DISSERTATION

THE PLASTID CASEINOLYTIC PROTEASE COMPLEX AS A MODEL FOR CYTONUCLEAR COEVOLUTION

Submitted by

Alissa Marie Williams

Graduate Degree Program in Cell and Molecular Biology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2021

Doctoral Committee:

Advisor: Daniel Sloan

Patricia Bedinger
Rachel Mueller
Marinus Pilon
Mark Stenglein

ABSTRACT

THE PLASTID CASEINOLYTIC PROTEASE COMPLEX AS A MODEL FOR CYTONUCLEAR COEVOLUTION

Coevolution, or evolution in response to reciprocal selective pressures, is important to biological function and the persistence of populations. Competition or mutualisms between organisms can drive coevolution, as can predatory or parasitic relationships. However, coevolution also occurs within cells, as coevolution can result from the interactions between proteins within complexes as well as between the multiple genomes within eukaryotic cells.

In protein complexes, subunits must bind tightly and specifically to one another. Changes in one protein subunit are often correlated with changes in the other subunits to preserve the functionality of the complex. Thus, in many protein complexes, correlated rates of evolution are found between the sequences of component subunits. This covariation is strong enough to be used as a method to predict which proteins are connected physically and/or functionally.

The coevolution between multiple genomes in eukaryotic cells is known as cytonuclear coevolution. Plants, for example, have a nuclear genome and two cytoplasmic genomes found in the plastid (chloroplast) and mitochondrion. Many protein complexes within these organelles consist of subunits deriving from both the nucleus and the organelle itself. Since the nuclear genome and organelle genomes differ in modes of transmission, mutation rates, and selective pressures, partnerships between proteins originating from two cellular compartments are great models for understanding protein complex evolution.

Protein complexes are frequently shaped via gene duplication. Many protein complexes contain paralogous proteins at their cores; the duplication of a self-binding protein leads to dimerization of the

paralogous proteins and subsequent recruitment of additional subunits. Gene duplication after establishment of a heteromeric complex allows subunits to specialize.

The plastid caseinolytic protease (Clp) complex provides a model system for studying protein complex evolution, in the context of cytonuclear interactions, gene duplication, and evolutionary rate variation. This complex is highly conserved across bacteria and consists of adaptors, chaperones, and a proteolytic core. It is present in both plastids and mitochondria because these organelles are derived from ancient bacterial endosymbionts. The Clp core contains 14 subunits; in mitochondria and most bacteria, all 14 subunits are encoded by the same gene. However, in the cyanobacterial and plastid lineage, multiple rounds of gene duplication have led to a core encoded by nine different genes in the model plant species *Arabidopsis thaliana*. Further, only one of these plastid Clp core subunit genes is encoded by the plastid itself—the remaining eight are encoded by the nucleus, the result of gene transfers from the organelle to the nucleus early in the history of green plants. In addition to representing multiple rounds of gene duplication, the plastid Clp core also demonstrates extreme rate variation across green plants. The plastid-encoded subunit (ClpP1) is typically highly conserved across species. However, in some species, ClpP1 is one of the most rapidly evolving genes across all three genomes.

In this dissertation, I use these features of the plastid Clp complex to shed light on protein complex evolution in various contexts. After a general introduction to the field in Chapter 1, Chapter 2 focuses on the evolutionary history of ClpP1, looking at rate variation and the loss of introns, RNA editing sites, and catalytic sites across green plants. Through mass spectrometry, I determine that ClpP1 is still a functional protein in *Silene noctiflora*, which has one of the most divergent plastid Clp complexes known. This work also includes an evolutionary rate covariation analysis between ClpP1 and the nuclear-encoded Clp core genes. Chapter 3 provides genomic resources, including a high-quality, long-read transcriptome, for *S. noctiflora*, which is a species of interest for the reason outlined above. Analysis of the transcriptome revealed a triplication of one of the nuclear-encoded Clp core genes in this species. Chapter 4 discusses the recent duplication history of the nuclear-encoded Clp core genes across a broad range of flowering plants. I use these data to examine and characterize post-duplication evolutionary fates

of paralogs. These analyses are extended to another plastid complex, acetyl-CoA carboxylase (ACCase). Taken together, these chapters elucidate various features of plastid Clp complex evolution as well as provide insight into the possible causes and consequences of rate variation and gene duplication in the coevolution of protein complex subunits.

ACKNOWLEDGEMENTS

Work in this dissertation was supported by National Science Foundation (NSF) grants MCB-1733227 and NSF MCB-1614629, start-up funds from Colorado State University, the Wolves to Rams undergraduate research program (NSF Grant Numbers 1930150 and 19300092, NIH Grant Number 1T34GM137861-01), and graduate fellowships from NSF (DGE-1321845), the National Institutes of Health (T32-GM132057), and the Department of Education (GAANN).

First, I would like to thank my scientific community at CSU. I have been surrounded by great scientists who also happen to be great people. I have really enjoyed lab meetings, journal clubs, and impromptu conversations with everyone, particularly my lab mates and committee members. So much of the graduate school experience is dependent on the adviser and I feel incredibly lucky to have had such an amazing one. Dan, thank you for being an insightful, supportive, dedicated, and overall wonderful adviser. You have helped me grow as a scientist while also knowing and appreciating me as a person. You are both one of the most brilliant scientists I know and one of the nicest scientists I know. To my committee: Pat—I've always loved scientific conversations with you because you are always excited about my research (and that of everyone else). I appreciate your ideas and excitement. I'm glad that I got to work in your lab for a rotation. Rachel—you have an outstanding ability to ask incredibly good questions without making the presenter feel overwhelmed. I appreciate all of your feedback from all of the lab meetings I've given over the years. Rien—you have been a vital part of my committee as a resident biochemist and plastid expert. Your questions are perceptive and keep me on my toes, which is always a good thing. Mark—I really enjoyed my rotation with you and I'm glad that you've been a part of my science since then. You are one of the best bioinformaticians I know. Hanging out with you and Dan in Todos Santos was a highlight of my graduate career. To my lab mates: thank you all for the intellectual stimulation and conversations as well as your friendship. I have loved being a part of the lab and the various lab activities and parties. To Amanda Broz—thank you for all of your mentorship at the bench. I couldn't have done it without you. And to Johanna Michelsohn, Olivia Carter, Greg Noe, and Hannah

Mendoza—thank you for being awesome undergraduate researchers. Being a mentor and working with you was one of my favorite experiences at CSU.

I also want to thank all of the other people I've met as a student here, particularly my IM sports teammates. Starting an IM team was one of the best things I did as a graduate student. I loved having a group of like-minded people to meet with every week as we calmly or not-so-calmly participated in sports. I've also really enjoyed the city's kickball league and I looked forward to it every summer.

Next, I'd like to thank the people who got me to this point in my life. Thanks to my parents (Bart and Wendy Williams), siblings (Kaitlin and Josh Williams plus my brother-in-law Nihal Studden), and extended family for all of the support throughout the years. Thanks also to my in-laws (Ray and Lauren Whitaker) for their encouragement. I had a great set of teachers starting from elementary school through high school and awesome professors at Wofford College as an undergraduate. My educational path and achievements would not have been possible without them.

Finally, thank you to my friends. I am lucky to have great friends in my life both near and far. Thank you to my friends from childhood, Wofford, and CSU for enriching my life. I have enjoyed phone calls, video calls, hikes, movie nights, etc. with many including Emily Bell, Kristina Lee, Katie Krebs, Kim Alexander, Amy Danson, Julia Nguyen, and others. I have also enjoyed spending time with my cat, Gordon. Special thanks to two friends in particular: Sean Freeman—thank you for always being willing to help me at the drop of a hat. You are an incredibly kind person as well as a perceptive one. Thank you for your insights, perspective, and camaraderie (particularly regarding graduate school things). Tyler Slonecki—I am incredibly grateful for a decade of friendship. Thank you for your support and steadiness whenever I've needed it. Despite living far apart for the last six years, we've only gotten closer. Thanks for being an awesome friend.

My last thank you is to my very best friend and husband, Justin Whitaker. Justin—thank you for being a wonderful partner and friend. I feel very fortunate that we happened to end up at the same undergraduate institution, where we immediately clicked as friends, and that we developed a strong friendship for years before we began dating. When I wanted to start somewhere new for graduate school,

I chose to come to CSU because you were here, which was probably the best decision I've made, both personally and professionally. I've loved my time here as a graduate student for many reasons, but especially because of you. We will always have many great memories in Colorado—we got engaged and married here and got to see a lot of cool places. You've been with me through good times and trying times. I appreciate your compassion, dedication, wit, humor, and intelligence. Thanks for being the best teammate I could have ever asked for. I will miss Colorado, but I look forward to our new life in Tennessee and everything yet to come—particularly meeting our first child in a few months.

TABLE OF CONTENTS

CHAPTER 1: AN INTRODUCTION TO MOLECULAR COEVOLUTION

The term "coevolution" was defined in the 1950s and 1960s; the first two major instances of the term were used to describe host-parasite interactions and reciprocal evolution between butterflies and their food plants, respectively (Ehrlich and Raven, 1964; Mode, 1958). Since then, it has become an important area of study from the population level to the cellular and molecular levels (Atchley et al., 2000; Codoñer and Fares, 2008; Juan et al., 2013; Lovell and Robertson, 2010; Thompson, 2009, 1989; Thompson and Burdon, 1992). Intracellular coevolution exists in multiple forms, include protein-protein coevolution and cytonuclear coevolution.

A main source of protein-protein interactions in the cell are protein complexes, which consist of physically associated protein subunits (Jones and Thornton, 1996; Marsh and Teichmann, 2015; Nooren and Thornton, 2003). Complexes can be homomeric or heteromeric; homomeric complexes are formed by identical subunits encoded by the same gene while heteromeric complexes are formed by different subunits (Jones and Thornton, 1996; Marsh and Teichmann, 2015; Nooren and Thornton, 2003). Frequently, heteromeric complexes are evolutionarily derived from homomers—often homodimers, which are formed by two self-interacting proteins. (Andreeva and Murzin, 2006; Finnigan et al., 2012; Ispolatov et al., 2005; Lynch, 2012; Pereira-Leal et al., 2007; Yosef et al., 2009). Upon duplication of the gene encoding the homodimer, the products of the paralogous genes may form heterodimers. Selective pressures on each paralog can preserve the original self-binding site, leading to the formation of heterodimers while allowing for divergence in other protein domains. Thus, over evolutionary time, the paralogous proteins may recruit different sets of additional proteins to the complex while maintaining their ability to bind to one another, leading to large heteromeric complexes (Andreeva and Murzin, 2006; Ispolatov et al., 2005; Lynch, 2012; Pereira-Leal et al., 2007; Yosef et al., 2009). The asymmetry acquired via paralogous proteins in a complex can also occur in a nonadaptive manner, for instance through drift or mutation accumulation (Hochberg et al., 2020; Lynch, 2012).

The genes encoding interacting members of heteromeric complexes often demonstrate evolutionary rate covariation (ERC), where rates of evolution are correlated between different components (Clark et al., 2012; Clark and Aquadro, 2010; Forsythe et al., 2021; Goh et al., 2000; Juan et al., 2013; Ramani and Marcotte, 2003; Sato et al., 2005; Williams et al., 2019). These correlated changes often allow function to be maintained, whether they occur through compensation, antagonistic interactions, or general selective pressures acting on the entire complex. ERC analyses are used in two directions—we can look for rate correlations between known interacting partners to test for the evolutionary consequences of molecular interactions or use rate covariation to predict which proteins interact with one another (Clark et al., 2012; Clark and Aquadro, 2010; Findlay et al., 2014; Forsythe et al., 2021; Goh et al., 2000; Juan et al., 2013; Ramani and Marcotte, 2003; Raza et al., 2019; Sato et al., 2005). These methods have been applied in many contexts, including cytonuclear coevolution, or the coevolution between the nuclear genome and the cytoplasmic genomes (plastids and mitochondria) (Barreto and Burton, 2013; Beck et al., 2015; Forsythe et al., 2021; Gong et al., 2014; Havird et al., 2017, 2015; Osada and Akashi, 2012; Rockenbach et al., 2016; Sloan et al., 2014a, 2014b; Weng et al., 2016; Williams et al., 2019; Yan et al., 2018; Zhang et al., 2016, 2015).

Cytonuclear interactions provide a particularly useful case study for evolutionary rate covariation due to the many differences between the nuclear and cytoplasmic genomes. While the nuclear genome is inherited equally from each parent, plastids and mitochondria are almost exclusively inherited maternally, meaning that they are essentially inherited in an asexual manner (Birky, 1995; Caspari, 1948; Mogensen, 1996; Murlas Cosmides and Tooby, 1981). Therefore, uniparental inheritance changes the selective pressures that act on the cytoplasmic genomes (Birky, 1995). Further, the plastid and mitochondrial genomes have a different set of error checking mechanisms than the nucleus, leading to differences in mutation rates; in plants, the cytoplasmic genomes generally have lower rates of mutation and gene sequence evolution than the nuclear genome (Drouin et al., 2008; Palmer and Herbon, 1988; Smith and Keeling, 2015; Wolfe et al., 1987). However, there are exceptions to this trend—some flowering plant species have greatly accelerated evolutionary rates across the entire mitochondrial genome and/or in a

subset of plastid-encoded genes, including ClpP1, which is part of the plastid caseinolytic protease (Clp) complex (Barnard-Kubow et al., 2014; Erixon and Oxelman, 2008; Guisinger et al., 2008; Haberle et al., 2008; Jansen et al., 2007; Knox, 2014; Magee et al., 2010; Park et al., 2017; Rockenbach et al., 2016; Sloan et al., 2014a, 2012b, 2012a; Weng et al., 2016; Williams et al., 2019, 2015).

The plastid Clp complex is essential in plants; it is one of the most abundant proteases in the stroma and degrades a variety of targets (Apitz et al., 2016; Bouchnak and van Wijk, 2021; Majeran et al., 2000; Montandon et al., 2019; Nishimura et al., 2017; Nishimura and van Wijk, 2015; Welsch et al., 2018). This complex demonstrates evolutionary convergence with the eukaryotic proteasome and is important for protein homeostasis in the plastid. Proteins targeted for degradation are identified by Clp adapter proteins, which deliver them to the Clp chaperone complex. The chaperone proteins use ATP to unfold protein targets into the core, which is the site of proteolysis (Nishimura et al., 2017; Nishimura and van Wijk, 2015). The Clp complex is widely conserved across bacteria; in *E. coli* and most other bacterial species, the core consists of 14 copies of a single subunit (Nishimura and van Wijk, 2015; Yu and Houry, 2007). However, gene duplication in the cyanobacterial/plastid lineage has led to a core consisting of nine different subunits present in one to three copies each in *Arabidopsis thaliana* (Nishimura and van Wijk, 2015; Olinares et al., 2011; Peltier et al., 2004; Sjögren et al., 2006; Stanne et al., 2007). These gene duplications occurred prior to the divergence of land plants (Olinares et al., 2011). In addition to gene duplication, the evolutionary history of this complex includes gene transfer from the plastid to the nucleus, which is a phenomenon common to both mitochondria and plastids (Keeling and Palmer, 2008; Timmis et al., 2004). Among plastid Clp complex components, only one of the core subunits is encoded by the plastid itself—the other eight core subunits, as well as all of the adapter, chaperone, and accessory proteins, are encoded by the nucleus (Nishimura and van Wijk, 2015).

The plastid Clp complex serves as a model system for studying molecular coevolution for several reasons. First, this complex represents an example of protein complex evolution via gene duplication—the core subunits, as well as the chaperone and accessory subunits, have all been shaped through gene duplication and subsequent divergence of paralogs (Nishimura and van Wijk, 2015). In particular, the Clp

complex core is an extreme example of the shift from homomeric to heteromeric complex through the duplication of self-interacting proteins; what was once a homomeric complex of 14 identical subunits now consists of nine different types of subunits (Nishimura and van Wijk, 2015; Olinares et al., 2011). Second, this complex is an example of a cytonuclear collaboration, which means that the evolutionary pressures acting on ClpP1, the plastid-encoded subunit, and the nuclear-encoded subunits vary dramatically. These cytonuclear interactions are especially predominant in the core of the complex, where ClpP1 must tightly associate with multiple types of nuclear-encoded core subunits (Nishimura and van Wijk, 2015). Third, genes encoding the Clp complex core display significant rate variation across flowering plants (Rockenbach et al., 2016; Sloan et al., 2014a; Williams et al., 2019). This variation presents yet another dynamic that can illuminate mechanisms of coevolution.

