rRNA genes; and 5) a concatenation of all introns. The *accD* gene is too structurally divergent in *S. noctiflora* and *S. conica* to produce a useful alignment that includes both species. However, a large portion of *accD* from each of

these species can be separately aligned against the remaining species in the analysis. Therefore, two separate analyses of *accD* sequence divergence were performed, one involving *S. noctiflora* and one involving *S. conica*.

To test for evidence of positive selection acting on individual genes or sets of genes, all loci with estimated $d_N/d_S$ ratios greater than one in any *Silene* species were re-analyzed with the $d_N/d_S$ ratio constrained to a value of one for that species. Log-likelihood ratio tests were performed to compare the constrained and unconstrained analyses and determine whether the estimated $d_N/d_S$ ratios significantly exceed one (Yang 1998). Because we performed a total of 72 rate analyses in *Silene* protein genes (18 genes or gene sets for each of four *Silene* species), a Bonferroni correction factor of 72 was applied to all *p*-values from likelihood ratio tests to account for multiple comparisons.

**RNA editing**

In land plants, mitochondrial and plastid mRNA transcripts undergo systematic conversion of cytidines to uridines (C-to-U editing), restoring conserved codons (Knoop 2011). RNA editing sites were previously identified by cDNA sequencing for a subset of plastid genes in all four of the *Silene* species analyzed in this study (Sloan et al. 2010a). To predict editing sites in other plastid genes, we aligned *Silene* genes against all protein-coding sequences from *Arabidopsis thaliana*, *Nicotiana tabacum*, and *Zea mays* that are known to undergo RNA editing. Editing data for these three species were obtained from REDIdb (Picardi et al. 2007) and other published sources (Tillich et al. 2005; Chateigner-Boutin and Small 2007). Any site that is edited in one or more of these outgroups was predicted to be edited in *Silene* species that have a C at the corresponding genomic

position (Table S1). A number of editing sites appear to have been lost in one or more *Silene* species as a result of C-to-T substitutions at the genomic level. For any site that was predicted to vary in its editing status among the four *Silene* species, cDNA sequencing was performed as described previously (Sloan et al. 2010a) to confirm editing in at least one species. The results of cDNA sequencing confirmed editing in all cases except for *rps14* (nucleotide position 80). This site was predicted to be edited in *S. latifolia*, *S. vulgaris*, and *S. conica* but to have been lost by a genomic C-to-T substitution in *S. noctiflora*. However, cDNA sequencing in *S. latifolia* found no evidence of editing. Therefore, this site was excluded from the counts shown in Table 1 and Table S1.

ACKNOWLEDGEMENTS

We thank Dave McCauley for providing *S. vulgaris* tissue for DNA extraction and John Chuckalovcak for his 454 sequencing efforts. This study was supported by the NSF (DEB-0808452, MCB-1022128) and the Jefferson Scholars Foundation.

TABLES

**Table 1.** Summary of *Silene* plastid genomes.

| | ***Silene latifolia*** | ***Silene vulgaris*** | ***Silene noctiflora*** | ***Silene conica*** |
|---|---|---|---|---|
| **Genome Size (bp)** | 151,736 | 151,583 | 151,639 | 147,208 |
| Inverted Repeat | 25,906 | 26,008 | 29,891 | 26,858 |
| Large Single-Copy Region | 82,704 | 82,258 | 79,475 | 80,129 |
| Small Single-Copy Region | 17,220 | 17,309 | 12,382 | 13,363 |
| **G+C Content (%)** | 36.43 | 36.25 | 36.51 | 36.12 |
| **Protein Genes**[a] | 77 | 77 | 77 | 77 |
| **tRNA Genes**[a] | 30 | 30 | 30 | 30 |
| **rRNA Genes**[a] | 4 | 4 | 4 | 4 |
| **Introns**[a,b] | 20 | 20 | 16 | 16 |
| **RNA Editing Sites**[c] | 25 | 26 | 24 | 24 |

[a]Gene and intron counts exclude putative pseudogenes and duplicate copies in the IR.
[b]Intron counts include one *trans*-spliced intron in *rps12*
[c]Editing site counts include predicted sites that have not been confirmed by cDNA sequencing (see Methods)

**Table 2.** Patterns of organelle genome divergence in *S. noctiflora* and *S. conica*.

| | Mitochondrial | Plastid |
|---|---|---|
| **Sequence** | | |
| Major, genome-wide increase in synonymous substitution rate | Yes | No |
| Large increases in $d_N/d_S$ in a subset of protein genes | No | Yes |
| Large decrease in the frequency of RNA editing | Yes | No |
| **Structural** | | |
| Genomic expansion | Yes | No |
| Evolution of multichromosomal genome structure | Yes | No |
| Gene duplication | Yes | No |
| Elevated indel rate | Yes | Yes |
| Intron losses | Only One | Yes |
| Inversions | N/A[a] | Yes |
| Shifts in IR boundaries | N/A | Yes |

[a]The typical rates of genome rearrangement between and even within species are so high in angiosperm mitochondrial genomes that estimating the number or rate of inversions in any given lineage is not feasible.

**Table 3.** Positive selection on *Silene* plastid genes. All genes (or sets of concatenated genes belonging to a single complex) with estimated $d_N/d_S$ values greater than 1 are shown. Estimates that are significantly greater than 1 are shown in bold with Bonferroni-corrected *p*-values in parentheses. Values that are significant based on an uncorrected p-value of 0.05 but not after Bonferroni correction are marked with an asterisk.

| Gene/Complex | *Silene noctiflora* | *Silene conica* |
|---|---|---|
| *accD* | 2.20 | 0.98 |
| *cemA* | 1.21 | 1.48 |
| *clpP* | 1.19 | 1.31 |
| *rps* (concatenated) | **2.23** (0.002) | 1.17 |
| *ycf1* | 1.6* | **2.33** ($6 \times 10^{-6}$) |
| *ycf2* | **1.87** (0.02) | 1.39* |

**Table S1.** C-to-U RNA editing sites in *Silene* plastid genes. 'E' indicates that RNA editing has been confirmed by cDNA sequencing, while 'e' indicates that editing is predicted based on confirmed editing at the same site in other species. 'T' indicates that the editing site has been lost because of a C-to-T substitution at the genomic level.

| Gene | Position (nt)[a] | *S. latifolia* | *S. vulgaris* | *S. noctiflora* | *S. conica* |
|---|---|---|---|---|---|
| *atpA* | 795 | e | e | e | e |
| *clpP* | 559 | E | e | **T** | **T** |
| *matK* | 637 | e | e | e | e |
| *ndhA* | 341 | e | e | e | e |
| *ndhA* | 566 | e | e | e | e |
| *ndhA* | 1073 | e | e | e | e |
| *ndhB* | 149 | E | E | E | E |
| *ndhB* | 467 | E | E | E | E |
| *ndhB* | 586 | E | E | E | E |
| *ndhB* | 737 | E | E | E | E |
| *ndhB* | 746 | E | E | E | E |
| *ndhB* | 830 | E | E | E | E |
| *ndhB* | 836 | E | E | E | E |
| *ndhB* | 1481 | E | E | E | E |
| *ndhD* | 2 | e | e | e | e |
| *ndhD* | 383 | e | e | e | e |
| *ndhD* | 887 | e | e | e | e |
| *ndhG* | 50 | **T** | E | e | e |
| *petL* | 5 | e | e | e | e |
| *psbF* | 77 | E | **T** | **T** | **T** |
| *psbL* | 2 | E | E | E | **T** |
| *psbZ* | 50 | e | e | e | e |
| *rpoA* | 200 | **T** | E | e | e |
| *rpoB* | 473 | E | E | **T** | E |
| *rpoB* | 551 | E | E | E | E |
| *rpoB* | 566 | E | E | E | E |
| *rps2* | 248 | e | e | e | e |
| **Total Edits** | | 25 | 26 | 24 | 24 |

[a]Reported nucleotide position is based on *S. latifolia* gene sequence.