This dissertation explores elements of coevolution in a variety of ways. Chapter 2 characterizes the variation in ClpP1 evolutionary rates across green plants as well as structural changes in its encoding gene, including the loss of introns, catalytic sites, and RNA editing sites. In addition to an evolutionary history of ClpP1, this chapter also provides an ERC analysis comparing ClpP1 with its nuclear-encoded counterparts in the Clp core. I also report proteomic data for the *Silene noctiflora* plastid Clp complex; this species is of special interest because it has a plastid Clp complex with one of the most highly accelerated rates of evolution across all green plants. Chapter 3 provides genomic resources for *S. noctiflora* in addition to identifying a triplication of *CLPR2* in this species. I also report a nuclear genome size for this species and compare it to other members of the genus, including a second *Silene* species with a high rate of evolution in the plastid Clp complex. Chapter 4 describes the duplication history of the eight nuclear-encoded plastid Clp core subunits and characterizes rate patterns of recent paralogs post-duplication. This chapter also analyzes evolutionary rates of another set of gene duplicates encoding a related plastid enzyme. Overall, this dissertation explores many layers of molecular coevolution and lays the foundation for future studies in this protease complex and beyond.

LITERATURE CITED

Andreeva, A., Murzin, A.G., 2006. Evolution of protein fold in the presence of functional constraints. Current Opinion in Structural Biology, Nucleic acids/Sequences and topology 16, 399–408. https://doi.org/10.1016/j.sbi.2006.04.003

Apitz, J., Nishimura, K., Schmied, J., Wolf, A., Hedtke, B., Wijk, K.J. van, Grimm, B., 2016. Posttranslational Control of ALA Synthesis Includes GluTR Degradation by Clp Protease and Stabilization by GluTR-Binding Protein. Plant Physiology 170, 2040–2051. https://doi.org/10.1104/pp.15.01945

Atchley, W.R., Wollenberg, K.R., Fitch, W.M., Terhalle, W., Dress, A.W., 2000. Correlations Among Amino Acid Sites in bHLH Protein Domains: An Information Theoretic Analysis. Molecular Biology and Evolution 17, 164–178. https://doi.org/10.1093/oxfordjournals.molbev.a026229

Barnard-Kubow, K.B., Sloan, D.B., Galloway, L.F., 2014. Correlation between sequence divergence and polymorphism reveals similar evolutionary mechanisms acting across multiple timescales in a rapidly evolving plastid genome. BMC Evol Biol 14. https://doi.org/10.1186/s12862-014-0268-y

Barreto, F.S., Burton, R.S., 2013. Evidence for compensatory evolution of ribosomal proteins in response to rapid divergence of mitochondrial rRNA. Mol. Biol. Evol. 30, 310–314. https://doi.org/10.1093/molbev/mss228

Beck, E.A., Thompson, A.C., Sharbrough, J., Brud, E., Llopart, A., 2015. Gene flow between Drosophila yakuba and Drosophila santomea in subunit V of cytochrome c oxidase: A potential case of cytonuclear cointrogression. Evolution 69, 1973–1986. https://doi.org/10.1111/evo.12718

Birky, C.W., 1995. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. PNAS 92, 11331–11338. https://doi.org/10.1073/pnas.92.25.11331

Bouchnak, I., van Wijk, K.J., 2021. Structure, Function and Substrates of Clp AAA+ protease systems in cyanobacteria, plastids and apicoplasts; a comparative analysis. Journal of Biological Chemistry 100338. https://doi.org/10.1016/j.jbc.2021.100338

Caspari, E., 1948. Cytoplasmic Inheritance, in: Demerec, M. (Ed.), Advances in Genetics. Academic Press, pp. 1–66. https://doi.org/10.1016/S0065-2660(08)60465-4

Clark, N.L., Alani, E., Aquadro, C.F., 2012. Evolutionary rate covariation reveals shared functionality and coexpression of genes. Genome Res 22, 714–720. https://doi.org/10.1101/gr.132647.111

Clark, N.L., Aquadro, C.F., 2010. A novel method to detect proteins evolving at correlated rates: identifying new functional relationships between coevolving proteins. Mol. Biol. Evol. 27, 1152–1161. https://doi.org/10.1093/molbev/msp324

Codoñer, F.M., Fares, M.A., 2008. Why Should We Care about Molecular Coevolution? Evol Bioinform Online 4, 117693430800400000. https://doi.org/10.1177/117693430800400003

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. https://doi.org/10.1016/j.ympev.2008.09.009

Ehrlich, P.R., Raven, P.H., 1964. Butterflies and Plants: A Study in Coevolution. Evolution 18, 586–608. https://doi.org/10.1111/j.1558-5646.1964.tb01674.x

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. https://doi.org/10.1371/journal.pone.0001386

Findlay, G.D., Sitnik, J.L., Wang, W., Aquadro, C.F., Clark, N.L., Wolfner, M.F., 2014. Evolutionary Rate Covariation Identifies New Members of a Protein Network Required for Drosophila melanogaster Female Post-Mating Responses. PLOS Genetics 10, e1004108. https://doi.org/10.1371/journal.pgen.1004108

Finnigan, G.C., Hanson-Smith, V., Stevens, T.H., Thornton, J.W., 2012. Evolution of increased complexity in a molecular machine. Nature 481, 360–364. https://doi.org/10.1038/nature10724

Forsythe, E.S., Williams, A.M., Sloan, D.B., 2021. Genome-wide signatures of plastid-nuclear coevolution point to repeated perturbations of plastid proteostasis systems across angiosperms. The Plant Cell 33, 980–997. https://doi.org/10.1093/plcell/koab021

Goh, C.S., Bogan, A.A., Joachimiak, M., Walther, D., Cohen, F.E., 2000. Co-evolution of proteins with their interaction partners. J Mol Biol 299, 283–293. https://doi.org/10.1006/jmbi.2000.3732

Gong, L., Olson, M., Wendel, J.F., 2014. Cytonuclear evolution of rubisco in four allopolyploid lineages. Mol Biol Evol 31, 2624–2636. https://doi.org/10.1093/molbev/msu207

Guisinger, M.M., Kuehl, J.V., Boore, J.L., Jansen, R.K., 2008. Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. PNAS 105, 18424–18429. https://doi.org/10.1073/pnas.0806759105

Haberle, R.C., Fourcade, H.M., Boore, J.L., Jansen, R.K., 2008. Extensive rearrangements in the chloroplast genome of Trachelium caeruleum are associated with repeats and tRNA genes. J. Mol. Evol. 66, 350–361. https://doi.org/10.1007/s00239-008-9086-4

Havird, J.C., Trapp, P., Miller, C.M., Bazos, I., Sloan, D.B., 2017. Causes and Consequences of Rapidly Evolving mtDNA in a Plant Lineage. Genome Biol Evol 9, 323–336. https://doi.org/10.1093/gbe/evx010

Havird, J.C., Whitehill Nicholas S., Snow Christopher D., Sloan Daniel B., 2015. Conservative and compensatory evolution in oxidative phosphorylation complexes of angiosperms with highly divergent rates of mitochondrial genome evolution. Evolution 69, 3069–3081. https://doi.org/10.1111/evo.12808

Hochberg, G.K.A., Liu, Y., Marklund, E.G., Metzger, B.P.H., Laganowsky, A., Thornton, J.W., 2020. A hydrophobic ratchet entrenches molecular complexes. Nature 1–6. https://doi.org/10.1038/s41586-020-3021-2

Ispolatov, I., Yuryev, A., Mazo, I., Maslov, S., 2005. Binding properties and evolution of homodimers in protein–protein interaction networks. Nucleic Acids Research 33, 3629–3635. https://doi.org/10.1093/nar/gki678

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., dePamphilis, C.W., Leebens-Mack, J., Müller, K.F., Guisinger-Bellian, M., Haberle, R.C., Hansen, A.K., Chumley, T.W., Lee, S.-B., Peery, R., McNeal, J.R., Kuehl, J.V., Boore, J.L., 2007. Analysis of 81 genes from 64 plastid genomes

resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. PNAS 104, 19369–19374. https://doi.org/10.1073/pnas.0709121104

Jones, S., Thornton, J.M., 1996. Principles of protein-protein interactions. PNAS 93, 13–20. https://doi.org/10.1073/pnas.93.1.13

Juan, D. de, Pazos, F., Valencia, A., 2013. Emerging methods in protein co-evolution. Nature Reviews Genetics 14, 249–261. https://doi.org/10.1038/nrg3414

Keeling, P.J., Palmer, J.D., 2008. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 9, 605–618. https://doi.org/10.1038/nrg2386

Knox, E.B., 2014. The dynamic history of plastid genomes in the Campanulaceae sensu lato is unique among angiosperms. PNAS 111, 11097–11102. https://doi.org/10.1073/pnas.1403363111

Lovell, S.C., Robertson, D.L., 2010. An Integrated View of Molecular Coevolution in Protein–Protein Interactions. Molecular Biology and Evolution 27, 2567–2575. https://doi.org/10.1093/molbev/msq144

Lynch, M., 2012. The Evolution of Multimeric Protein Assemblages. Mol Biol Evol 29, 1353–1366. https://doi.org/10.1093/molbev/msr300

Magee, A.M., Aspinall, S., Rice, D.W., Cusack, B.P., Sémon, M., Perry, A.S., Stefanović, S., Milbourne, D., Barth, S., Palmer, J.D., Gray, J.C., Kavanagh, T.A., Wolfe, K.H., 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Res. 20, 1700–1710. https://doi.org/10.1101/gr.111955.110

Majeran, W., Wollman, F.-A., Vallon, O., 2000. Evidence for a Role of ClpP in the Degradation of the Chloroplast Cytochrome b6f Complex. The Plant Cell 12, 137–149. https://doi.org/10.1105/tpc.12.1.137

Marsh, J.A., Teichmann, S.A., 2015. Structure, dynamics, assembly, and evolution of protein complexes. Annu. Rev. Biochem. 84, 551–575. https://doi.org/10.1146/annurev-biochem-060614-034142

Mode, C.J., 1958. A Mathematical Model for the Co-Evolution of Obligate Parasites and Their Hosts. Evolution 12, 158–165. https://doi.org/10.2307/2406026

Mogensen, H.L., 1996. The hows and whys of cytoplasmic inheritance in seed plants. American Journal of Botany 83, 383–404. https://doi.org/10.1002/j.1537-2197.1996.tb12718.x

Montandon, C., Friso, G., Liao, J.-Y.R., Choi, J., van Wijk, K.J., 2019. In Vivo Trapping of Proteins Interacting with the Chloroplast CLPC1 Chaperone: Potential Substrates and Adaptors. J. Proteome Res. 18, 2585–2600. https://doi.org/10.1021/acs.jproteome.9b00112

Murlas Cosmides, L., Tooby, J., 1981. Cytoplasmic inheritance and intragenomic conflict. Journal of Theoretical Biology 89, 83–129. https://doi.org/10.1016/0022-5193(81)90181-8

Nishimura, K., Kato, Y., Sakamoto, W., 2017. Essentials of Proteolytic Machineries in Chloroplasts. Molecular Plant 10, 4–19. https://doi.org/10.1016/j.molp.2016.08.005

Nishimura, K., van Wijk, K.J., 2015. Organization, function and substrates of the essential Clp protease system in plastids. Biochimica et Biophysica Acta (BBA) - Bioenergetics, SI: Chloroplast Biogenesis 1847, 915–930. https://doi.org/10.1016/j.bbabio.2014.11.012

Nooren, I.M.A., Thornton, J.M., 2003. Diversity of protein-protein interactions. EMBO J 22, 3486–3492. https://doi.org/10.1093/emboj/cdg359

Olinares, P.D.B., Kim, J., Davis, J.I., van Wijk, K.J., 2011. Subunit stoichiometry, evolution, and functional implications of an asymmetric plant plastid ClpP/R protease complex in Arabidopsis. Plant Cell 23, 2348–2361. https://doi.org/10.1105/tpc.111.086454

Osada, N., Akashi, H., 2012. Mitochondrial-nuclear interactions and accelerated compensatory evolution: evidence from the primate cytochrome C oxidase complex. Mol. Biol. Evol. 29, 337–346. https://doi.org/10.1093/molbev/msr211

Palmer, J.D., Herbon, L.A., 1988. Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J. Mol. Evol. 28, 87–97.

Park, S., Ruhlman, T.A., Weng, M.-L., Hajrah, N.H., Sabir, J.S.M., Jansen, R.K., 2017. Contrasting Patterns of Nucleotide Substitution Rates Provide Insight into Dynamic Evolution of Plastid and Mitochondrial Genomes of Geranium. Genome Biol Evol 9, 1766–1780. https://doi.org/10.1093/gbe/evx124

Peltier, J.-B., Ripoll, D.R., Friso, G., Rudella, A., Cai, Y., Ytterberg, J., Giacomelli, L., Pillardy, J., Wijk, K.J. van, 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. https://doi.org/10.1074/jbc.M309212200

Pereira-Leal, J.B., Levy, E.D., Kamp, C., Teichmann, S.A., 2007. Evolution of protein complexes by duplication of homomeric interactions. Genome Biology 8, R51. https://doi.org/10.1186/gb-2007-8-4-r51

Ramani, A.K., Marcotte, E.M., 2003. Exploiting the co-evolution of interacting proteins to discover interaction specificity. J Mol Biol 327, 273–284. https://doi.org/10.1016/s0022-2836(03)00114-1

Raza, Q., Choi, J.Y., Li, Y., O'Dowd, R.M., Watkins, S.C., Chikina, M., Hong, Y., Clark, N.L., Kwiatkowski, A.V., 2019. Evolutionary rate covariation analysis of E-cadherin identifies Raskol as a regulator of cell adhesion and actin dynamics in Drosophila. PLOS Genetics 15, e1007720. https://doi.org/10.1371/journal.pgen.1007720

Rockenbach, K., Havird, J.C., Monroe, J.G., Triant, D.A., Taylor, D.R., Sloan, D.B., 2016. Positive Selection in Rapidly Evolving Plastid–Nuclear Enzyme Complexes. Genetics 204, 1507–1522. https://doi.org/10.1534/genetics.116.188268

Sato, T., Yamanishi, Y., Kanehisa, M., Toh, H., 2005. The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics 21, 3482–3489. https://doi.org/10.1093/bioinformatics/bti564

Sjögren, L.L.E., Stanne, T.M., Zheng, B., Sutinen, S., Clarke, A.K., 2006. Structural and Functional Insights into the Chloroplast ATP-Dependent Clp Protease in Arabidopsis. The Plant Cell 18, 2635–2649. https://doi.org/10.1105/tpc.106.044594

Sloan, D.B., Alverson, A.J., Chuckalovcak, J.P., Wu, M., McCauley, D.E., Palmer, J.D., Taylor, D.R., 2012a. Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria with Exceptionally High Mutation Rates. PLOS Biology 10, e1001241. https://doi.org/10.1371/journal.pbio.1001241

Sloan, D.B., Alverson, A.J., Wu, M., Palmer, J.D., Taylor, D.R., 2012b. Recent Acceleration of Plastid Sequence and Structural Evolution Coincides with Extreme Mitochondrial Divergence in the Angiosperm Genus Silene. Genome Biol Evol 4, 294–306. https://doi.org/10.1093/gbe/evs006

Sloan, D.B., Triant, D.A., Forrester, N.J., Bergner, L.M., Wu, M., Taylor, D.R., 2014a. A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae (Caryophyllaceae). Molecular Phylogenetics and Evolution 72, 82–89. https://doi.org/10.1016/j.ympev.2013.12.004

Sloan, D.B., Triant, D.A., Wu, M., Taylor, D.R., 2014b. Cytonuclear interactions and relaxed selection accelerate sequence evolution in organelle ribosomes. Mol. Biol. Evol. 31, 673–682. https://doi.org/10.1093/molbev/mst259

Smith, D.R., Keeling, P.J., 2015. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. PNAS 112, 10177–10184. https://doi.org/10.1073/pnas.1422049112

Stanne, T.M., Pojidaeva, E., Andersson, F.I., Clarke, A.K., 2007. Distinctive Types of ATP-dependent Clp Proteases in Cyanobacteria. J. Biol. Chem. 282, 14394–14402. https://doi.org/10.1074/jbc.M700275200

Thompson, J.N., 2009. The Coevolutionary Process, The Coevolutionary Process. University of Chicago Press.