FIGURE LEGENDS

**Figure 1.** Plastid genome map for *Silene latifolia*. Boxes inside and outside the circle correspond to genes on the clockwise and anti-clockwise strand, respectively. The inner circle depicts GC content. The positions of the IR are labeled on the inner circle and noted with thicker black lines on the outer circle. All differences >100 bp in IR boundary positions among *Silene* species are labeled on the outer circle. Asterisks indicate genes that have lost introns in *S. noctiflora* and/or *S. conica*. Maps of the all four *Silene* plastid genomes are provided as supplementary material (Fig. S1). This figure was generated with OGDraw v1.2 (Lohse et al. 2007).

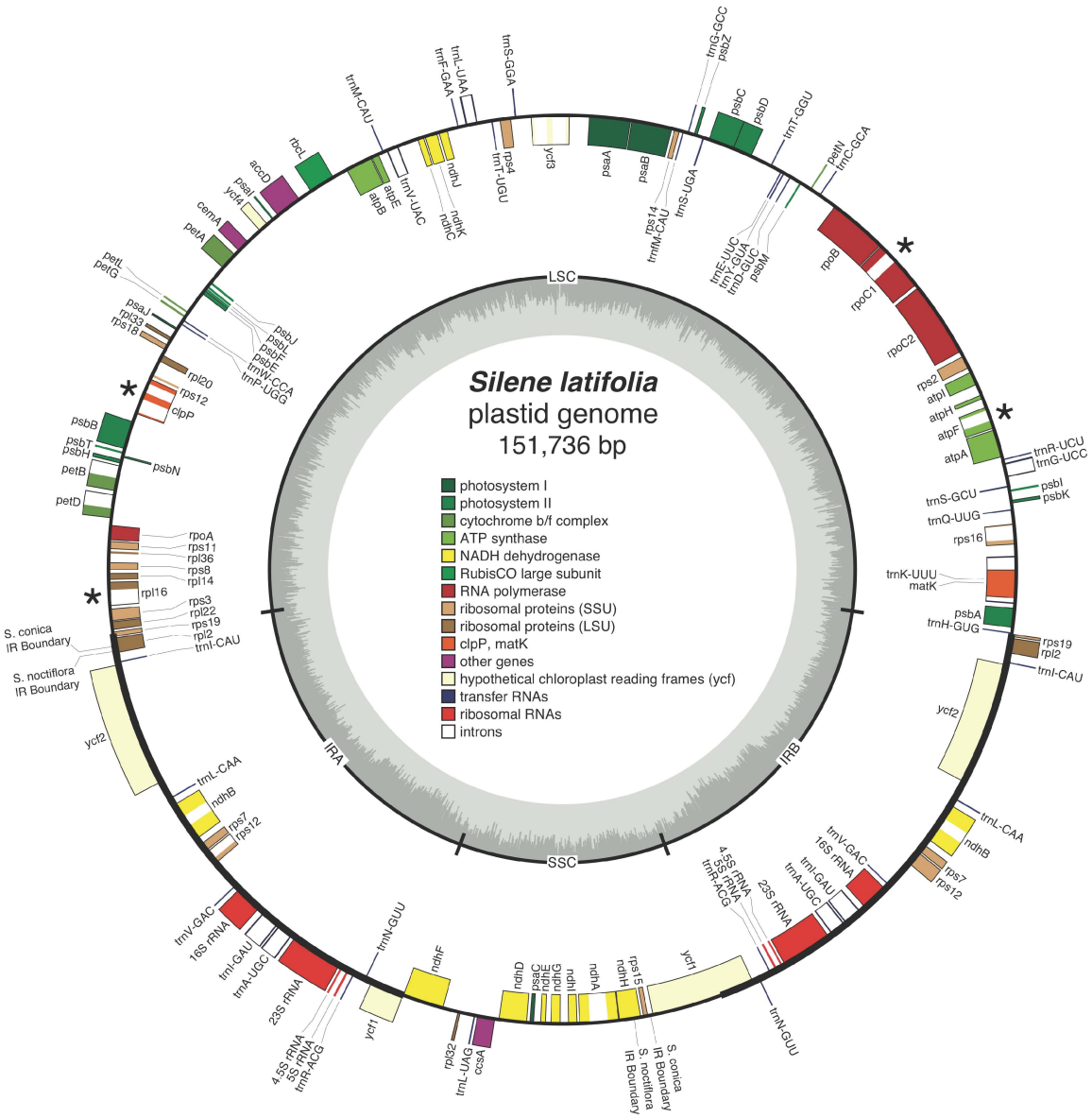

Silene latifolia
plastid genome
151,736 bp
photosystem I
photosystem II
cytochrome b/f complex
ATP synthase
NADH dehydrogenase
RubisCO large subunit
RNA polymerase
ribosomal proteins (SSU)
ribosomal proteins (LSU)
clpP, matK
other genes
hypothetical chloroplast reading frames (ycf)
transfer RNAs
ribosomal RNAs
introns
LSC
IRA
IRB
SSC
S. conica IR Boundary
S. noctiflora IR Boundary

**Figure 2.** Structural alignments of *Silene* and *Spinacia* plastid genomes. The coloring identifies collinear sequence blocks shared by all five genomes. Bars drawn below the black line indicate sequences found in inverted orientation. The height of each bar reflects sequence similarity. The eight inversion breakpoints identified by GRIMM are labeled below. Only a single copy of the inverted repeat was included in each genome, and the orientation of the small single-copy region was reversed relative to its conventional presentation to minimize complexities associated with changes in the IR boundaries. This figure was generated with MAUVE v2.3.1 (Darling et al. 2010).