Thompson, J.N., 1989. Concepts of coevolution. Trends in Ecology & Evolution 4, 179–183. https://doi.org/10.1016/0169-5347(89)90125-0

Thompson, J.N., Burdon, J.J., 1992. Gene-for-gene coevolution between plants and parasites. Nature 360, 121–125. https://doi.org/10.1038/360121a0

Timmis, J.N., Ayliffe, M.A., Huang, C.Y., Martin, W., 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet 5, 123–135. https://doi.org/10.1038/nrg1271

Welsch, R., Zhou, X., Yuan, H., Álvarez, D., Sun, T., Schlossarek, D., Yang, Y., Shen, G., Zhang, H., Rodriguez-Concepcion, M., Thannhauser, T.W., Li, L., 2018. Clp Protease and OR Directly Control the Proteostasis of Phytoene Synthase, the Crucial Enzyme for Carotenoid Biosynthesis in Arabidopsis. Mol Plant 11, 149–162. https://doi.org/10.1016/j.molp.2017.11.003

Weng, M.-L., Ruhlman, T.A., Jansen, R.K., 2016. Plastid–Nuclear Interaction and Accelerated Coevolution in Plastid Ribosomal Genes in Geraniaceae. Genome Biol Evol 8, 1824–1838. https://doi.org/10.1093/gbe/evw115

Williams, A.M., Friso, G., Wijk, K.J. van, Sloan, D.B., 2019. Extreme variation in rates of evolution in the plastid Clp protease complex. The Plant Journal 98, 243–259. https://doi.org/10.1111/tpj.14208

Williams, A.V., Boykin, L.M., Howell, K.A., Nevill, P.G., Small, I., 2015. The Complete Sequence of the Acacia ligulata Chloroplast Genome Reveals a Highly Divergent clpP1 Gene. PLOS ONE 10, e0125768. https://doi.org/10.1371/journal.pone.0125768

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

Yan, Z., Ye, G., Werren, J., 2018. Evolutionary rate coevolution between mitochondria and mitochondria-associated nuclear-encoded proteins in insects. bioRxiv 288456. https://doi.org/10.1101/288456

Yosef, N., Kupiec, M., Ruppin, E., Sharan, R., 2009. A complex-centric view of protein network evolution. Nucleic Acids Res 37, e88. https://doi.org/10.1093/nar/gkp414

Yu, A.Y.H., Houry, W.A., 2007. ClpP: A distinctive family of cylindrical energy-dependent serine proteases. FEBS Letters 581, 3749–3757. https://doi.org/10.1016/j.febslet.2007.04.076

Zhang, J., Ruhlman, T.A., Sabir, J., Blazier, J.C., Jansen, R.K., 2015. Coordinated Rates of Evolution between Interacting Plastid and Nuclear Genes in Geraniaceae. The Plant Cell 27, 563–573. https://doi.org/10.1105/tpc.114.134353

Zhang, J., Ruhlman, T.A., Sabir, J.S.M., Blazier, J.C., Weng, M.-L., Park, S., Jansen, R.K., 2016. Coevolution between Nuclear-Encoded DNA Replication, Recombination, and Repair Genes and Plastid Genome Complexity. Genome Biol Evol 8, 622–634. https://doi.org/10.1093/gbe/evw033

CHAPTER 2: EXTREME VARIATION IN RATES OF EVOLUTION IN THE PLASTID CLP PROTEASE COMPLEX[1]

Summary

Eukaryotic cells represent an intricate collaboration between multiple genomes, even down to the level of multisubunit complexes in mitochondria and plastids. One such complex in plants is the caseinolytic protease (Clp), which plays an essential role in plastid protein turnover. The proteolytic core of Clp comprises subunits from one plastid-encoded gene (*clpP1*) and multiple nuclear genes. The *clpP1* gene is highly conserved across most green plants, but it is by far the fastest evolving plastid-encoded gene in some angiosperms. To better understand these extreme and mysterious patterns of divergence, we investigated the history of *clpP1* molecular evolution across green plants by extracting sequences from 988 published plastid genomes. We find that *clpP1* has undergone remarkably frequent bouts of accelerated sequence evolution and architectural changes (e.g., loss of introns and RNA-editing sites) within seed plants. Although *clpP1* is often assumed to be a pseudogene in such cases, multiple lines of evidence suggest that this is rarely true. We applied comparative native gel electrophoresis of chloroplast protein complexes followed by protein mass spectrometry in two species within the angiosperm genus *Silene*, which has highly elevated and heterogeneous rates of *clpP1* evolution. We confirmed that *clpP1* is expressed as a stable protein and forms oligomeric complexes with the nuclear-encoded Clp subunits, even in one of the most divergent *Silene* species. Additionally, there is a tight correlation between amino-acid substitution rates in *clpP1* and the nuclear-encoded Clp subunits across a broad sampling of angiosperms, suggesting ongoing selection on interactions within this complex.

Introduction

Rates of sequence evolution vary dramatically across genes and genomes. Understanding the

[1] Published in *The Plant Journal*, Volume 98, April 2019.

Authors: Alissa M. Williams, Giulia Friso, Klaas J. van Wijk, Daniel B. Sloan

forces that determine such variation is one of the defining goals in the field of molecular evolution. In seed plants, the plastid genome (plastome) generally evolves two- to six-fold more slowly than the nuclear genome (Drouin et al., 2008; Smith and Keeling, 2015; Wolfe et al., 1987). However, among angiosperms, there is considerable heterogeneity in the rate of plastome evolution. Many lineages have maintained a slowly-evolving plastome, while others have experienced drastic rate increases (Jansen et al., 2007). For instance, among close relatives within the tribe *Sileneae* there have been at least three recent and independent increases in plastome evolutionary rate (Erixon and Oxelman, 2008; Sloan et al., 2014a). Similar accelerations have been documented in the Campanulaceae (Barnard-Kubow et al., 2014; Haberle et al., 2008; Knox, 2014)(Barnard-Kubow et al., 2014) , Geraniaceae (Guisinger et al., 2008; Weng et al., 2014), Fabaceae (Magee et al., 2010), and Poaceae (Guisinger et al., 2010), among other taxa. At a structural level, plastome gene order has largely been conserved, with most angiosperms retaining the structural organization that was present in the most recent common ancestor of this group (Raubeson and Jansen, 2005). However, the sporadic increases in rates of plastome sequence evolution have often been accompanied by structural changes, including indels, inversions, duplications, shifts in inverted-repeat boundaries, and gene and intron loss (Guisinger et al., 2010; Jansen et al., 2007; Sloan et al., 2014a; Weng et al., 2014).

Interestingly, increases in plastome evolutionary rate are often not genome-wide; rather, these increases tend to affect a subset of genes (Guisinger et al., 2008; Sloan et al., 2014a). Commonly affected genes include those encoding the essential chloroplast factors Ycf1 and Ycf2 (Sloan et al., 2012b), RNA polymerase subunits (Blazier et al., 2016), ribosomal proteins (Guisinger et al., 2008), and the AccD subunit of the acetyl-CoA carboxylase enzyme complex (Rockenbach et al., 2016). Some of the most striking and extreme accelerations are found in *clpP1*, which encodes a core subunit of the plastid caseinolytic protease (Clp) (Erixon and Oxelman, 2008; Barnard-Kubow et al., 2014; Zhang et al., 2014; Williams et al., 2015).

The plastid Clp complex plays an important role in maintaining homeostasis in plant cells by stabilizing the plastid proteome. It is the most abundant stromal protease in developing chloroplasts and

has been shown to degrade various chloroplast proteins, e.g., the cytochrome b_6f complex, a copper transporter, glutamyl-tRNA reductase, and phytoene synthase (Apitz et al., 2016; Majeran et al., 2000; Nishimura et al., 2017; Nishimura and van Wijk, 2015; Welsch et al., 2018). Consistent with the significance of the Clp complex, plastid-encoded ClpP1 as well as most nuclear-encoded subunits are essential for plant growth and viability (Clarke et al., 2005; Kim et al., 2009; Koussevitzky et al., 2007; Kuroda and Maliga, 2003; Nishimura and van Wijk, 2015; Zheng et al., 2006).

The proteolytic core of the Clp complex is composed of two stacked heptameric rings (Nishimura and van Wijk, 2015; Yu and Houry, 2007) (**Figure 2.1a**). In most bacteria, including *E. coli*, these 14 subunits are identical, encoded by a single gene (*clpP*). The plastid Clp core is also composed of two heptameric rings, but there have been multiple duplications of the *clpP* gene throughout cyanobacterial and plastid evolution such that the rings now comprise numerous paralogous subunits (Majeran et al., 2005; Nishimura and van Wijk, 2015; Olinares et al., 2011a; Peltier et al., 2001). In green plants, only *clpP1* is retained in the plastome, and the other paralogs are found in the nucleus. In *Arabidopsis thaliana,* there are a total of eight nuclear paralogs, and subunit composition differs between the two core rings. The ClpP1/R ring contains three copies of the plastid-encoded ClpP1 subunit and one of each of four ClpR subunits (ClpR1-4), while the ClpP3-6 ring contains the nuclear-encoded ClpP subunits (ClpP3-6) in a 1:2:3:1 ratio (Olinares et al., 2011b) (**Figure 2.1a**). Catalytic activity in Clp subunits is conferred by an amino-acid triad (Ser 101, His 126, Asp 176) (**Figure 2.1b**). The distinguishing feature between ClpP and ClpR subunits is that the latter have each lost at least one catalytic residue and are thus assumed to be non-proteolytic; ClpR subunits are also found in the cyanobacterium *Synechococcus elongatus* and the apicoplast of the parasite *Plasmodium falsiparum* (El Bakkouri et al., 2013; Peltier et al., 2004; Schelin et al., 2002). Therefore, ClpP1 (the sole plastid-encoded subunit) appears to be the only catalytic member of the ClpP1/R ring. In addition to the core, the plastid Clp complex also contains several chaperone, accessory, and adaptor subunits required for proper Clp function, all of which are nuclear-encoded (Nishimura and van Wijk, 2015) (**Figure 2.1a**).

The importance of the plastid Clp system to cellular function makes accelerations in *clpP1* evolutionary rate particularly surprising. Various mechanisms have been hypothesized to explain such cases of rapid plastid gene evolution (Blazier et al., 2016; Erixon and Oxelman, 2008; Guisinger et al., 2008; Magee et al., 2010; Williams et al., 2015), but the relative contributions of adaptive evolution, relaxed selection, outright pseudogenization, and increased mutation rates remain unclear. Because Clp subunits are encoded by two different genomes, cytonuclear interactions are integral to the functioning of this complex. There is growing evidence that accelerated organelle genome evolution can lead to correlated increases in rates of evolution in interacting nuclear-encoded proteins. Such effects have been detected in the plastid Clp complex (Rockenbach et al., 2016), as well as plastid and mitochondrial ribosomes (Sloan et al., 2014a; Weng et al., 2016), the plastid-encoded RNA polymerase (Zhang et al., 2015), mitochondrial oxidative phosphorylation (OXPHOS) complexes (Havird et al., 2015; Li et al., 2017; Yan et al., 2018), and DNA replication and repair machinery that directly interacts with plastid and mitochondrial genomes (Havird et al., 2017; Zhang et al., 2016). In plants, however, such studies have been limited to close relatives within just two groups (Geraniaceae and *Silene*) and have not been examined at deeper timescales across angiosperms.

Here, to better understand the context and scope of plastid *clpP1* acceleration, we provide a detailed accounting of the molecular evolutionary history of *clpP1* across the entire green plant lineage. Further, we use proteomic techniques to assess the functional status of *clpP1* in two species within the genus *Silene*; this genus has some of the highest and most variable observed rates of divergence for this gene. Finally, we determine whether coevolution between plastid- and nuclear-encoded subunits of the Clp complex is a broad and repeatable pattern across angiosperm diversity.

Results

ClpP1 has undergone numerous and extreme accelerations in rates of amino-acid substitution across independent lineages of green plants

Massive acceleration in ClpP1 amino-acid substitution rate has occurred multiple times across green plants (**Figure 2.2, Figure S2.1**). These accelerations are most pronounced in seed plants, with striking examples found in conifers, gnetophytes, and numerous angiosperm lineages. Therefore, previous reports of divergent ClpP1 sequences in specific lineages (Erixon and Oxelman, 2008; Barnard-Kubow et al., 2014; Zhang et al., 2014; Williams et al., 2015) appear to be the result of an incredibly frequent occurrence of accelerations over the course of seed plant evolution.

The cases of ClpP1 rate acceleration in seed plants are especially remarkable when considered against the high level of conservation in many related land plant lineages. For instance, since the split at the base of the land plant phylogeny *ca.* 490 Mya (Morris et al., 2018), a representative liverwort (*Marchantia polymorpha*) and a representative hornwort (*Anthoceros angustus*) have exhibited a total rate of 0.1 amino-acid substitutions per site per billion years. At the other extreme, the angiosperms *Silene noctiflora* and *S. conica* diverged only about 5 million years ago (Rautenberg et al., 2012) and have since exhibited a rate of 335 amino acid substitutions per site per billion years – a >3000-fold increase relative to the rate of divergence between liverworts and hornworts.

To assess whether the observed rate variation in ClpP1 was the result of a plastome-wide effect, we repeated our analyses on a representative photosynthetic protein (PsaA) and found considerably less rate heterogeneity (**Figure 2.2**). This result is in line with the widespread conservation that is characteristic of plastid genes and especially those involved in photosynthesis (Guisinger et al., 2008). Rate increases in PsaA have been much smaller than those in ClpP1 (Newick tree files provided at https://github.com/alissawilliams/clpP1_2018). For example, using the same comparisons as above between bryophytes and *Silene*, there has been a rate of 0.22 amino acid substitutions per site per billion years between *Marchantia polymorpha* and *Anthoceros angustus*, whereas this rate is 2.9 amino acid substitutions per site per billion years between *Silene noctiflora* and *S. conica*. Therefore, the rate has increased roughly 10-fold in *Silene* relative to the liverwort-hornwort comparison, but it is not nearly the 3000-fold increase we observe in ClpP1.

Although the most dramatic cases of ClpP1 accelerations were found in seed plants, there was also rate heterogeneity among lineages of green algae, albeit at much deeper timescales than within seed plants. The most extreme examples of ClpP1 divergence in algae are found in lineages such as *Chlorokybus*, *Mesostigma*, *Stichococcus*, and Chlorophyceae. The rapid sequence evolution within Chlorophyceae is perhaps not surprising because large insertions in ClpP1 have been previously identified in this group (Derrien et al., 2012; Huang et al., 1994; Majeran et al., 2005).

ClpP1 acceleration is correlated with changes in structure and gene architecture

Our analysis demonstrated that accelerated rates of ClpP1 amino-acid substitution are also associated with broader changes at a structural level. For example, we confirmed that the few species with identified *clpP1* duplications within the plastome (**Figure S2.2)** also have high rates of amino acid substitution ($P = 0.006$) (Erixon and Oxelman, 2008; Park et al., 2017). Angiosperm examples of *clpP1* duplication include *Silene chalcedonica* (=*Lychnis chaldeconica*), *Carex siderosticta*, and multiple lineages within the Geraniaceae. By sampling a subset of angiosperm species, we also observed that high substitution rates tend to be associated with indels in *clpP1* (**Figure S2.3**). This relationship was not statistically significant regardless of whether we normalized by nuclear substitution rates (**Figure S2.3a**: Spearman's rho = 0.36*,* $P = 0.08$) or plastid substitution rates (**Figure S2.3b**: Spearman's rho = 0.20, $P = 0.34$). However, that may reflect limited power resulting from our conservative approach to identifying indel events in the gappy *clpP1* sequencing alignment because overlapping indels of different lengths were treated as a single indel type by the SeqState software tool.

Accelerated *clpP1* evolution is also associated with intron loss in many lineages (**Figure S2.4**). Most land plants share two *clpP1* introns, which appear to have been gained in series during the evolution of streptophytes. Intron 1 was likely gained in a common ancestor of Charophyceae, Coleochaetophyceae, Zygnemophyceae, and land plants based on its presence in *Chara vulgaris*, *Chaetosphaeridium globosum*, and *Mesotaenium endlicherianum*. Intron 2 appears to have been acquired more recently in a common ancestor of Zygnemophyceae and land plants because the only algal species in which it was identified was

M. endlicherianum, which is consistent with the conclusion that Zygnemophyceae is the sister lineage to all land plants (Wickett et al., 2014). The only other *clpP1* intron identified in our sample appears to be an independent acquisition in the streptophyte *Klebsormidium flaccidum* at a unique position within the gene. This inferred history of intron gains implies multiple secondary losses within green algae because multiple algal species within these groups lack one or both of the introns. Strikingly, we identified at least 31 independent losses of one or both introns in land plants (**Figure 2.3**), which are strongly associated with accelerated rates of ClpP1 evolution ($P << 0.001$ for both), including in multiple *Silene* species, the genus *Oenothera*, and *Plantago maritima* (**Figure S2.4**).