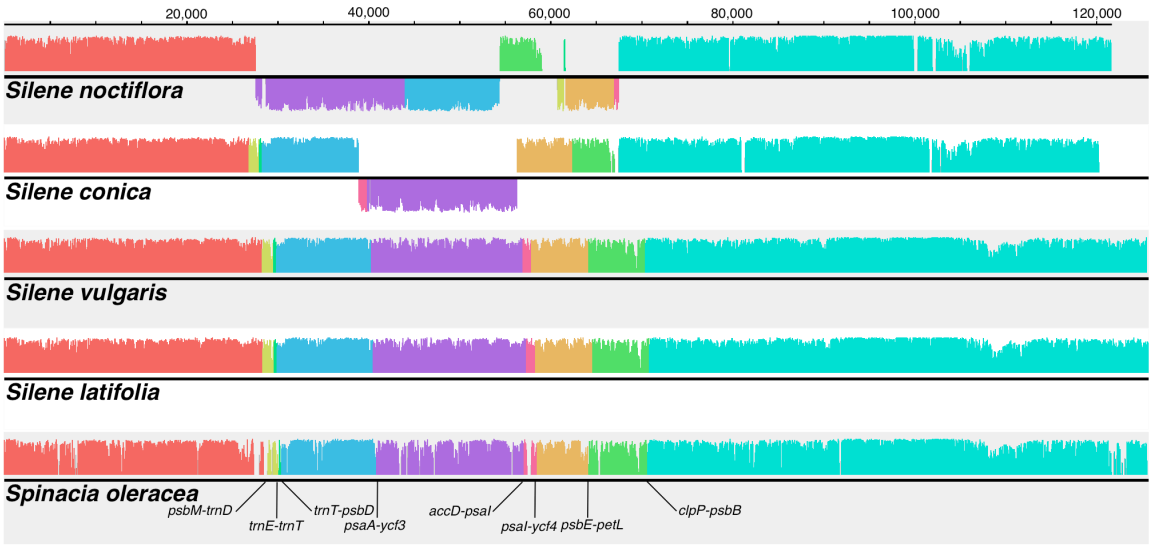
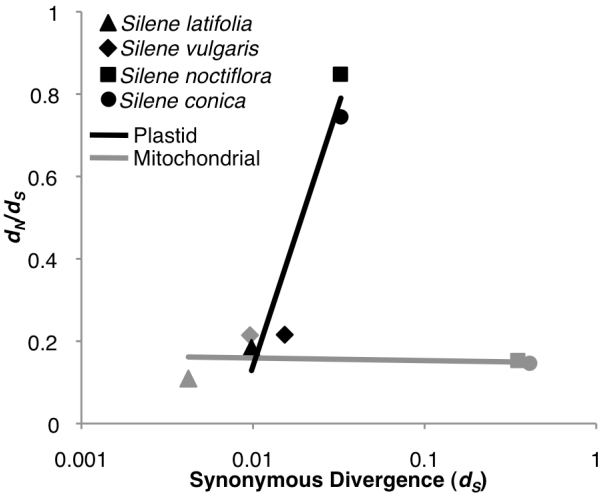


**Figure 3.** Synonymous ($d_S$) and non-synonymous divergence ($d_N$) in *Silene* mitochondrial and plastid genomes as estimated with PAML. Plastid data are based on an analysis of all protein genes except *accD*. Mitochondrial data are from Table 1 in chapter 7.



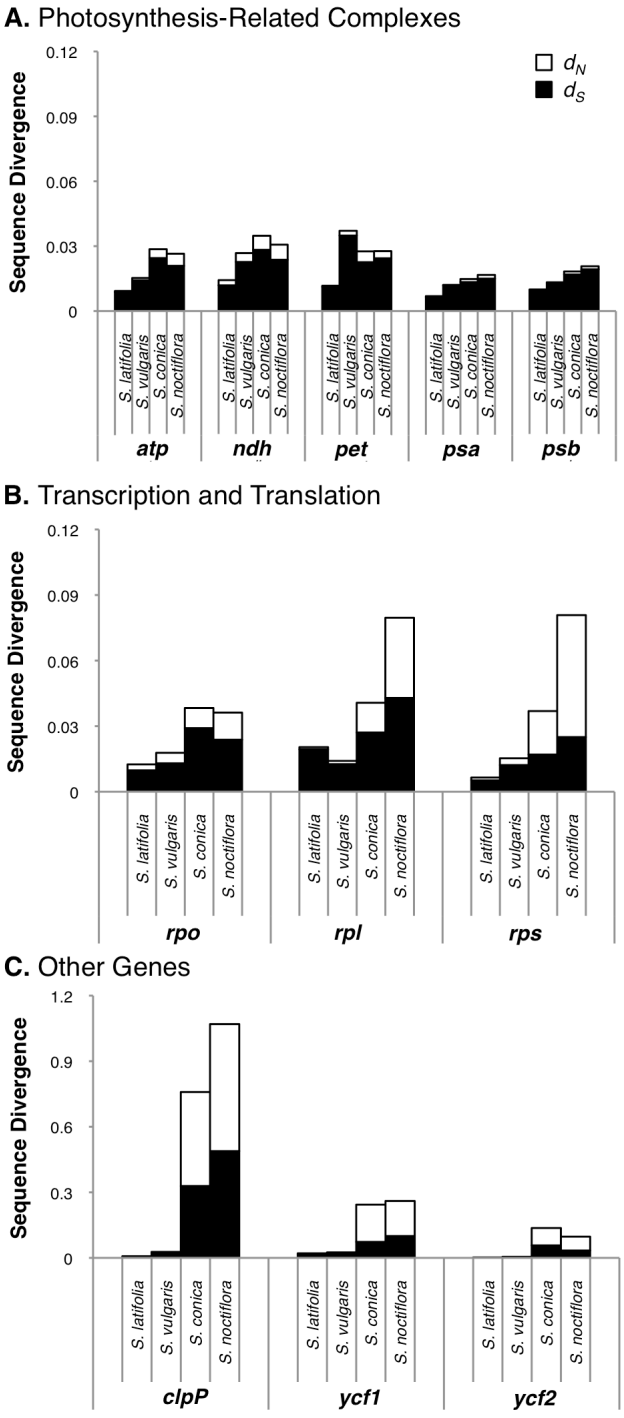



**Figure 4.** Synonymous and non-synonymous sequence divergence in *Silene* plastid genes as measured by the estimated number of substitutions per site in the terminal branch for each species. (A) Genes in major photosynthesis-related complexes. (B) Genes coding for RNA polymerase subunits and ribosomal proteins. (C) Three other highly divergent genes (note the ten-fold change in scale). Additional plots are available as supplementary material (Fig. S2) for individual protein genes, rRNA genes, and introns.

**Figure 5.** Phylogenetic relationships inferred from maximum likelihood analyses of a concatenation of all protein genes except the highly divergent *accD*, *clpP*, *ycf1*, and *ycf2* (A), a concatenation of protein genes in major photosynthesis-related complexes (B), and a concatenation of all shared *cis*-splicing introns (C). Branch lengths are defined in terms of substitutions per site. Values at each node indicate percentage support based on 1000 bootstrap replicates.

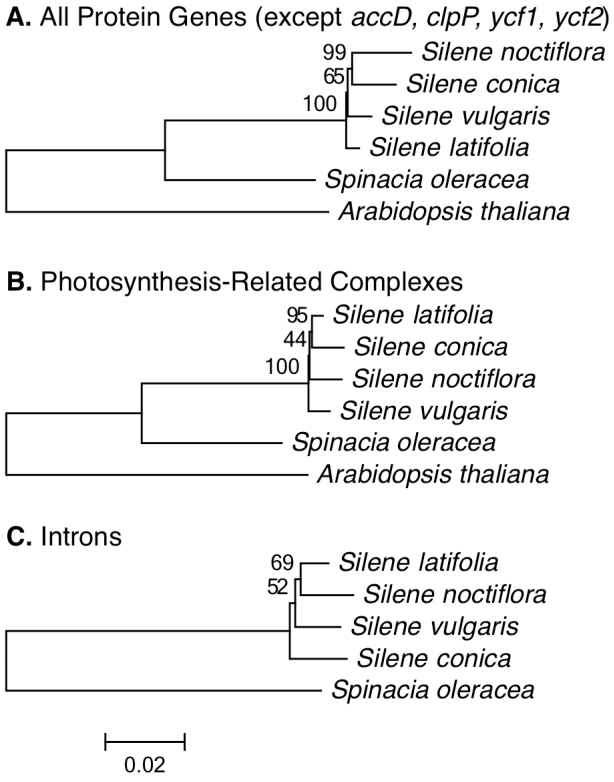


**Figure S1.** Plastid genome maps for all four *Silene* species. Boxes inside and outside the circle correspond to genes on the clockwise and anti-clockwise strand, respectively. The inner circle depicts GC content.

**Figure S1 A**

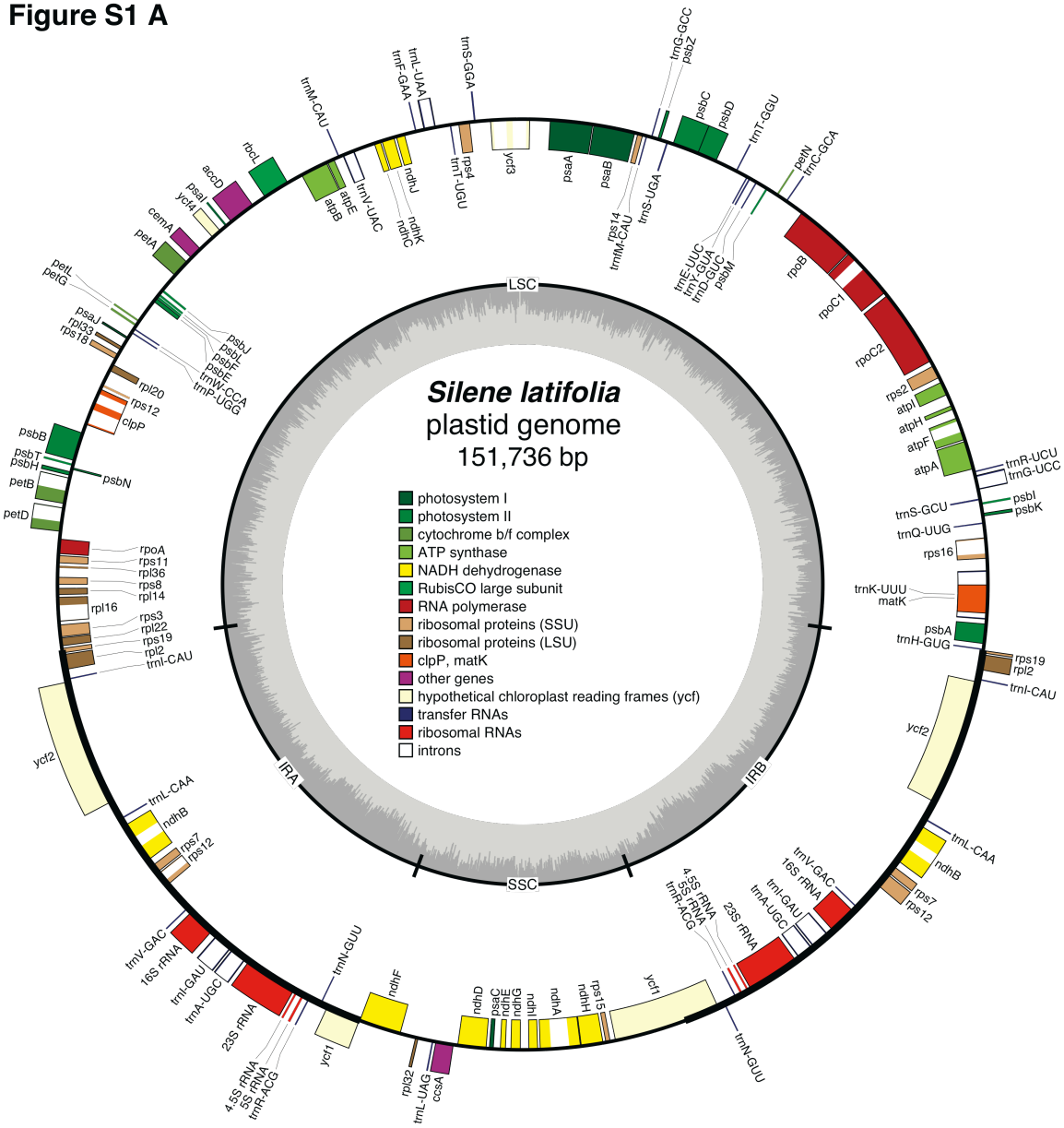


**Figure S1 B**

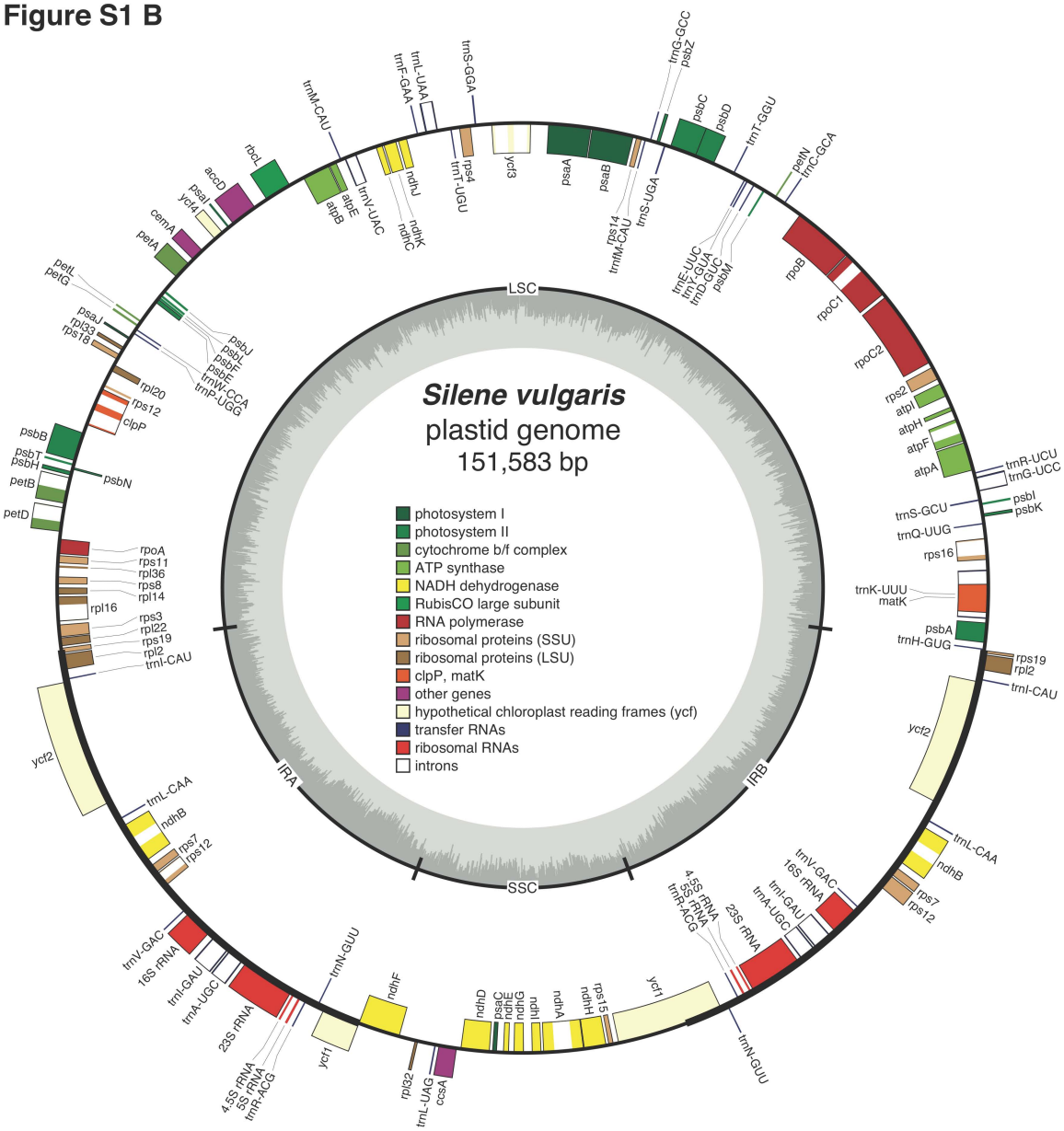

Silene vulgaris
plastid genome
151,583 bp
LSC
IRA
IRB
SSC
photosystem I
photosystem II
cytochrome b/f complex
ATP synthase
NADH dehydrogenase
RubisCO large subunit
RNA polymerase
ribosomal proteins (SSU)
ribosomal proteins (LSU)
clpP, matK
other genes
hypothetical chloroplast reading frames (ycf)
transfer RNAs
ribosomal RNAs
introns
trnG-GCC
psbZ
psbC
psbD
trnT-GGU
petN
trnC-GCA
rpoB
rpoC1
rpoC2
rps2
atpI
atpH
atpF
atpA
trnR-UCU
trnG-UCC
psbI
psbK
trnS-GCU
trnQ-UUG
rps16
trnK-UUU
matK
psbA
trnH-GUG
rps19
rpl2
trnI-CAU
ycf2
trnL-CAA
ndhB
rps7
rps12
trnV-GAC
16S rRNA
trnI-GAU
trnA-UGC
4.5S rRNA
5S rRNA
trnR-ACG
23S rRNA
trnN-GUU
ycf1
ndhF
rpl32
trnL-UAG
ccsA
ndhD
psaC
ndhE
ndhG
ndhI
ndhA
ndhH
rps15
trnM-CAU
atpE
atpB
rbcL
accD
psaI
ycf4
cemA
petA
petL
petG
psaJ
rpl33
rps18
rpl20
rps12
clpP
psbB
psbT
psbH
petB
petD