In land plant mitochondria and plastids, there is frequent C-to-U RNA editing (Freyer et al., 1997; Tsudzuki et al., 2001), and there is evidence that accelerated sequence evolution can also be associated with loss of editing sites (Guo et al., 2016; Parkinson et al., 2005; Sloan et al., 2010; Zhu et al., 2014). In *Arabidopsis thaliana*, there is a single RNA editing site in *clpP1* at codon 187, where the codon CAU (His) is edited to UAU (Tyr) (Tillich et al., 2005). This edited C has been replaced in most accelerated angiosperm species ($P << 0.001$), usually in favor of "hard-coding" a T (U) at this position (**Figure S2.5**). In contrast, the C is maintained in most non-accelerated lineages (though there are exceptions such as the loss of this site at the base of the asterids (Hein and Knoop, 2018) well before any rate accelerations in that group). We also examined a second site in codon 28, which also undergoes CAU (His) to UAU (Tyr) editing in angiosperms (Hein and Knoop, 2018). This editing site was replaced with a hard-coded T at the base of the core eudicots but appears to be retained in other major angiosperm groups, where there is a trend toward hard-coding in accelerated species (**Figure S2.6**). However, this relationship does not approach statistical significance ($P = 0.43$).

clpP1 loss or pseudogenization events in green plants

To assess whether the functional inactivation (i.e., pseudogenization) of *clpP1* may contribute to cases of extreme rate accelerations, we looked for evidence of potential pseudogenes and outright gene loss across green plants. Beyond the obvious loss of *clpP1* in the holoparasitic genera *Rafflesia* and

Polytomella, which lack any detectable plastomes whatsoever (Molina et al., 2014; Smith and Lee, 2014), we have identified an additional seven species that appear to lack *clpP1* in their plastomes (**Table 2.1**). Of these seven lineages, which encompass both green algae and angiosperms, five are holoparasitic or mycoheterotrophic. The loss of photosynthesis in holoparasites and mycoheterotrophs is typically associated with radical changes in the plastome (Bromham et al., 2013; Krause, 2008). Although *clpP1* is often retained in these species despite massive loss of the genes that encode photosynthetic machinery (Delannoy et al., 2011; Wolfe et al., 1992), our results indicate that non-photosynthetic genes such as *clpP1* also face increased probability of loss during the evolution of holoparasitism.

In addition to cases of outright loss, there are species in which *clpP1* is still present but may be a pseudogene. For instance, the reported sequence in the angiosperm *Epipremnum aureum* is radically altered as part of a partial tandem duplication, and there are other angiosperms in which the first exon has been lost (**Table 2.1**). There are also lineages in which *clpP1* contains an internal stop codon. However, these cases most likely reflect posttranscriptional modifications and changes in translation rather than actual pseudogenes. The internal stop codon in the hornwort *Anthoceros angustus* has been shown to be removed by U-to-C RNA editing (Kugita et al., 2003), and it is possible that a similar mechanism removes internal stop codons in the hornwort *Nothoceros aenigmaticus* and the fern *Diplopterygium glaucum* – lineages in which U-to-C plastid RNA editing is prevalent (Duff and Moore, 2005; Knie et al., 2016). In addition, the copy of *clpP1* in the chlorophyte *Jenufa minuta* contains multiple (UGA) stop codons, but these are found at positions normally encoding conserved Trp residues in numerous genes within this plastome (Lemieux et al., 2015), suggesting that *J. minuta* has undergone a change in the plastid genetic code in which the UGA codon has been reassigned to encode Trp.

ClpP1 is still expressed and likely assembles with nuclear-encoded Clp subunits in *Silene* species that exhibit extreme heterogeneity in rates of ClpP1 sequence evolution

To further assess whether observed increases in rates of ClpP1 sequence evolution reflect a loss of functionality, we took advantage of the rate heterogeneity within the angiosperm genus *Silene*. We

selected *S. latifolia* as a species that has retained a highly conserved copy of ClpP1 and *S. noctiflora* as a close relative with a recent and extreme rate acceleration that has resulted in one of the most divergent copies of ClpP1 in our dataset (**Figure 2.2**), including the substitution of both histidine and aspartate in its catalytic triad. We isolated intact chloroplasts and separated the native soluble stromal complexes by native gel electrophoresis (LB-Native-PAGE), which we have applied in the past to determine the assembly state of the ClpPR core and other stromal complexes in Arabidopsis (Olinares et al., 2011a; Peltier et al., 2006). Based on tandem mass spectrometry (MS/MS) of size-fractionated samples, over 97% of all adjusted spectral counts (AdjSPC) matched annotated plastid proteins, with 1.3 and 0.9% of the AdjSPC matching to the Clp protein family in *S. noctiflora* and *S. latifolia,* respectively. We identified (i.e., deduced from the *Silene* species sequence data) all predicted ClpPR proteins, including ClpP1 (**Figure 2.4**). Further, we detected ClpP1 in the same size fractions as the ClpR subunits in both species, which is consistent with the expectation of a mass of *ca.* 200 kDa for the ClpP1/R ring (Olinares et al., 2011a), providing indirect evidence that ClpP1 still assembles as part of the ring structure that makes up the proteolytic core of the plastid Clp complex (**Figure 2.4; Figure S2.7)**.

Signatures of selection and rate variation across ClpP1

In species with accelerated ClpP1 sequence evolution, substitutions were widely distributed across the length of the protein (**Figure 2.5a**), but individual sites varied substantially in their degree of conservation (**Figure 2.5b**). Only a single residue (Gly at position 110) was invariant across all sampled green plants. As expected, residues in the catalytic triad were broadly conserved, though Asp 176 has been lost in over 20 species (**Figure S2.8)**. Substitutions at Ser 101 and His 126 were less common but still observed in 5 and 6 species, respectively. Losses of catalytic residues were each significantly correlated with accelerated ClpP evolution (Ser: $P = 0.012$; His: $P = 0.008$; Asp: $P << 0.001$). Notably, some species have experienced substitutions at multiple sites in the catalytic triad; for example, *Plantago maritima* and *Vaccinium macrocarpon* have both lost all three residues.

We partitioned the ClpP1 sequence into major functional domains and applied maximum-likelihood models to compare rates of amino-acid substitution across these partitions. We found that rates differed significantly among regions ($\chi^2 = 337.9$; d.f. = 3; $P << 0.001$), with the highest rates observed in the predicted "handle" domain (**Figure 2.5a)**, which is likely involved in physical interactions between the two heptameric rings that make up the tetradecameric proteolytic core of the Clp complex (Yu and Houry, 2007). We also observed that the beginning of the handle domain appeared to be a hotspot for structural variants, including large insertions in *Carex*, *Eustrephus*, *Silene*, *Taxus*, and *Viviania* (alignment available at https://github.com/alissawilliams/clpP1_2018)**.**

To take advantage of independent acceleration events and avoid the saturation expected when examining deep splits in the green plant phylogeny, our further analyses focused on a broad sample of 25 angiosperm species representing a wide range of rate variation. We ran a PAML branch-site model on this sample, using ClpP1-accelerated lineages as the foreground, to determine whether there are signatures of positive selection on specific codons across multiple accelerated species. This analysis identified 32 amino acid residues with greater than 95 percent probability of being under positive selection. Based on the *E. coli* ClpP structure, these 32 sites are scattered across the protein in no obvious spatial pattern (**Figure 2.1b**).

We also calculated gene-wide average ratios of nonsynonymous to synonymous substitution rates (d_N/d_S) for *clpP1* and *psaA* on each branch in our angiosperm phylogeny. Across both accelerated and non-accelerated species, *clpP1* d_N/d_S ratios tend to be greater than the corresponding *psaA* ratios (**Figure 2.6**). As expected, the *clpP1* d_N/d_S ratios are higher in species that were identified as having highly divergent ClpP1 protein sequences. However, we found that some slow species (*Solanum lycopersicum, Cucumis sativus*, and *Vitis vinifera*) have d_N/d_S ratios more characteristic of accelerated species despite much lower overall levels of ClpP1 protein sequence divergence. We obtained gene-wide d_N/d_S ratios greater than 1 for *clpP1* in several species (**Table S2.1**) which could indicate extreme positive selection in these branches, but none of these results were statistically significant (uncorrected *p*-values all greater than 0.05). We found that accelerated species have also experienced increased synonymous substitution

rates, though not nearly to the extent of the increase in nonsynonymous substitution rates (**Figure S2.9**), which explains the observed increases in d_N/d_S (**Figure 2.6**).

ClpP1 accelerations are highly correlated with accelerations in nuclear-encoded Clp genes

Across our sample of 25 angiosperms, the amino-acid substitution rate in ClpP1 is strongly associated with the amino acid-substitution rate of the nuclear-encoded core Clp subunits (**Figure 2.7a-c**). Notably, the mirroring effect between the ClpP1 and nuclear-encoded Clp branch lengths does not occur in a random sample of nuclear-encoded proteins, indicating that the correlated accelerations are not due to genome-wide rates of evolution. After using branch lengths from random genes to account for background rate variation, the vast majority of the variation in branch length for nuclear-encoded Clp subunits can be explained by the ClpP1 branch length for a particular lineage. This was true regardless of whether ClpP1 rates were normalized with nuclear proteins ($R^2 = 0.88$, $P << 0.001$) (**Figure 2.7d**) or with a set of plastid-encoded photosynthetic proteins ($R^2 = 0.86$, $P << 0.001$) (**Figure S2.10**).

Discussion

Can pseudogenization explain massive accelerations in rates of *clpP1* evolution?

Our analysis shows that accelerated *clpP1* evolution has occurred frequently and independently across green plants—particularly among seed plants. Accelerations in *clpP1* are thus a striking feature of seed plant evolution, especially given that *clpP1* is highly conserved in a majority of plant species. Pseudogenization has often been hypothesized as an explanation for extreme *clpP1* divergence (Hirao et al., 2008; Williams et al., 2015; Zhang et al., 2014), and in many other plastome sequencing projects, *clpP1* gene sequences have gone completely unrecognized and unannotated because of their extreme divergence (Fajardo et al., 2013; Haberle et al., 2008; Straub et al., 2011; Yao et al., 2015). There has also been speculation in these cases that *clpP1* has been functionally transferred to the nucleus, as intracellular gene transfer is a common and ongoing phenomenon in plants (Adams et al., 2002b; Millen et al., 2001).

In the highly reorganized *Boodlea composita* plastome (Cortona et al., 2017), *clpP1* appears to have been lost and possibly transferred to the nucleus. However, in searching the assembled transcriptome (A. Del Cortona pers. comm.) of this unusual species of green algae, we found it difficult to unambiguously identify *clpP*-like sequences as derived from *clpP1* or its ancestral nuclear due to deep sequence divergence. Thus, we are not aware of any clear examples of plastid *clpP1* transfer to the nucleus in green plants.

Although we have identified probable cases of *clpP1* loss or pseudogenization within the plastome in a relatively small number of species (**Table 2.1**), it is unlikely that these processes represent a general explanation for the repeated and widespread pattern of substitution-rate acceleration in green plants. In the vast majority of our sampled species, *clpP1* reading frames have remained intact, even in species with extreme rates of indels and nucleotide substitutions. In the absence of functional constraints, such rapid change would quickly introduce internal stop codons and frameshifts, which suggests that there is still selection in these species to retain a functional gene copy.

Previous work has provided some evidence that even highly divergent *clpP1* genes may still be functional. For instance, the divergent copies of *clpP1* in *Acacia ligulata* and *Campanulastrum americanum* are still transcribed and spliced (Barnard-Kubow et al., 2014; Williams et al., 2015). In this study, we examined the plastid proteome of *Silene noctiflora*, a species with one of the most divergent known copies of *clpP1*, and found evidence that such divergent genes can still be expressed at the protein level and co-assemble with other Clp subunits. These results suggest that ClpP1 in *S. noctiflora* is still an important part of the core Clp structure, despite the fact that it has lost two members of its catalytic triad. If the subunit composition of ClpP1/R ring is indeed conserved, there are potentially important (but currently unknown) functional consequences of catalytic triad loss in the only catalytic member of this ring, which may affect overall Clp catalytic activity and/or complex structure (Andersson et al., 2009; Zeiler et al., 2013). Recently, inactivation of the catalytic site in *Arabidopsis thaliana* ClpP3 was found to have no phenotypic effect, whereas complete loss of ClpP3 resulted in a severe phenotype. In contrast, in

case of ClpP5, loss of the catalytic site or complete gene loss results in embryo lethality (Kim et al., 2013; Liao et al., 2018).

We found that several nuclear-encoded ClpP subunits in fast-evolving species have substitutions at catalytic sites (**Table S2.5**); these substitutions likely render their respective proteins proteolytically inactive. However, none of the sampled species have lost catalytic sites from all ClpP proteins, meaning that each likely has at least one fully catalytic ClpP subunit. We also considered the possibility that, in lineages with one or multiple non-catalytic ClpP subunits, ClpR subunits have regained catalytic activity. While there are a few cases in which a ClpR subunit has regained one of the three catalytic residues via an amino-acid substitution, we did not find any ClpR proteins with a fully restored catalytic triad (**Table S2.5**). Thus, replacement of ClpP proteins or their catalytic activity by their ClpR counterparts is unlikely.

Role of mutation rate vs. selection in *clpP1* accelerations

Another explanation for extreme divergence in *clpP1* (and other plastid genes) could be an increase in the underlying mutation rate (Park et al., 2017). However, there are apparent difficulties with interpretations based solely on changes in mutation rate. While we would generally expect an increase in mutation rate to affect the entire plastome, it is clear from whole-plastome sequencing efforts that *clpP1* acceleration often occurs with little or no change in rates for a large fraction of the plastome (Guisinger et al., 2008; Sloan et al., 2014a; Williams et al., 2015). The locus-specific nature of *clpP1* accelerations was supported by our analysis of *psaA* substitution rates across green plants, which were consistently low, even in species with extremely divergent *clpP1* sequences (**Figure 2.1**). Thus, if *clpP1* acceleration is caused by an increase in mutation rate, it would require a highly localized effect within the plastome. While such an effect is more difficult to explain than a genome-wide increase in mutation rate, "localized hypermutation" has been suggested previously as the cause of high divergence in the plastid gene *ycf4* (Magee et al., 2010) and may be associated with the presence of short, repetitive sequences (Stoike and Sears, 1998). In our case, the species with high rates of protein sequence evolution in ClpP1 do have elevated synonymous substitution rates at this locus (**Figure S2.9**), which is often taken as a proxy for

mutation rate. There are also several documented cases of extreme variation in synonymous substitution rates across individual mitochondrial genomes, demonstrating that mutation rates do likely vary within organelle genomes in some plants (Sloan et al., 2009; Zhu et al., 2014).

One possible mechanism for localized hypermutation is "mutagenic retroprocessing" (Park et al., 2017; Parkinson et al., 2005), which occurs when a mature transcript recombines back into the genome after reverse transcription. In this scenario, accelerated substitution rates could be explained by the relatively high error rates of reverse transcriptases (Preston, 1996; Sabot and Schulman, 2006) and/or RNA polymerases (Traverse and Ochman, 2016). Retroprocessing would be expected to affect exons and introns differently. If the recombination event involves the entire gene, introns would be lost because they are not included in mature transcripts. If the recombination event involves only part of a mature transcript, that portion would necessarily be an exon, meaning that the rate acceleration would be limited to exonic regions. Both of these predictions have empirical support. Species with accelerated *clpP1* sequences often lack *clpP1* introns (**Figure S2.4**) (Erixon and Oxelman, 2008; Park et al., 2017). Among the accelerated species that do retain their *clpP1* introns, rates of sequence evolution are much higher in exons than in introns (Barnard-Kubow et al., 2014; Erixon and Oxelman, 2008). Despite these observations, it is not clear why *clpP1* would specifically or preferentially undergo mutagenic retroprocessing in the plastome, particularly because most plastid genes have high transcription rates and many are transcribed at higher rates than *clpP1* (Mullet, 1993; Sanitá Lima and Smith, 2017).

Another difficulty with an explanation based solely on mutation is that, if *clpP1* has only been subject to an increased mutation rate, we would not necessarily expect an increase in d_N/d_S. The d_N/d_S statistic is typically interpreted as a measure of selection pressure, so low values are expected for genes under purifying selection, even if the mutation rate is high. Indeed, previous work has found that increased mutation pressure can even be associated with decreased d_N/d_S values in genes that remain under strong purifying selection (Havird and Sloan, 2016; Wolf et al., 2009). In contrast, we found a trend of increased d_N/d_S values in species with fast rates of *clpP1* evolution (**Figure 2.6**). While this result may not be surprising given that amino acid substitution rates and d_N/d_S values are interconnected, it does

indicate that the rates of nonsynonymous substitutions in these lineages have increased disproportionately relative to synonymous substitutions. This result suggests that *clpP1* acceleration is due, at least in part, to a change in selective pressures on protein sequence rather than simply an increase in mutation rate. This line of argument provides an alternative interpretation for the aforementioned observation that introns do not exhibit a similar degree of accelerated sequence evolution (Barnard-Kubow et al., 2014; Erixon and Oxelman, 2008). Importantly, any interpretation of d_N/d_S ratios comes with the caveat that they can be overestimated and model-dependent, especially in cases where there are multinucleotide mutations and/or indels in the gene of interest (De Maio et al., 2013; Li et al., 2009; Stoletzki and Eyre-Walker, 2011; Venkat et al., 2017).

Why might selection pressures on *clpP1* have changed?

While localized increases in mutation rate could be involved in *clpP1* accelerations, it is unlikely that mutation rates are the only contributing factor for the reasons described above. Rather, it is likely that changes in selection are involved, and increases in d_N/d_S are typically caused by some combination of relaxed selection and/or positive selection. One possible explanation is that an increased mutation rate has itself altered selection pressures. This mechanism has been previously hypothesized in legumes, where the plastid gene *ycf4* has undergone increases in evolutionary rate in several species potentially as a result of localized hypermutation (Magee et al., 2010). Because high mutation rates can lead to an accumulation of deleterious mutations, there may be selection for affected genes to "escape" this mutation pressure by being functionally transferred to the nucleus (Blanchard and Lynch, 2000; Magee et al., 2010). If functional replacement does occur, the plastid-encoded gene would no longer be needed for its original function and thus experience relaxed selection. In this scenario, the decrease in functional constraint would occur due to gene transfer/replacement, which was initially driven by increased mutation pressure (Magee et al., 2010).

Relaxed selection can also occur without a change in underlying mutation rate. For instance, it is possible that functional constraint on the entire Clp complex could be reduced if it simply becomes less

important to cellular functioning. Such a decrease in functional constraint could conceivably occur if some of the many other plastid proteases take precedence (Nishimura et al., 2017). Alternatively, functional constraint may be reduced specifically on *clpP1* (as opposed to the entire complex). As described above, a likely cause of relaxed selection (or outright pseudogenization) would be replacement in the Clp complex by a nuclear-transferred copy of *clpP1* or one of the existing nuclear-encoded subunits. Such transfers/replacements frequently occur for other genes in plant organelles even in the absence of increased mutational pressure (Adams et al., 2002b, 2002a; Millen et al., 2001). However, given that we did not find evidence for widespread *clpP1* pseudogenization and/or gene replacement and that the highly divergent ClpP1 subunit in *S. noctiflora* still appears to be associated with other plastid Clp core subunits, it is unlikely that functional replacement of ClpP1 in the Clp complex has broadly occurred in accelerated lineages.

The other form of selection that could be involved in *clpP1* rate accelerations and increases in d_N/d_S is positive selection. Under positive selection, there is selection for change, which can lead to a superficially similar pattern of accelerated protein sequence evolution as observed under relaxed selection. Previous work has found evidence of positive selection in both nuclear- and plastid-encoded Clp core subunits (Erixon and Oxelman, 2008; Rockenbach et al., 2016). Often, positive selection is assumed to reflect an adaptation for a novel function or a response to an environmental change. While there is no obvious shared background environment or biological feature among *clpP1*-accelerated lineages, our understanding of Clp function (including the identities of many of its target substrates) remains incomplete, so adaptive change of the whole complex is still a viable hypothesis. Further, multiple bacterial lineages including *Bacillus thuringiensis* and cyanobacteria have undergone major Clp complex reorganizations as the result of core subunit duplication and diversification, suggesting selection to deviate from the conserved ancestral functions of Clp (Fedhila et al., 2002; Stanne et al., 2007).

Positive selection could also be related to the intimate interactions between the plastid Clp subunits encoded by different genomes. Using an evolutionary rate correlation analysis, we have shown that accelerations in *clpP1* were paralleled by similar accelerations in nuclear-encoded Clp genes across a

broad range of angiosperms (**Figure 2.7**). Our analysis represents a general class of computational methods that detect correlated rate changes between residues, genes, and/or complexes across a phylogeny as a means to identify genes that share a functional relationship and may be coevolving or at least responding to the same selection pressures (Clark and Aquadro, 2010; Dutheil and Galtier, 2007; Juan et al., 2013; Yeang and Haussler, 2007). Therefore, the fact that *clpP1* and the other core plastid Clp subunits had a significant rate correlation demonstrates that the plastid- and nuclear-encoded Clp subunits are subject to shared variation in selection pressures. This result could be simply due to selection acting on the complex as a whole (as described above), or it could indicate that *clpP1* and its nuclear-encoded counterparts are coevolving because of their direct interactions within the complex. Thus, coevolution between Clp subunits could be a driver of *clpP1* acceleration. A change in one subunit could introduce pressure on the other subunits to change in response—and these subsequent changes could drive further change, creating a chain reaction. Regardless of the initial trigger, this mechanism could explain both previous observations of positive selection on Clp subunits and the high correlation between their evolutionary rates. There has been speculation that such a positive feedback loop in the plastid Clp could be due to antagonistic interactions between the plastid and nuclear genomes (Rockenbach et al., 2016), and recent studies have implicated other plastid loci in selfish interactions with the nucleus (Bogdanova et al., 2015; Sobanski et al., 2018), but direct evidence for this or any other trigger for coevolutionary change is currently lacking.

In summary, our analysis has characterized the remarkable extent and repeatability of *clpP1* acceleration, provided evidence of retained functionality at the protein level even for one of the most extreme cases of *clpP1* divergence, and revealed an exceptionally strong, angiosperm-wide rate correlation within this complex. These results point to multiple possible mechanisms that should be investigated as researchers continue to disentangle the causes of variable evolutionary rates in plastid genomes.

Experimental Procedures

Extraction and filtering of plastid gene sequences

All 988 complete Viridiplantae plastome sequences available in the NCBI RefSeq collection as of May 2, 2016 were downloaded from GenBank and parsed with a custom BioPerl script to extract the annotated coding sequences and number of exons for *clpP1* and *psaA*. Nucleotide sequences for species missing after this initial step were manually extracted, either via inspection of the GenBank annotation or after a tblastn v2.2.30+ (Gertz et al., 2006) search using the corresponding *Arabidopsis thaliana* protein sequence as a query. Extracted sequences were then screened with custom Perl scripts to identify missing or internal stop codons and identify potential annotation errors based on gene-length outliers. Corrections were made to annotations with the aid of NCBI ORFfinder (Rombel et al., 2002), except in cases where internal stop codons were known or inferred to be due to U-to-C RNA editing.

To reduce redundancy in the dataset, only a single sample was chosen from genera that were represented by more than one species, except in cases where substantial variation in ClpP1 sequence was observed among congeners. This down-sampling reduced the *clpP1* dataset to 480 species (and 483 sequences because we retained divergent *clpP1* copies found within the plastomes of *Carex siderosticta* and *Silene chalcedonica*). We used the same set of species for the *psaA* analysis, except that it did not include 16 holoparasitic species that have lost *psaA* along with most or all of their photosynthesis-related genes (Bromham et al., 2013; Krause, 2008) or the lycophyte *Selaginella moellendorffii*, which exhibits extreme levels of RNA editing (Smith, 2009), making it difficult to estimate rates of protein sequence evolution. The resulting *psaA* dataset contained 463 sequences.

In the Methods below, we will refer to the set of Viridiplantae species described above as the "large dataset" (**Table S2.2**). For some analyses, we used a more targeted sampling of 25 angiosperms, which we will refer to as the "small dataset" (**Table S2.3**). The species in the small dataset were selected to 1) span the phylogenetic diversity of angiosperms, 2) capture multiple independent accelerations in plastome evolutionary rate as well as related species with slow evolutionary rates, and 3) only include species for which nuclear genome/transcriptome resources were available (for use in subsequent analyses

of cytonuclear coevolution; see below). All scripts, sequence data, and alignments are available at https://github.com/alissawilliams/clpP1_2018.

Sequence alignment and tree construction

To assess variation in rates of ClpP1 and PsaA protein sequence evolution across green plants, the nucleotide sequences of the large dataset were translated into protein sequences in MEGA v7.0.21 (Kumar et al., 2016). These protein sequences were then aligned using the MAFFT v7.222 einsi option (Katoh and Standley, 2013). Constraint trees were manually constructed based on established phylogenetic relationships, using NCBI taxonomy and the Angiosperm Phylogeny Website v13. Branch lengths were estimated using codeml in the PAML v4.9a package (Yang, 2007) with an LG substitution matrix and rate variation among sites estimated with a gamma distribution. For this analysis, the ClpP1 alignment was trimmed to remove all insertions relative to the 196-aa *Nicotiana tabacum* reference sequence.

Analysis of substitution rate variation and tests for selection

Variation among sites. To generate site-specific estimates of amino-acid substitution rate across the ClpP1 subunit, we applied a partitioned model in codeml. Using option G, we specified a separate partition for each position in the 196-aa alignment. The Mgene parameter was set to 0 such that total tree length could vary across partitions (i.e., different sites could have different rates), but branch lengths had to remain proportional. The complexity of this model necessitated that it be run under a simple Poisson model of amino acid substitutions. To assess whether certain regions of the protein exhibited disproportionate accelerations in fast lineages, we performed this analysis on two different subsets of the large dataset. The first was a set of 27 slow "background" lineages sampled from across green plants. The second was a set of 60 angiosperms that contained 38 "accelerated" lineages and 22 interspersed slow lineages that were included to increase the probability of detecting parallel amino-acid substitutions in fast lineages (**Table S2.4**). Site-specific rate estimates were summarized using a sliding-window analysis

with a window size of 21 aa. This rate analysis was performed both on raw tree lengths and on normalized rates that were scaled to the average tree length across the data set. Site-specific variation was also summarized with WebLogo v3.5.0 (Crooks et al., 2004), using default settings and the trimmed 196-aa alignment of 483 ClpP1 sequences described above.

To further investigate rate variation within ClpP1, we partitioned the subunit into the following four regions based on characterized structural domains in *E. coli* (Yu and Houry, 2007): 1) the "handle domain", consisting of positions 129-162; 2) the "head domain", consisting of positions and 32-124 and 165-193; 3) the N-terminal "axial loops", consisting of positions 7-20 (although the alignment between *E. coli* and the plastid ClpP1 is weak in this region, so it not clear whether there is a conserved functional role); and 4) "other" spacer regions, consisting of all remaining positions in the 196-aa alignment. We repeated the codeml analysis described above but used the full 483-sequence alignment and specified these four partitions with option G. As a basis of comparison, we performed the same analysis without any partitions. To test for evidence of significant rate heterogeneity among the four regions, we performed a likelihood ratio test (LRT) that compared these two models with three degrees of freedom.

Variation among branches. To examine differences in d_N/d_S across species, we used codeml to determine d_N and d_S (for both *clpP1* and *psaA*) for each branch of the small dataset. We converted our previously obtained ClpP1 and PsaA amino-acid alignments into codon-based nucleotide alignments. In the codeml run, we used the parameters model=1 and fix_omega=0, which together specify estimation of an individual d_N/d_S value for each branch in the tree. For terminal branches with a d_N/d_S estimate > 1, we assessed statistical significance by constraining each branch of interest (separately) to a d_N/d_S value of 1 (model=2, fix_omega=1). We determined whether the unconstrained PAML model was a significantly better fit to the data than each constrained model by performing LRTs with one degree of freedom. To summarize these data with boxplots, species in the small dataset were partitioned into one of two categories based on fast vs. slow rate of ClpP1 evolution. The fast species are those with a cumulative

distance (branch length) from the root of the 25-species angiosperm tree of at least 0.44 amino-acid substitutions per site, an arbitrary cutoff.

Branch-site models. Using the codon-based alignment of *clpP1* for the small dataset (described above), we used a codeml branch-site model (Zhang et al., 2005) to infer whether any amino acid sites in ClpP1 have been subject to positive selection. We used the same partition of fast and slow species as described above. The fast species were specified as the foreground, which means that the analysis identified sites under positive selection in this species subset. This analysis used the parameters model=2, NSsites=2, fix_omega=1, and omega=1 for the null model, and the parameters model=2, NSsites=2, fix_omega=0, and omega=1 for the alternative model. To determine whether the alternative model was a significantly better fit to the data than the null model, we performed an LRT with one degree of freedom. We mapped sites with a Bayes empirical Bayes (BEB) posterior probability of ≥ 0.95 for positive selection based on this analysis to the homologous positions in the *E. coli* ClpP structure (Protein Data Bank 1YG6) visualized in Chimera v1.11.2 (Pettersen et al., 2004).

Evolutionary rate covariation between the plastid- and nuclear-encoded Clp subunits.

To determine whether the rate of amino-acid substitution in ClpP1 is correlated with the rate in nuclear-encoded Clp subunits across angiosperms, we compiled protein sequences of all nine core Clp subunits (ClpP1,3-6, ClpR1-4) and 20 non-Clp nuclear-encoded genes for each species in the small dataset, using a custom Python script to reciprocally blast *A. thaliana* protein sequences against predicted protein sequences from each of the other species. Predicted protein sequence data were collected from various sources, including sequenced genomes on Phytozome (https://phytozome.jgi.doe.gov/), transcriptomes from the 1KP project (Wickett et al., 2014), and various other transcriptome sequencing projects (**Table S2.3**). The 20 non-Clp genes represent the subset of the 50 control genes examined by Rockenbach *et al*. (2016) for which we could recover orthologs in all species of interest. Protein sequences were aligned with MAFFT (Katoh and Standley, 2013) and trimmed at the N- and C-terminal

ends as needed (in cases of poor alignment) to avoid overestimation of branch lengths. In rare cases, internal trimming was required for the same reason. In two cases, nuclear sequences from one species were paired with a ClpP1 sequence from a closely related species (*Acacia aulacocarpa* was paired with *A. ligulata*, and *Mimulus guttatus* was paired with *Erythranthe lutea*) due to the lack of a published plastome (**Table S2.3**).

We estimated branch lengths both for individual proteins and for sets of concatenated amino-acid sequences using codeml with the LG substitution matrix and a gamma distribution. The concatenated sets of sequences were 1) all nuclear-encoded core Clp proteins, 2) all nuclear-encoded non-Clp proteins, and 3) two randomly divided halves of the 20 nuclear-encoded non-Clp proteins (10 proteins each). We used correlation analysis to compare the branch lengths of ClpP1 and the concatenated nuclear-encoded Clp proteins, each normalized by dividing by the branch length of one set of concatenated non-Clp proteins in that species. By using the two different halves of the dataset for normalization, we avoided introducing statistical non-independence between our variables in the correlation analysis. Only terminal branches were used in the correlation analysis, with the exception of the grasses (*Oryza sativa* and *Sorghum bicolor*). In that case, ClpP1 acceleration occurred before the split between the two species; thus, the branch leading to the grasses was used in place of the two terminal branches of those species. We log-transformed the normalized branch lengths for ClpP1 and the concatenation of nuclear-encoded Clp subunits and calculated the Pearson correlation coefficient across branches with R v3.4.1.

In order to specifically control for background rates of evolution in the plastid genome, we repeated this statistical analysis using a concatenated set of 44 plastid-encoded photosynthetic proteins to normalize ClpP1 rates. These 44 photosynthetic protein sequences were extracted from GenBank plastomes of the 25 species using a custom Perl script. Alignment and branch-length estimates were generated as described above. For this analysis, we used the concatenation of all 20 nuclear-encoded non-Clp proteins to normalize the nuclear-encoded Clp rates.

Analysis of indels

To determine the relationship between rates of amino-acid substitution and rates of indels in *clpP1*, we used the codon-based alignment for the small dataset. Indels were coded using the modified complex coding option in SeqState v1.0 (Müller, 2005). The indel data were visualized in Mesquite v3.31 using "Trace Character History → Parsimony Ancestral States." This visualization plotted indel states on each branch of the constraint tree at each indel site. Using these plots, we counted the number of novel indels on each branch of the tree. In this case, all novel indels occurred on terminal branches. Correlation analysis was performed similarly to the plastid-nuclear Clp correlation described above. Once again, the substitution-based branch lengths for ClpP1 were normalized with one concatenated set of non-Clp sequences. The branch-specific ClpP1 indel counts were normalized with the branch lengths from the other concatenated set of non-Clp sequences. We tested for a significant association between these normalized rates of ClpP1 sequence and structural evolution, using a Spearman Rank Correlation analysis in R. We completed this same statistical analysis using two independent sets of plastid-encoded photosynthetic proteins (n = 22 each); these sets were the result of division of the photosynthetic protein set (n = 44) described above.