psbN
rpoA
rps11
rpl36
rps8
rpl14
rpl16
rps3
rpl22
rps19
rpl2
trnI-CAU
ycf2
trnL-CAA
ndhB
rps7
rps12
trnV-GAC
16S rRNA
trnI-GAU
trnA-UGC
23S rRNA
4.5S rRNA
5S rRNA
trnR-ACG
trnN-GUU
ycf1
rpl32
trnL-UAG
ccsA
ndhD
psaC
ndhE
ndhG
ndhI
ndhA
ndhH
rps15
ndhF
trnL-UAA
trnF-GAA
ndhJ
ndhK
ndhC
trnV-UAC
trnM-CAU
atpE
atpB
rbcL
accD
psaI
ycf4
cemA
petA
petL
petG
psaJ
rpl33
rps18
rpl20
rps12
clpP
psbB
psbT
psbH
petB
petD
psbN
psbJ
psbL
psbF
psbE
trnW-CCA
trnP-UGG
trnS-GGA
trnT-UGU
rps4
ycf3
psaA
psaB
rps14
trnM-CAU
trnS-UGA
trnE-UUC
trnY-GUA
trnD-GUC
psbM

**Figure S1 C**

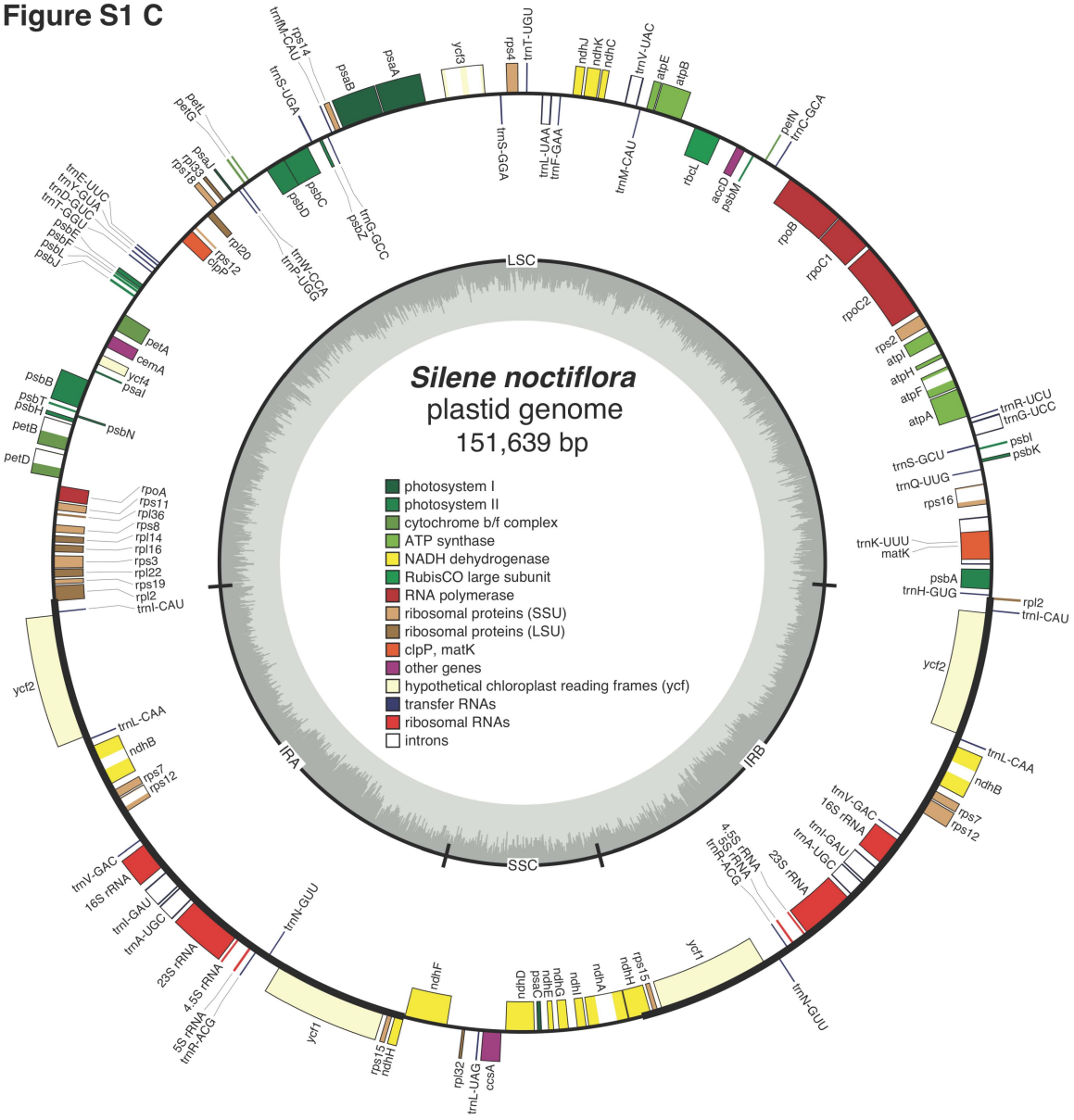

Silene noctiflora
plastid genome
151,639 bp
LSC
IRA
IRB
SSC
photosystem I
photosystem II
cytochrome b/f complex
ATP synthase
NADH dehydrogenase
RubisCO large subunit
RNA polymerase
ribosomal proteins (SSU)
ribosomal proteins (LSU)
clpP, matK
other genes
hypothetical chloroplast reading frames (ycf)
transfer RNAs
ribosomal RNAs
introns