Correlation between ClpP1 evolutionary rates and character states

To determine whether accelerated ClpP1 evolutionary rates is correlated with *clpP1* duplication, the loss of RNA editing sites, the loss of introns, and substitutions in the catalytic triad, we used binary phylogenetic generalized linear mixed models to account for pseudoreplication due to shared phylogenetic history (Ives and Garland, 2014). These tests were implemented with the binaryPGLMM function in the R ape package and applied to the angiosperm species from the large dataset. The independent variable for these tests was the cumulative branch length to the root of the 483-species Viridiplantae ClpP1 tree. We tested for a significant relationship between this rate variable and each binary independent variable: the presence/absence of *clpP1* duplications, each intron, each RNA editing site, and each catalytic residue. Our input tree for this analysis was the angiosperm portion of the PsaA

evolutionary rate tree (**Figure 2.2**), so the small number of angiosperms that lack a PsaA sequence were excluded.

Proteome analysis of Clp core complexes in *Silene species*

Tissue collection and chloroplast isolation. A total of 60 g of leaf tissue was collected from mature rosettes from four *S. noctiflora* individuals and from ten *S. latifolia* individuals. The *S. noctiflora* individuals were derived from the BRP line previously used for mitochondrial genome sequencing (Wu et al., 2015), and the *S. latifolia* individuals were derived from the line previously used for plastid genome sequencing (Sloan et al., 2012b). Plants were grown in Fafard 2SV Mix supplemented with vermiculite and perlite in the Colorado State University greenhouses with supplemental lighting on a 16/8-hr light/dark cycle and regular watering and fertilization. Chloroplasts were isolated following a protocol based on van Wijk et al. (van Wijk et al., 2007). In brief, rinsed leaf tissue was disrupted with a blender in 100 ml of grinding buffer for each 10 g of tissue (50 mM HEPES-KOH pH 8.0, 330 mM sorbitol, 2 mM EDTA, 5 mM ascorbic acid, 5 mM cysteine, 0.05% BSA) and filtered through two layers of Miracloth. The resulting samples were centrifuged at 1300 x g for 4 min in a fixed-angle rotor. Pellets were resuspended in wash buffer (50 mM HEPES-KOH pH 8.0, 330 mM sorbitol, 2 mM EDTA), loaded onto 40%/85% Percoll step gradients, and centrifuged at 3750 x g for 10 min in a swinging-bucket rotor. Intact chloroplasts were harvested from the interface of the step-gradient, diluted in wash buffer, and centrifuged at 1300 x g for 3 min in a fixed-angle rotor. The resulting chloroplast pellets were flash frozen in liquid nitrogen and stored at -80 °C. All isolation steps were performed at 4 °C under dim green light.

Isolation of stromal protein fractions and native polyacrylamide gel electrophoresis (Native PAGE). Frozen chloroplast pellets were resuspended in HEPES 50 mM pH 8, $MgCl_2$ 10 mM, glycerol 15% and protease inhibitors, and the soluble stromal proteomes were isolated from the resuspended, broken chloroplasts by centrifugation at 100,000 x g for 30 min at 4 °C. The supernatant containing the chloroplast soluble proteomes were concentrated by Amicon 10 kDa filter units and proteins were

quantified by BCA Protein Assay Kit (Thermo Fisher). For light-blue native PAGE analysis, 50 µg of stromal protein of each species in 50 mM BisTris-HCl, 50 mM NaCl, 10% w/v glycerol and 0.001% Ponceau S (pH 7.2) was loaded per lane using the NativePage Novex gel system with precast 4 - 16% acrylamide Bis-Tris gels (Invitrogen). The upper buffer contained 0.002% Coomassie G. A total of three lanes were loaded for each species**.**

Mass spectrometry (MS) and Data analysis. Each gel lane was cut into 19 bands followed by reduction, alkylation, and in-gel digestion with trypsin as described in (Friso et al., 2011; Shevchenko et al., 2006). For replicate 1, we used one gel lane for each species, whereas we pooled two gel lanes for replicate 2 to increase protein identifications and sequence coverage. The resuspended peptide extracts were analyzed by data-dependent MS/MS using an on-line LC-LTQ-Orbitrap (Thermo Electron Corp.) with details as described in (Kim et al., 2015). Hence, a total of 38 MS/MS runs for each species was carried out. MS data searching against assembled databases for *S. noctiflora* (88,166 sequences; 20,816,406 residues) and *S. latifolia* (101,108 sequences; 20,447,864 residues) was done using Mascot, followed by filtering, grouping of closely related sequences based on matched MS/MS spectra and quantification based on normalized AdjSPC (NadjSPC) as in (Friso et al., 2011; Kim et al., 2015). For each species, databases were a merger of annotated proteins from organellar genomes (Sloan et al., 2012a, 2012b) and protein sequence predictions generated by TransDecoder (Haas et al., 2013) from transcriptome assemblies (Sloan et al., 2014b). Protein annotations are based on homology to Arabidopsis and taken from the Plant Proteome Data Base (http://ppdb.tc.cornell.edu/).

To determine the assembly state of ClpP1 and other ClpPR subunits, native masses were calibrated with endogenous stromal complexes for which we and/or others previously determined the native mass in Arabidopsis, in particular CPN60 (800 kDa), RUBISCO holocomplex (550 kDa), glutamate-ammonia ligase (GS2; 240 kDa), CLPC/D (200 kDa), transketolase-1 (TKL-1; 150 kDa), thiazole biosynthetic enzyme 1 (THI1; 245 kDa), and metalloprotease PREP1/2 (110 kDa) (see (Peltier et al., 2006) and references therein).

.

Table 2.1. Examples of *clpP1* loss or putative pseudogenization

Genus or species	Classification	*clpP1* status	Evidence
Bathycoccus prasinos	Green alga	Lost	
Boodlea composita	Green alga	Lost (possibly transferred to nucleus)	
Helicosporidium sp.	Holoparasitic green alga	Lost	
Pilostyles aethiopica	Holoparasitic angiosperm	Lost	
Pilostyles hamiltonii	Holoparasitic angiosperm	Lost	
Hydnora visseri	Holoparasitic angiosperm	Lost	
Sciaphila thaidanica	Mycoheterotrophic angiosperm	Lost	
Epipremnum aureum	Angiosperm	Apparent pseudogenization	Large truncation/structural change
Hanabusaya asiatica	Angiosperm	Apparent pseudogenization	Loss of first exon
Phelipanche purpurea	Holoparasitic angiosperm	Apparent pseudogenization	Loss of first exon
Actinidia (except *tetramera*)	Angiosperm	Apparent pseudogenization	Loss of first exon

*note: *Scaevola* has also been suggested to lack *clpP1* (Jansen et al., 2007), but does not have a complete plastome assembly.

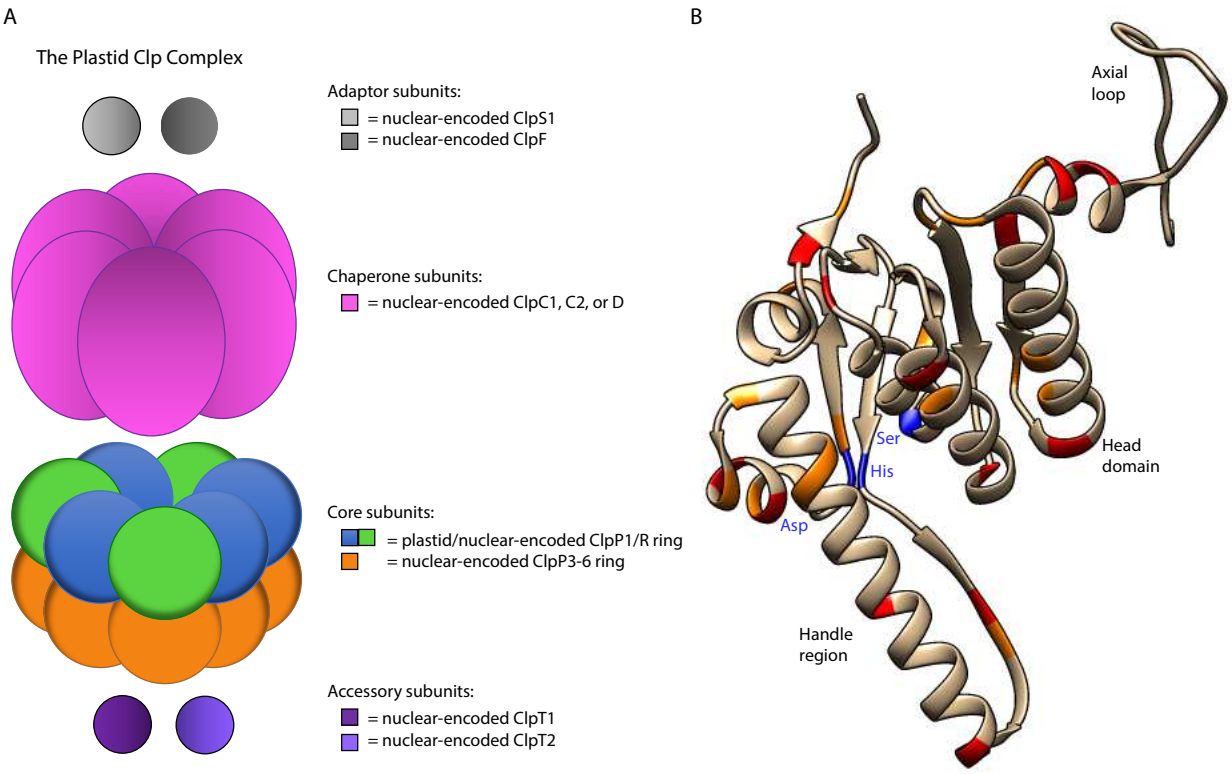


Figure 2.1: Depiction of the plastid Clp complex. A) Elements of the plastid Clp complex. Adaptor subunits deliver substrates to the homohexameric chaperone, which uses ATP to unfold them into the Clp core. The Clp core consists of two heptameric rings with different compositions, which stack together to form a barrel shape. In plants, the ClpT subunits associate with the ClpPR core. Figure adapted from Nishimura and van Wijk, 2015. B) A single plastid Clp core subunit, as mapped onto the *E. coli* ClpP structure (PDB accession 1YG6). The three catalytic sites of ClpP1 are colored in blue. The head domain is involved in intra-ring interactions, the handle domain is involved in inter-ring interactions, and the axial loop is thought to play a role in binding to non-core subunits. The red and orange spots represent amino acid residues under selection in ClpP1, as determined via a PAML branch-site analysis across accelerated species. Residues with posterior probabilities between 0.95 and 0.99 are colored in orange, and residues with posterior probabilities of 0.99 and above are colored in red.

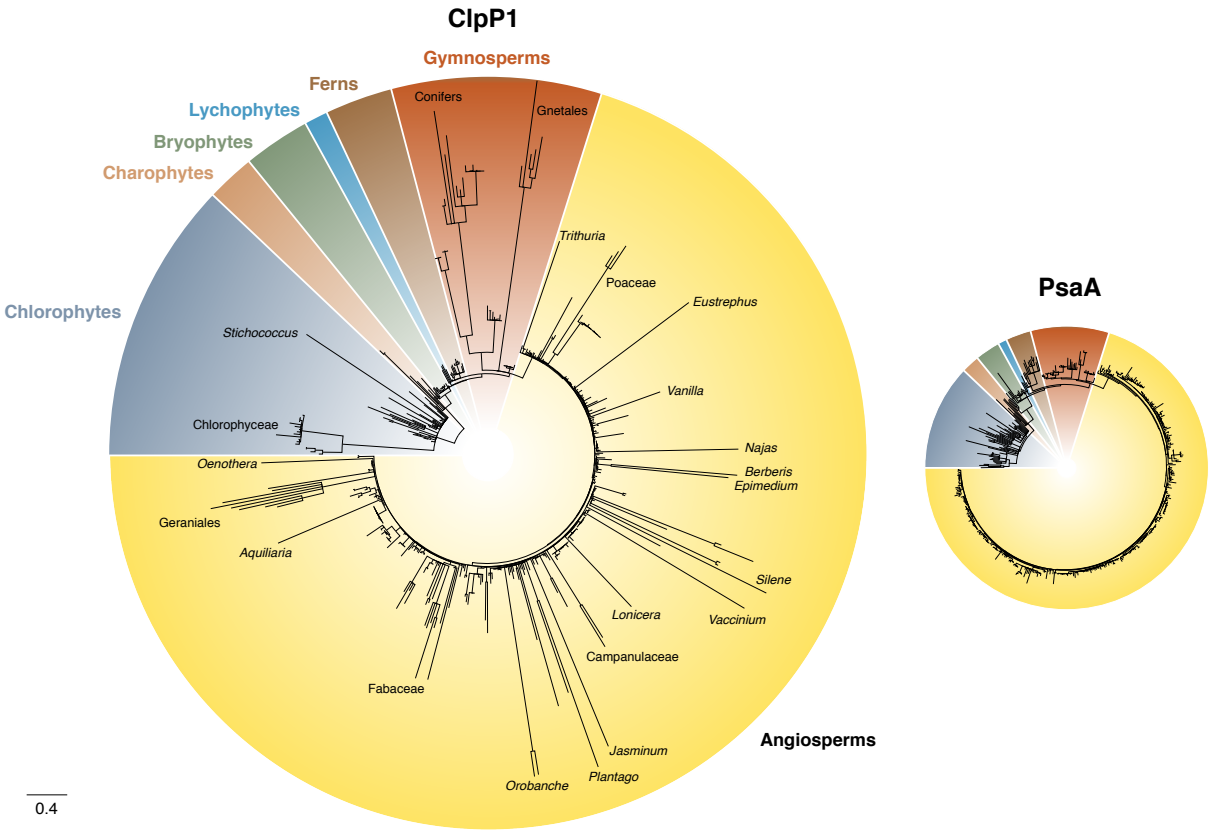
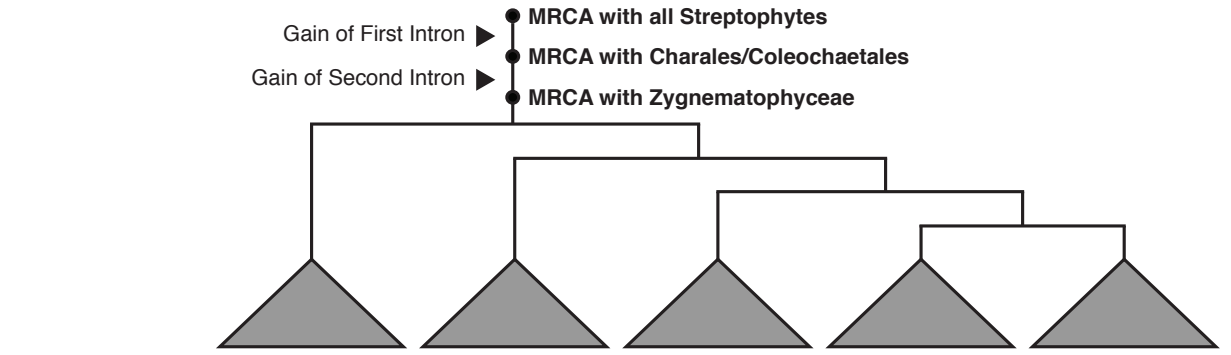


Figure 2.2: Comparison of evolutionary rates between ClpP1 and PsaA across green plants. Branch lengths represent amino acid substitutions per site. The species sampling between the trees is nearly identical (see Main Text for description of differences). Taxon names are included for select "fast" branches in the ClpP1 tree. See Figure S1 for a tree with full species labeling.



Intron Loss Events	Bryophytes (14)	Lycophytes (3)	Ferns (18)	Gymnosperms (73)	Angiosperms (795)
• First Intron Only	0	0	0	0	3
• Second Intron Only	0	1	1	0	9
• Both Introns	0	0	0	1	17

Figure 2.3: Intron gain and loss in *clpP1* across green plants. Number of species sampled is included parenthetically for each group. Columns contain the number of each type of loss in each group. MRCA, most recent common ancestor.

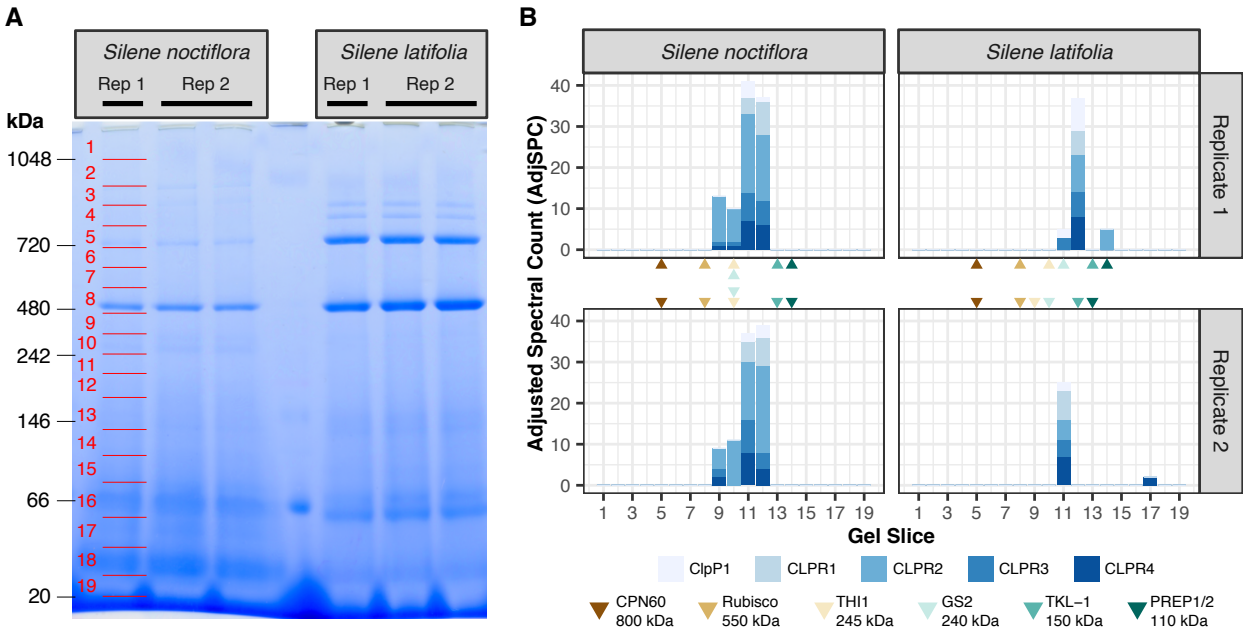


Figure 2.4: Native-gel and mass spectrometry analysis of *Silene* plastid proteins. A) LB-Native-PAGE performed on stromal protein fraction from *S. noctiflora* and *S. latifolia*. Red lines indicate approximate positions of gel slices for MS/MS analysis, but native masses were more finely calibrated with known stromal complexes. B) AdjSPC for subunits of the ClpP1/R ring, including ClpP1. Triangles indicate the gel slice corresponding to peak detection for native complexes used for internal calibration. For more details see Figure S7 and Supplemental Dataset 1.

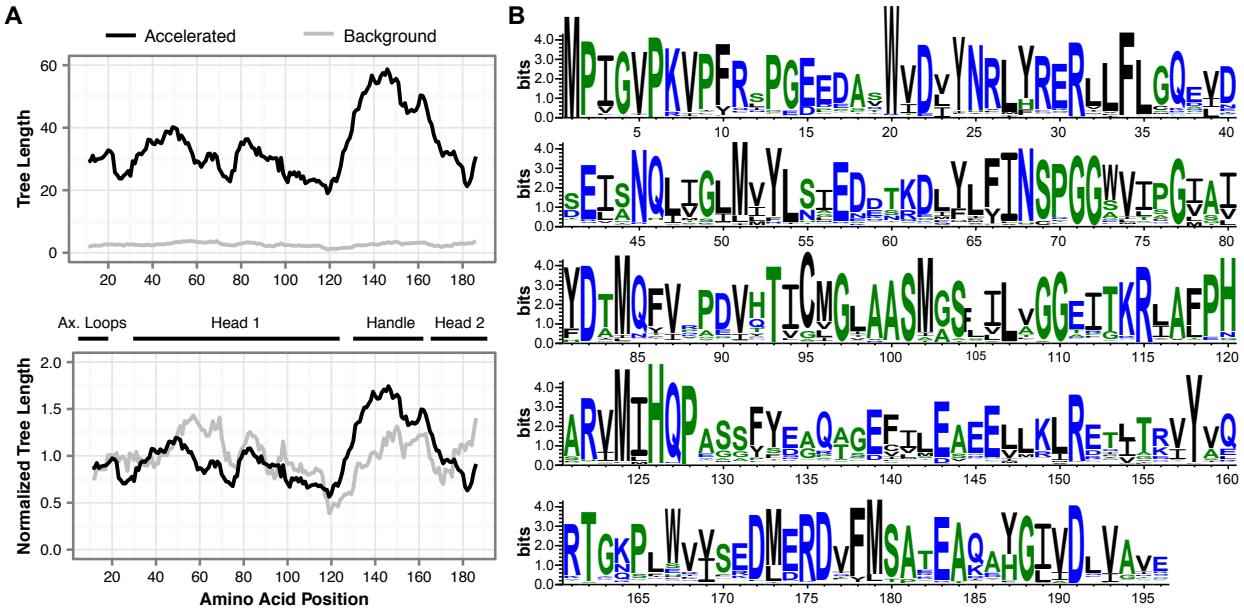


Figure 2.5: Rate variation across ClpP1. a) Sliding window analysis of rate variation across a diverse subsample of angiosperms, using a window size of 21 aa and tree length measured as amino acid substitutions per site. Normalized tree lengths (bottom plot) were calculated by dividing each window by the average tree length of the entire protein. b) WebLogo representation of sequence conservation across green plants. The size of the letter at each amino acid position is indicative of the level of conservation.

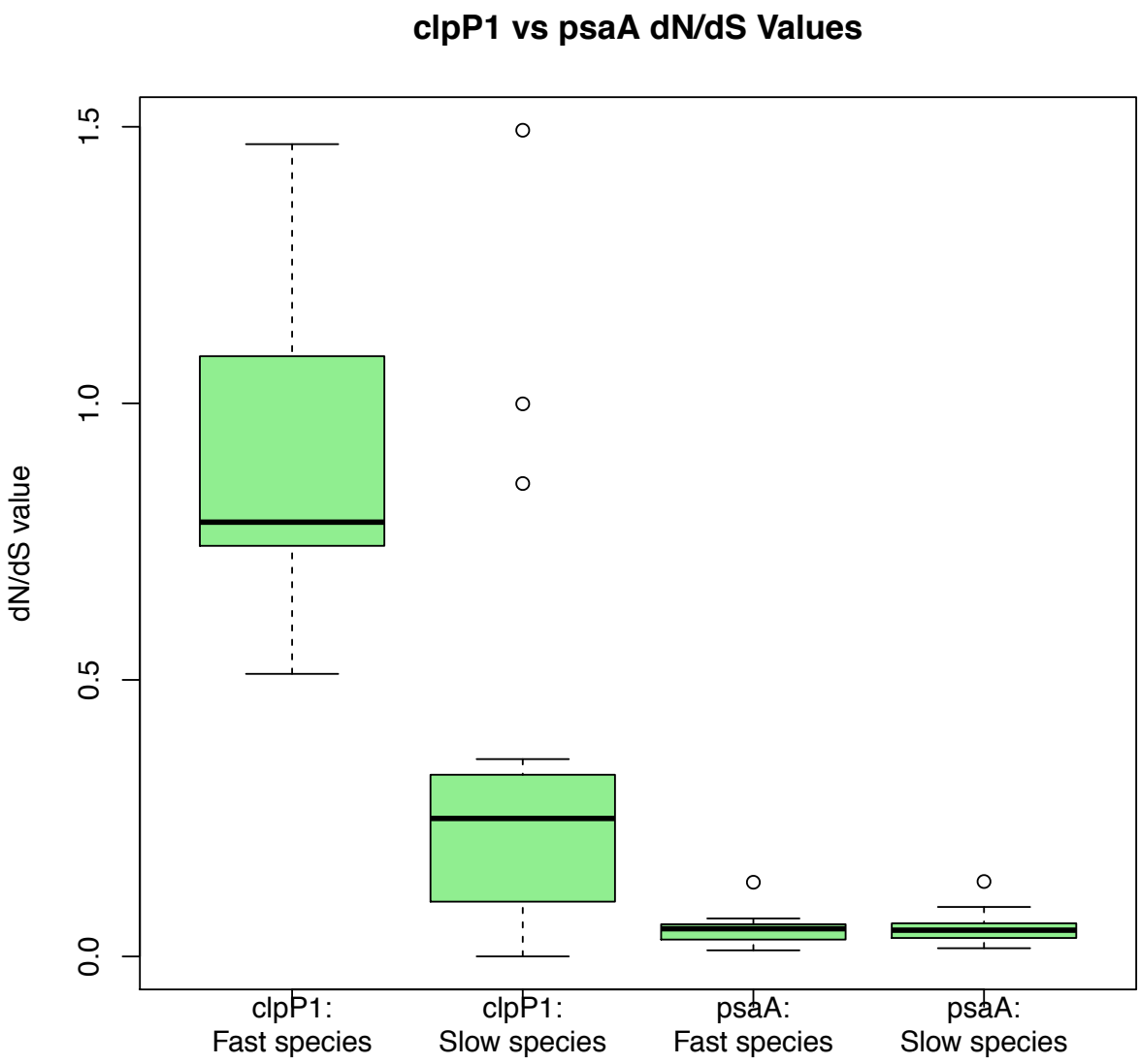


Figure 2.6: d_N/d_S values for *clpP1* and *psaA* in a sample of 25 angiosperms; shown are the values obtained for terminal branches only (see Table S1). Species were designated as "slow" or "fast" based on ClpP1 amino acid substitution rates. The top and bottom of each box represent the upper and lower quartiles, respectively. The line contained within the box represents the median. The dotted lines connect the full range of points, apart from outliers, which are represented by dots.

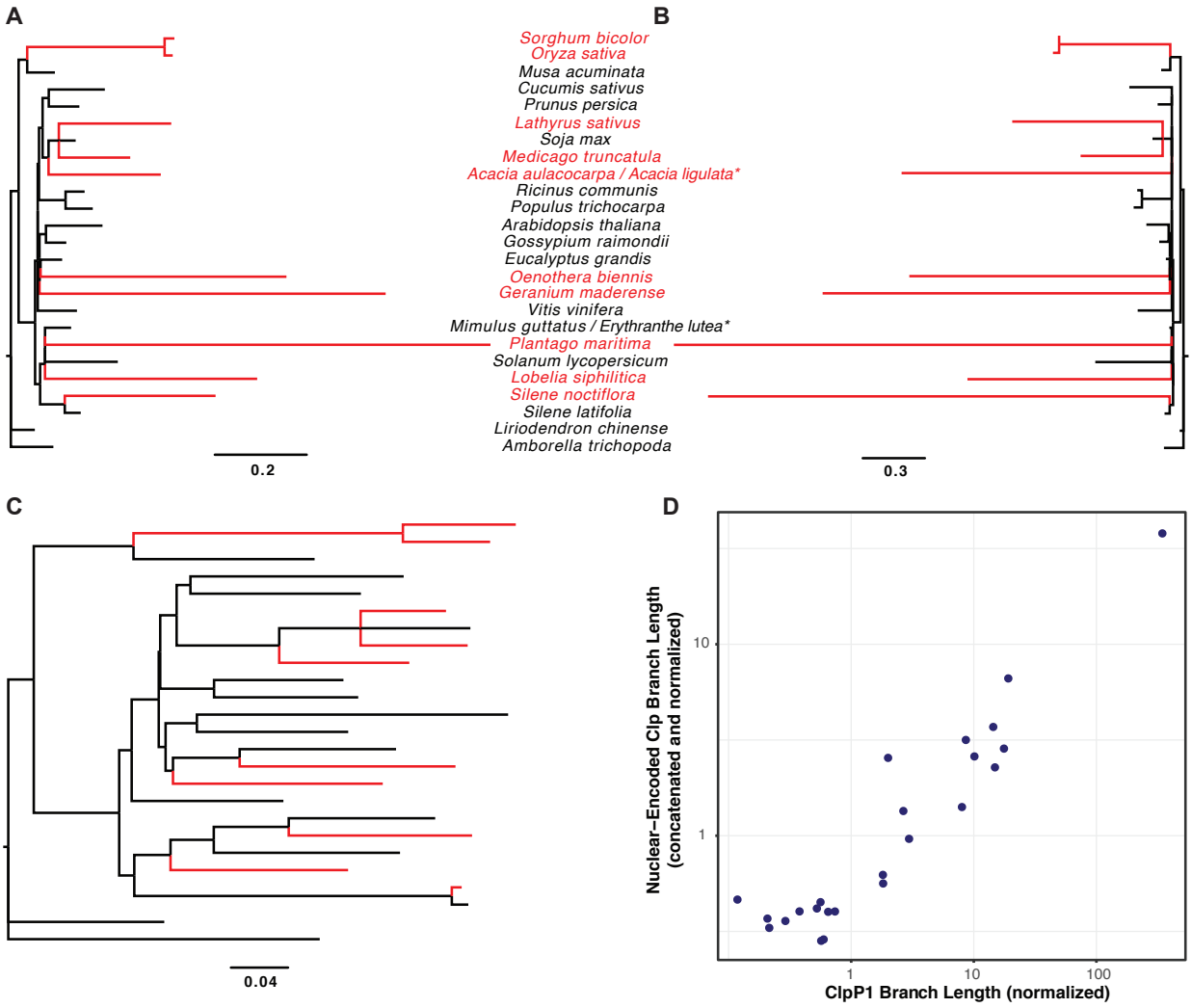


Figure 2.7: Comparison of evolutionary rates of ClpP1, the nuclear-encoded core plastid Clp subunits, and non-Clp-related nuclear-encoded genes. Fast species, as defined by distance to the root, are indicated in red. Branch lengths represent amino acid substitutions per site. A) Nuclear-encoded core plastid Clp subunits (n=8, concatenated), B) ClpP1, C) non-Clp nuclear-encoded proteins (n=20, concatenated), D) Scatterplot comparison of branch lengths. Normalization of both axes was achieved using independent sets of non-Clp nuclear-encoded proteins (n=10 each, concatenated).

LITERATURE CITED

Adams, K.L., Daley, D.O., Whelan, J., Palmer, J.D., 2002a. Genes for two mitochondrial ribosomal proteins in flowering plants are derived from their chloroplast or cytosolic counterparts. Plant Cell 14, 931–943.