**Figure S1 D**

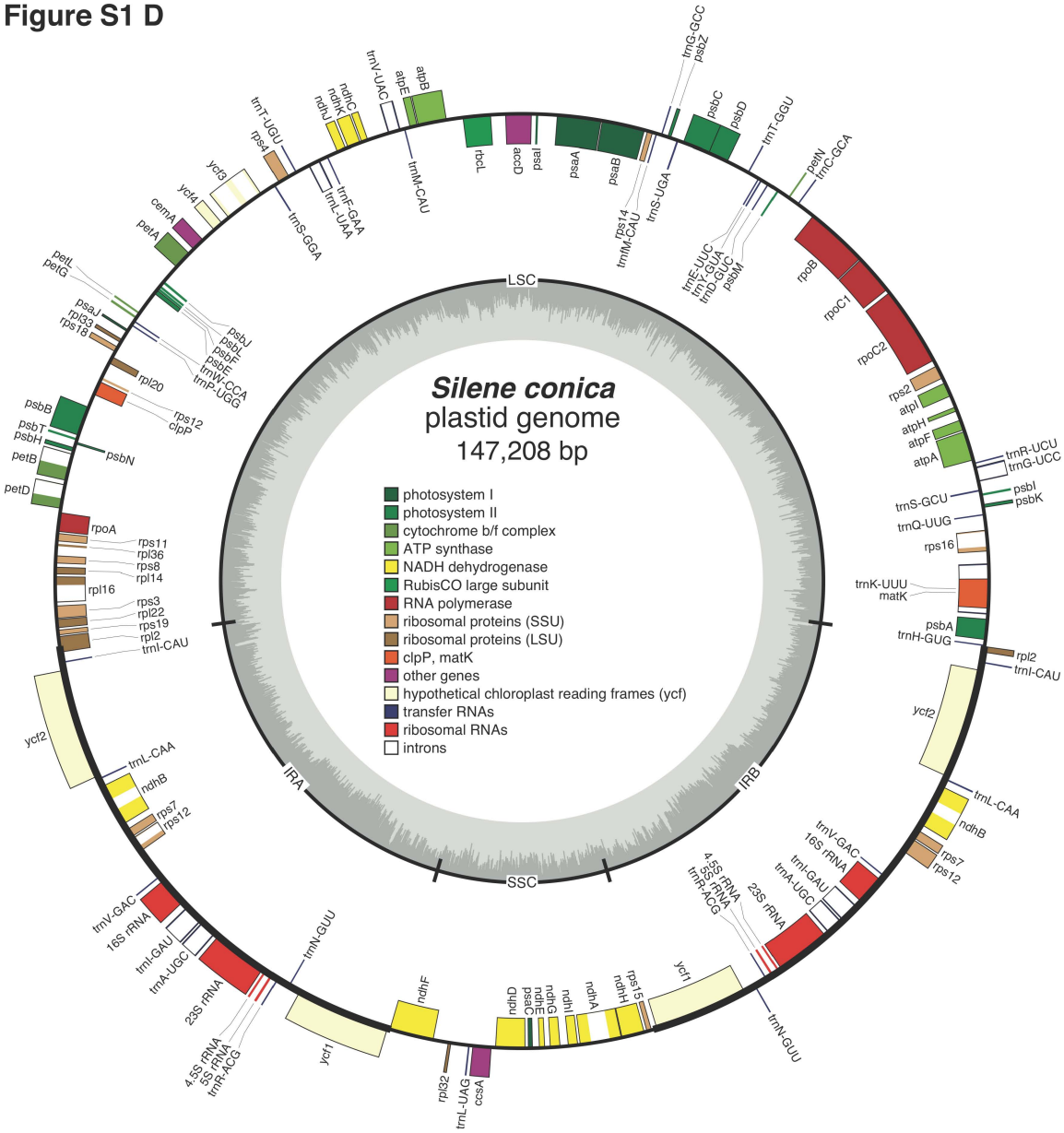
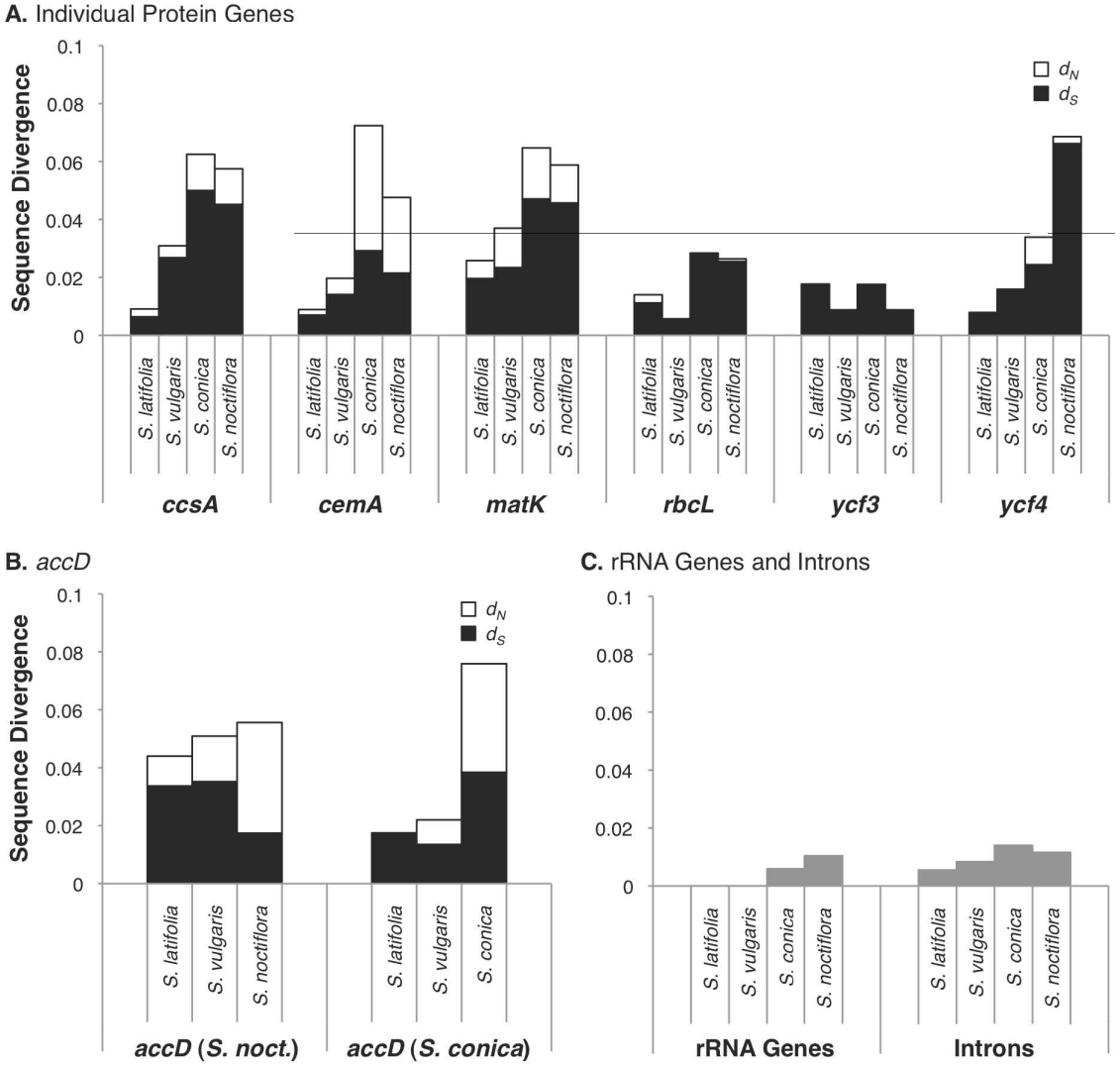

Silene conica
plastid genome
147,208 bp
LSC
IRA
IRB
SSC
photosystem I
photosystem II
cytochrome b/f complex
ATP synthase
NADH dehydrogenase
RubisCO large subunit
RNA polymerase
ribosomal proteins (SSU)
ribosomal proteins (LSU)
clpP, matK
other genes
hypothetical chloroplast reading frames (ycf)
transfer RNAs
ribosomal RNAs
introns

**Figure S2.** *Silene* sequence divergence as measured by the estimated number of substitutions per site in the terminal branch for each species based on individual protein genes (A), *accD* (B), and rRNA genes and introns (C). Protein genes were analyzed with codon-based models of evolution and estimates are provided for both synonymous ($d_S$) and non-synonymous ($d_N$) divergence. Introns and rRNA genes were analyzed with nucleotide-based models of evolution and are reported in terms of total number of substitutions per site. Large deletions in *accD* in both *S. noctiflora* and *S. conica* precluded performing a single analysis with both species, so separate analyses were run using the portions of the gene retained in each genome.


A. Individual Protein Genes
Sequence Divergence
0.1
0.08
0.06
0.04
0.02
0
$d_N$
$d_S$
S. latifolia
S. vulgaris
S. conica
S. noctiflora
ccsA
cemA
matK
rbcL
ycf3
ycf4
B. accD
accD (S. noct.)
accD (S. conica)
C. rRNA Genes and Introns
rRNA Genes
Introns

REFERENCES

Abdelnoor R. V., R. Yule, A. Elo, A. C. Christensen, G. Meyer-Gauen, and S. A. Mackenzie. 2003. Substoichiometric shifting in the plant mitochondrial genome is influenced by a gene homologous to MutS. Proc Natl Acad Sci 100:5968-5973.

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

Alverson A. J., S. Zhuo, D. W. Rice, D. B. Sloan, and J. D. Palmer. 2011. The mitochondrial genome of the legume *Vigna radiata* and the analysis of recombination across short mitochondrial repeats. PLoS One 6:e16404.

Andersson S. G., and C. G. Kurland. 1998. Reductive evolution of resident genomes. Trends Microbiol. 6:263-268.

Arrieta-Montiel M. P., V. Shedge, J. Davila, A. C. Christensen, and S. A. Mackenzie. 2009. Diversity of the Arabidopsis mitochondrial genome occurs via nuclear-controlled recombination activity. Genetics 183:1261-1268.

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

Barbrook A. C., C. J. Howe, and S. Purton. 2006. Why are plastid genomes retained in non-photosynthetic organisms? Trends Plant Sci. 11:101-108.

Bell C. D., D. E. Soltis, and P. S. Soltis. 2010. The age and diversification of the angiosperms re-revisited. Am. J. Bot. 97:1296-1303.

Benning C., C. Xu, and K. Awai. 2006. Non-vesicular and vesicular lipid trafficking involving plastids. Curr. Opin. Plant Biol. 9:241-247.

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

Bubunenko M. G., J. Schmidt, and A. R. Subramanian. 1994. Protein substitution in chloroplast ribosome evolution. A eukaryotic cytosolic protein has replaced its organelle homologue (L23) in spinach. J. Mol. Biol. 240:28-41.

Campagna M. L., and S. R. Downie. 1998. The intron in chloroplast gene *rpl16* is missing from the flowering plant families Geraniaceae, Goodeniaceae and Plumbaginaceae. Trans. Illin. Acad. Sci. 91:1-11.

Cappadocia L., A. Marechal, J. S. Parent, E. Lepage, J. Sygusch, and N. Brisson. 2010. Crystal structures of DNA-Whirly complexes and their role in *Arabidopsis* organelle genome repair. Plant Cell 22:1849-1867.

Carrie C., E. Giraud, and J. Whelan. 2009. Protein transport in organelles: Dual targeting of proteins to mitochondria and chloroplasts. FEBS J. 276:1187-1195.

Chateigner-Boutin A. L., and I. Small. 2007. A rapid high-throughput method for the detection and quantification of RNA editing based on high-resolution melting of amplicons. Nucleic Acids Res. 35:e114.

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

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

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

Day A., and P. Madesis. 2007. DNA replication, recombination, and repair in plastids. Pp. 65-119 *in* R. Bock, ed. Cell and Molecular Biology of Plastids. Springer, .

Delannoy E., S. Fujii, C. C. des Francs, M. Brundrett, and I. Small. In Press. Rampant gene loss in the underground orchid *Rhizanthella gardneri* highlights evolutionary constraints on plastid genomes. Mol. Biol. Evol. .

Downie S. R., E. Llanas, and D. S. Katz-Downie. 1996. Multiple independent losses of the *rpoC1* intron in angiosperm chloroplast DNA's. Syst. Bot. 21:135-151.

Downie S. R., R. G. Olmstead, G. Zurawski, D. E. Soltis, P. S. Soltis, J. C. Watson, and J. D. Palmer. 1991. Six independent losses of the chloroplast DNA *rpl2* intron in dicotyledons: molecular and phylogenetic implications. Evolution 45:1245-1259.

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

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

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

Erixon P., and B. Oxelman. 2008a. Reticulate or tree-like chloroplast DNA evolution in Sileneae (Caryophyllaceae)? Mol. Phylogenet. Evol. 48:313-325.

--- 2008b. Whole-gene positive selection, elevated synonymous substitution rates, duplication, and indel evolution of the chloroplast clpP1 gene. PLoS ONE 3:e1386.

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

Gould S. B., R. F. Waller, and G. I. McFadden. 2008. Plastid evolution. Annu. Rev. Plant. Biol. 59:491-517.

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

Greiner S., X. Wang, R. G. Herrmann, U. Rauwolf, K. Mayer, G. Haberer, and J. Meurer. 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., J. V. Kuehl, J. L. Boore, and R. K. Jansen. 2008. Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. Proc. Natl. Acad. Sci. 105:18424-18429.

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

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

Jansen R. K., Z. Cai, L. A. Raubeson, and H. Daniell. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc. Natl. Acad. Sci. 104:19369.

Jansen R. K., M. F. Wojciechowski, E. Sanniyasi, S. B. Lee, and H. Daniell. 2008. Complete plastid genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic distribution of *rps12* and *clpP* intron losses among legumes (Leguminosae). Mol. Phylogenet. Evol. 48:1204-1217.

Jansen R. K., L. A. Raubeson, J. L. Boore, C. W. dePamphilis, T. W. Chumley, R. C. Haberle, S. K. Wyman, A. J. Alverson, R. Peery, S. J. Herman, H. M. Fourcade, J. V. Kuehl, J. R. McNeal, J. Leebens-Mack, and L. Cui. 2005. Methods for obtaining and analyzing whole chloroplast genome sequences. Methods Enzymol. 395:348-384.

Katayama H., and Y. Ogihara. 1996. Phylogenetic affinities of the grasses to other monocots as revealed by molecular analysis of chloroplast DNA. Curr. Genet. 29:572-581.

Kim E., and J. Archibald. 2009. Diversity and evolution of plastids and their genomes. Pp. 1-39 *in* H. Aronsson and A. S. Sandelius, eds. *The Chloroplast—interactions with the environment*. Springer-Verlag, Berlin.

Knoop V. 2011. When you can't trust the DNA: RNA editing changes transcript sequences. Cell. Mol. Life Sci. 68:567-586.

Knox E. B., and J. D. Palmer. 1999. The chloroplast genome arrangement of *Lobelia thuliniana* (Lobeliaceae): Expansion of the inverted repeat in an ancestor of the Campanulales. Plant Syst. Evol. 214:49-64.

Knox E. B., S. R. Downie, and J. D. Palmer. 1993. Chloroplast genome rearrangements and the evolution of giant Lobelias from herbaceous ancestors. Mol. Biol. Evol. 10:414-430.

Kode V., E. A. Mudd, S. Iamtham, and A. Day. 2005. The tobacco plastid *accD* gene is essential and is required for leaf development. Plant J. 44:237-244.

Koller B., H. Fromm, E. Galun, and M. Edelman. 1987. Evidence for in vivo trans splicing of pre-mRNAs in tobacco chloroplasts. Cell 48:111-119.

Kuroda H., and P. Maliga. 2003. The plastid clpP1 protease gene is essential for plant development. Nature 425:86-89.

Li R., C. Yu, Y. Li, T. W. Lam, S. M. Yiu, K. Kristiansen, and J. Wang. 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966-1967.

Logacheva M. D., T. H. Samigullin, A. Dhingra, and A. A. Penin. 2008. Comparative chloroplast genomics and phylogenetics of *Fagopyrum esculentum* ssp. *ancestrale*–A wild ancestor of cultivated buckwheat. BMC Plant Biol. 8:59.

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

Magee A. M., S. Aspinall, D. W. Rice, B. P. Cusack, M. Sémon, A. S. Perry, S. Stefanović, D. Milbourne, S. Barth, J. D. Palmer, J. C. Gray, T. A. Kavanagh, and K. H. Wolfe. 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Res. 20:1700-1710.

Martinez-Zapater J. M., P. Gil, J. Capel, and C. R. Somerville. 1992. Mutations at the *Arabidopsis CHM* locus promote rearrangements of the mitochondrial genome. Plant Cell 4:889-899.

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

Millen R. S., R. G. Olmstead, K. L. Adams, J. D. Palmer, N. T. Lao, L. Heggie, T. A. Kavanagh, J. M. Hibberd, J. C. Gray, C. W. Morden, P. J. Calie, L. S. Jermiin, and K. H. Wolfe. 2001. Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. Plant Cell 13:645-658.

Moran N. A., and J. J. Wernegreen. 2000. Lifestyle evolution in symbiotic bacteria: insights from genomics. Trends Ecol. Evol. 15:321-326.

Mower J. P., D. B. Sloan, and A. J. Alverson. In press. Plant mitochondrial diversity – the genomics revolution. *in* J. F. Wendel, ed. Plant Genome Diversity. Springer, .

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

Ohta T. 1992. The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. 23:263-286.

Palmer J. D. 1986. Isolation and structural analysis of chloroplast DNA. Meth. Enzymol. 118:167-186.

Palmer J. D., and L. A. Herbon. 1988. Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J. Mol. Evol. 28:87-97.

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

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

Picardi E., T. M. Regina, A. Brennicke, and C. Quagliariello. 2007. REDIdb: the RNA editing database. Nucleic Acids Res. 35:D173-7.

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

Raubeson L. A., and R. K. Jansen. 2005. Chloroplast genomes of plants. Pp. 45-68 *in* R. J. Henry, ed. Plant diversity and evolution: genotypic and phenotypic variation in higher plants. CABI, Wallingford, UK.

Redei G. P. 1973. Extra-chromosomal mutability determined by a nuclear gene locus in *Arabidopsis*. Mutation Res. 18:149-162.

Roussell D. L., D. L. Thompson, S. G. Pallardy, D. Miles, and K. J. Newton. 1991. Chloroplast structure and function is altered in the NCS2 maize mitochondrial mutant. Plant Physiol. 96:232-238.

Rowan B. A., D. J. Oldenburg, and A. J. Bendich. 2010. RecA maintains the integrity of chloroplast DNA molecules in *Arabidopsis*. J. Exp. Bot. 61:2575-2588.

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

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

Shikanai T., K. Shimizu, K. Ueda, Y. Nishimura, T. Kuroiwa, and T. Hashimoto. 2001. The chloroplast *clpP* gene, encoding a proteolytic subunit of ATP-dependent protease, is indispensable for chloroplast development in tobacco. Plant Cell Physiol. 42:264-273.

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

Sloan D. B., A. H. MacQueen, A. J. Alverson, J. D. Palmer, and D. R. Taylor. 2010a. 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., A. J. Alverson, H. Storchova, J. D. Palmer, and D. R. Taylor. 2010b. Extensive loss of translational genes in the structurally dynamic mitochondrial genome of the angiosperm *Silene latifolia*. BMC Evol. Biol. 10:274.

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

Smith S. A., and M. J. Donoghue. 2008. Rates of molecular evolution are linked to life history in flowering plants. Science 322:86-89.

Stanne T. M., L. L. Sjogren, S. Koussevitzky, and A. K. Clarke. 2009. Identification of new protein substrates for the chloroplast ATP-dependent Clp protease supports its constitutive role in *Arabidopsis*. Biochem. J. 417:257-268.

Tesler G. 2002. GRIMM: genome rearrangements web server. Bioinformatics 18:492-493.

Tillich M., H. T. Funk, C. Schmitz-Linneweber, P. Poltnigg, B. Sabater, M. Martin, and R. M. Maier. 2005. Editing of plastid RNA in Arabidopsis thaliana ecotypes. Plant J. 43:708-715.

Timmis J. N., M. A. Ayliffe, C. Y. Huang, and W. Martin. 2004. Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet. 5:123-135.

Whittle C. A., and M. O. Johnston. 2002. Male-driven evolution of mitochondrial and chloroplastidial DNA sequences in plants. Mol. Biol. Evol. 19:938-949.

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

Wolfe K. H., C. W. Morden, S. C. Ems, and J. D. Palmer. 1992. Rapid evolution of the plastid translational apparatus in a nonphotosynthetic plant: loss or accelerated sequence evolution of tRNA and ribosomal protein genes. J. Mol. Evol. 35:304-317.

Woodson J. D., and J. Chory. 2008. Coordination of gene expression between organellar and nuclear genomes. Nat. Rev. Genet. 9:383-395.

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

Wyman S. K., R. K. Jansen, and J. L. Boore. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252-3255.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24:1586-1591.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15:568-573.

Yoshida K., and K. Noguchi. 2011. Interaction between chloroplasts and mitochondria: activity, function, and regulation of the mitochondrial respiratory system during photosynthesis. Pp. 383-409 *in* F. Kempken, ed. Plant Mitochondria. Springer.

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

Zurawski G., and M. T. Clegg. 1987. Evolution of higher-plant chloroplast DNA-encoded genes: implications for structure-function and phylogenetic studies. Annu. Rev. Plant. Phys. 38:391-418.