Adams, K.L., Qiu, Y.-L., Stoutemyer, M., Palmer, J.D., 2002b. 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. 99, 9905–9912. https://doi.org/10.1073/pnas.042694899

Andersson, F.I., Tryggvesson, A., Sharon, M., Diemand, A.V., Classen, M., Best, C., Schmidt, R., Schelin, J., Stanne, T.M., Bukau, B., Robinson, C.V., Witt, S., Mogk, A., Clarke, A.K., 2009. Structure and function of a novel type of ATP-dependent Clp protease. J. Biol. Chem. 284, 13519–13532. https://doi.org/10.1074/jbc.M809588200

Apitz, J., Nishimura, K., Schmied, J., Wolf, A., Hedtke, B., Wijk, K.J. van, Grimm, B., 2016. Posttranslational Control of ALA Synthesis Includes GluTR Degradation by Clp Protease and Stabilization by GluTR-Binding Protein. Plant Physiol. 170, 2040–2051. https://doi.org/10.1104/pp.15.01945

Barnard-Kubow, K.B., Sloan, D.B., Galloway, L.F., 2014. Correlation between sequence divergence and polymorphism reveals similar evolutionary mechanisms acting across multiple timescales in a rapidly evolving plastid genome. BMC Evol. Biol. 14. https://doi.org/10.1186/s12862-014-0268-y

Blanchard, J.L., Lynch, M., 2000. Organellar genes: why do they end up in the nucleus? Trends Genet. 16, 315–320. https://doi.org/10.1016/S0168-9525(00)02053-9

Blazier, J.C., Ruhlman, T.A., Weng, M.-L., Rehman, S.K., Sabir, J.S.M., Jansen, R.K., 2016. Divergence of RNA polymerase α subunits in angiosperm plastid genomes is mediated by genomic rearrangement. Sci. Rep. 6, 24595. https://doi.org/10.1038/srep24595

Bogdanova, V.S., Zaytseva, O.O., Mglinets, A.V., Shatskaya, N.V., Kosterin, O.E., Vasiliev, G.V., 2015. Nuclear-Cytoplasmic Conflict in Pea (Pisum sativum L.) Is Associated with Nuclear and Plastidic Candidate Genes Encoding Acetyl-CoA Carboxylase Subunits. PLOS ONE 10, e0119835. https://doi.org/10.1371/journal.pone.0119835

Bromham, L., Cowman, P.F., Lanfear, R., 2013. Parasitic plants have increased rates of molecular evolution across all three genomes. BMC Evol. Biol. 13, 126. https://doi.org/10.1186/1471-2148-13-126

Clark, N.L., Aquadro, C.F., 2010. A novel method to detect proteins evolving at correlated rates: identifying new functional relationships between coevolving proteins. Mol. Biol. Evol. 27, 1152–1161. https://doi.org/10.1093/molbev/msp324

Clarke, A.K., MacDonald, T.M., Sjögren, L.L.E., 2005. The ATP-dependent Clp protease in chloroplasts of higher plants. Physiol. Plant. 123, 406–412. https://doi.org/10.1111/j.1399-3054.2005.00452.x

Cortona, A.D., Leliaert, F., Bogaert, K.A., Turmel, M., Boedeker, C., Janouškovec, J., Lopez-Bautista, J.M., Verbruggen, H., Vandepoele, K., Clerck, O.D., 2017. The Plastid Genome in Cladophorales Green Algae Is Encoded by Hairpin Chromosomes. Curr. Biol. 27, 3771-3782.e6. https://doi.org/10.1016/j.cub.2017.11.004

Crooks, G.E., Hon, G., Chandonia, J.-M., Brenner, S.E., 2004. WebLogo: A Sequence Logo Generator. Genome Res. 14, 1188–1190. https://doi.org/10.1101/gr.849004

De Maio, N., Holmes, I., Schlötterer, C., Kosiol, C., 2013. Estimating empirical codon hidden Markov models. Mol. Biol. Evol. 30, 725–736. https://doi.org/10.1093/molbev/mss266

Delannoy, E., Fujii, S., Colas des Francs-Small, C., Brundrett, M., Small, I., 2011. Rampant Gene Loss in the Underground Orchid Rhizanthella gardneri Highlights Evolutionary Constraints on Plastid Genomes. Mol. Biol. Evol. 28, 2077–2086. https://doi.org/10.1093/molbev/msr028

Derrien, B., Majeran, W., Effantin, G., Ebenezer, J., Friso, G., Wijk, K.J. van, Steven, A.C., Maurizi, M.R., Vallon, O., 2012. The purification of the <Emphasis Type="Italic">Chlamydomonas

reinhardtii</Emphasis> chloroplast ClpP complex: additional subunits and structural features. Plant Mol. Biol. 80, 189–202. https://doi.org/10.1007/s11103-012-9939-5

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. https://doi.org/10.1016/j.ympev.2008.09.009

Duff, R.J., Moore, F.B.-G., 2005. Pervasive RNA Editing Among Hornwort <Emphasis Type="Italic">rbc</Emphasis>L Transcripts Except <Emphasis Type="Italic">Leiosporoceros</Emphasis>. J. Mol. Evol. 61, 571–578. https://doi.org/10.1007/s00239-004-0146-0

Dutheil, J., Galtier, N., 2007. Detecting groups of coevolving positions in a molecule: a clustering approach. BMC Evol. Biol. 7, 242. https://doi.org/10.1186/1471-2148-7-242

El Bakkouri, M., Rathore, S., Calmettes, C., Wernimont, A.K., Liu, K., Sinha, D., Asad, M., Jung, P., Hui, R., Mohmmed, A., Houry, W.A., 2013. Structural insights into the inactive subunit of the apicoplast-localized caseinolytic protease complex of Plasmodium falciparum. J. Biol. Chem. 288, 1022–1031. https://doi.org/10.1074/jbc.M112.416560

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. https://doi.org/10.1371/journal.pone.0001386

Fajardo, D., Senalik, D., Ames, M., Zhu, H., Steffan, S.A., Harbut, R., Polashock, J., Vorsa, N., Gillespie, E., Kron, K., Zalapa, J.E., 2013. Complete plastid genome sequence of <Emphasis Type="Italic">Vaccinium macrocarpon</Emphasis>: structure, gene content, and rearrangements revealed by next generation sequencing. Tree Genet. Genomes 9, 489–498. https://doi.org/10.1007/s11295-012-0573-9

Fedhila, S., Msadek, T., Nel, P., Lereclus, D., 2002. Distinct clpP Genes Control Specific Adaptive Responses in Bacillus thuringiensis. J. Bacteriol. 184, 5554–5562. https://doi.org/10.1128/JB.184.20.5554-5562.2002

Freyer, R., Kiefer-Meyer, M.-C., Kössel, H., 1997. Occurrence of plastid RNA editing in all major lineages of land plants. Proc. Natl. Acad. Sci. 94, 6285–6290.

Friso, G., Olinares, P.D.B., van Wijk, K.J., 2011. The workflow for quantitative proteome analysis of chloroplast development and differentiation, chloroplast mutants, and protein interactions by spectral counting. Methods Mol. Biol. Clifton NJ 775, 265–282. https://doi.org/10.1007/978-1-61779-237-3_14

Gertz, E.M., Yu, Y.-K., Agarwala, R., Schäffer, A.A., Altschul, S.F., 2006. Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. BMC Biol. 4, 41. https://doi.org/10.1186/1741-7007-4-41

Guisinger, M.M., Chumley, T.W., Kuehl, J.V., Boore, J.L., Jansen, R.K., 2010. Implications of the Plastid Genome Sequence of <Emphasis Type="Italic">Typha</Emphasis> (Typhaceae, Poales) for Understanding Genome Evolution in Poaceae. J. Mol. Evol. 70, 149–166. https://doi.org/10.1007/s00239-009-9317-3

Guisinger, M.M., Kuehl, J.V., Boore, J.L., Jansen, R.K., 2008. Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. Proc. Natl. Acad. Sci. 105, 18424–18429. https://doi.org/10.1073/pnas.0806759105

Guo, W., Grewe, F., Fan, W., Young, G.J., Knoop, V., Palmer, J.D., Mower, J.P., 2016. Ginkgo and Welwitschia Mitogenomes Reveal Extreme Contrasts in Gymnosperm Mitochondrial Evolution. Mol. Biol. Evol. 33, 1448–1460. https://doi.org/10.1093/molbev/msw024

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N., Regev, A., 2013. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat. Protoc. 8. https://doi.org/10.1038/nprot.2013.084

Haberle, R.C., Fourcade, H.M., Boore, J.L., Jansen, R.K., 2008. Extensive rearrangements in the chloroplast genome of Trachelium caeruleum are associated with repeats and tRNA genes. J. Mol. Evol. 66, 350–361. https://doi.org/10.1007/s00239-008-9086-4

Havird, J.C., Sloan, D.B., 2016. The Roles of Mutation, Selection, and Expression in Determining Relative Rates of Evolution in Mitochondrial versus Nuclear Genomes. Mol. Biol. Evol. 33, 3042–3053. https://doi.org/10.1093/molbev/msw185

Havird, J.C., Trapp, P., Miller, C.M., Bazos, I., Sloan, D.B., 2017. Causes and Consequences of Rapidly Evolving mtDNA in a Plant Lineage. Genome Biol. Evol. 9, 323–336. https://doi.org/10.1093/gbe/evx010

Havird, J.C., Whitehill Nicholas S., Snow Christopher D., Sloan Daniel B., 2015. Conservative and compensatory evolution in oxidative phosphorylation complexes of angiosperms with highly divergent rates of mitochondrial genome evolution. Evolution 69, 3069–3081. https://doi.org/10.1111/evo.12808

Hein, A., Knoop, V., 2018. Expected and unexpected evolution of plant RNA editing factors CLB19, CRR28 and RARE1: retention of CLB19 despite a phylogenetically deep loss of its two known editing targets in Poaceae. BMC Evol. Biol. 18, 85. https://doi.org/10.1186/s12862-018-1203-4

Hirao, T., Watanabe, A., Kurita, M., Kondo, T., Takata, K., 2008. Complete nucleotide sequence of the Cryptomeria japonicaD. Don. chloroplast genome and comparative chloroplast genomics: diversified genomic structure of coniferous species. BMC Plant Biol. 8, 70. https://doi.org/10.1186/1471-2229-8-70

Huang, C., Wang, S., Chen, L., Lemieux, C., Otis, C., Turmel, M., Liu, X.-Q., 1994. The <Emphasis Type="Italic">Chlamydomonas</Emphasis> chloroplast <Emphasis Type="Italic">clpP</Emphasis> gene contains translated large insertion sequences and is essential for cell growth. Mol. Gen. Genet. MGG 244, 151–159. https://doi.org/10.1007/BF00283516

Ives, A.R., Garland, T., 2014. Phylogenetic Regression for Binary Dependent Variables, in: Garamszegi, L.Z. (Ed.), Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 231–261. https://doi.org/10.1007/978-3-662-43550-2_9

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., dePamphilis, C.W., Leebens-Mack, J., Müller, K.F., Guisinger-Bellian, M., Haberle, R.C., Hansen, A.K., Chumley, T.W., Lee, S.-B., Peery, R., McNeal, J.R., Kuehl, J.V., Boore, J.L., 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc. Natl. Acad. Sci. 104, 19369–19374. https://doi.org/10.1073/pnas.0709121104

Juan, D. de, Pazos, F., Valencia, A., 2013. Emerging methods in protein co-evolution. Nat. Rev. Genet. 14, 249–261. https://doi.org/10.1038/nrg3414

Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30, 772–780. https://doi.org/10.1093/molbev/mst010

Kim, J., Kimber, M.S., Nishimura, K., Friso, G., Schultz, L., Ponnala, L., Wijk, K.J. van, 2015. Structures, Functions, and Interactions of ClpT1 and ClpT2 in the Clp Protease System of Arabidopsis Chloroplasts. Plant Cell 27, 1477–1496. https://doi.org/10.1105/tpc.15.00106

Kim, J., Olinares, P.D., Oh, S., Ghisaura, S., Poliakov, A., Ponnala, L., Wijk, K.J. van, 2013. Modified Clp Protease Complex in the ClpP3 Null Mutant and Consequences for Chloroplast Development and Function in Arabidopsis. Plant Physiol. 162, 157–179. https://doi.org/10.1104/pp.113.215699

Kim, J., Rudella, A., Rodriguez, V.R., Zybailov, B., Olinares, P.D.B., Wijk, K.J. van, 2009. Subunits of the Plastid ClpPR Protease Complex Have Differential Contributions to Embryogenesis, Plastid Biogenesis, and Plant Development in Arabidopsis. Plant Cell 21, 1669–1692. https://doi.org/10.1105/tpc.108.063784

Knie, N., Grewe, F., Fischer, S., Knoop, V., 2016. Reverse U-to-C editing exceeds C-to-U RNA editing in some ferns – a monilophyte-wide comparison of chloroplast and mitochondrial RNA editing

suggests independent evolution of the two processes in both organelles. BMC Evol. Biol. 16, 134. https://doi.org/10.1186/s12862-016-0707-z

Knox, E.B., 2014. The dynamic history of plastid genomes in the Campanulaceae sensu lato is unique among angiosperms. Proc. Natl. Acad. Sci. 111, 11097–11102. https://doi.org/10.1073/pnas.1403363111

Koussevitzky, S., Stanne, T.M., Peto, C.A., Giap, T., Sjögren, L.L.E., Zhao, Y., Clarke, A.K., Chory, J., 2007. An <Emphasis Type="Italic">Arabidopsis thaliana</Emphasis> virescent mutant reveals a role for ClpR1 in plastid development. Plant Mol. Biol. 63, 85–96. https://doi.org/10.1007/s11103-006-9074-2

Krause, K., 2008. From chloroplasts to "cryptic" plastids: evolution of plastid genomes in parasitic plants. Curr. Genet. 54, 111. https://doi.org/10.1007/s00294-008-0208-8

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

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol. Biol. Evol. 33, 1870–1874. https://doi.org/10.1093/molbev/msw054

Kuroda, H., Maliga, P., 2003. The plastid *clpP1* protease gene is essential for plant development. Nature 425, 86. https://doi.org/10.1038/nature01909

Lemieux, C., Vincent, A.T., Labarre, A., Otis, C., Turmel, M., 2015. Chloroplast phylogenomic analysis of chlorophyte green algae identifies a novel lineage sister to the Sphaeropleales (Chlorophyceae). BMC Evol. Biol. 15, 264. https://doi.org/10.1186/s12862-015-0544-5

Li, J., Zhang, Z., Vang, S., Yu, J., Wong, G.K.-S., Wang, J., 2009. Correlation between Ka/Ks and Ks is related to substitution model and evolutionary lineage. J. Mol. Evol. 68, 414–423. https://doi.org/10.1007/s00239-009-9222-9

Li, Y., Zhang, R., Liu, S., Donath, A., Peters, R.S., Ware, J., Misof, B., Niehuis, O., Pfrender, M.E., Zhou, X., 2017. The molecular evolutionary dynamics of oxidative phosphorylation (OXPHOS) genes in Hymenoptera. BMC Evol. Biol. 17, 269. https://doi.org/10.1186/s12862-017-1111-z

Liao, J.-Y.R., Friso, G., Kim, J., Wijk, K.J. van, 2018. Consequences of the loss of catalytic triads in chloroplast CLPPR protease core complexes in vivo. Plant Direct 2, e00086. https://doi.org/10.1002/pld3.86

Magee, A.M., Aspinall, S., Rice, D.W., Cusack, B.P., Sémon, M., Perry, A.S., Stefanović, S., Milbourne, D., Barth, S., Palmer, J.D., Gray, J.C., Kavanagh, T.A., Wolfe, K.H., 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Res. 20, 1700–1710. https://doi.org/10.1101/gr.111955.110

Majeran, W., Friso, G., van Wijk, K.J., Vallon, O., 2005. The chloroplast ClpP complex in Chlamydomonas reinhardtii contains an unusual high molecular mass subunit with a large apical domain. FEBS J. 272, 5558–5571. https://doi.org/10.1111/j.1742-4658.2005.04951.x

Majeran, W., Wollman, F.-A., Vallon, O., 2000. Evidence for a Role of ClpP in the Degradation of the Chloroplast Cytochrome b6f Complex. Plant Cell 12, 137–149. https://doi.org/10.1105/tpc.12.1.137

Millen, R.S., Olmstead, R.G., Adams, K.L., Palmer, J.D., Lao, N.T., Heggie, L., Kavanagh, T.A., Hibberd, J.M., Gray, J.C., Morden, C.W., Calie, P.J., Jermiin, L.S., Wolfe, K.H., 2001. Many Parallel Losses of infA from Chloroplast DNA during Angiosperm Evolution with Multiple Independent Transfers to the Nucleus. Plant Cell 13, 645–658. https://doi.org/10.1105/tpc.13.3.645

Molina, J., Hazzouri, K.M., Nickrent, D., Geisler, M., Meyer, R.S., Pentony, M.M., Flowers, J.M., Pelser, P., Barcelona, J., Inovejas, S.A., Uy, I., Yuan, W., Wilkins, O., Michel, C.-I., LockLear, S., Concepcion, G.P., Purugganan, M.D., 2014. Possible Loss of the Chloroplast Genome in the Parasitic Flowering Plant Rafflesia lagascae (Rafflesiaceae). Mol. Biol. Evol. 31, 793–803. https://doi.org/10.1093/molbev/msu051

Morris, J.L., Puttick, M.N., Clark, J.W., Edwards, D., Kenrick, P., Pressel, S., Wellman, C.H., Yang, Z., Schneider, H., Donoghue, P.C.J., 2018. The timescale of early land plant evolution. Proc. Natl. Acad. Sci. 115, E2274–E2283. https://doi.org/10.1073/pnas.1719588115

Müller, K., 2005. SeqState. Appl. Bioinformatics 4, 65–69. https://doi.org/10.2165/00822942-200504010-00008

Mullet, J.E., 1993. Dynamic regulation of chloroplast transcription. Plant Physiol. 103, 309–313.

Nishimura, K., Kato, Y., Sakamoto, W., 2017. Essentials of Proteolytic Machineries in Chloroplasts. Mol. Plant 10, 4–19. https://doi.org/10.1016/j.molp.2016.08.005

Nishimura, K., van Wijk, K.J., 2015. Organization, function and substrates of the essential Clp protease system in plastids. Biochim. Biophys. Acta BBA - Bioenerg., SI: Chloroplast Biogenesis 1847, 915–930. https://doi.org/10.1016/j.bbabio.2014.11.012

Olinares, P.D.B., Kim, J., Davis, J.I., van Wijk, K.J., 2011a. Subunit stoichiometry, evolution, and functional implications of an asymmetric plant plastid ClpP/R protease complex in Arabidopsis. Plant Cell 23, 2348–2361. https://doi.org/10.1105/tpc.111.086454

Olinares, P.D.B., Kim, J., van Wijk, K.J., 2011b. The Clp protease system; a central component of the chloroplast protease network. Biochim. Biophys. Acta BBA - Bioenerg., Regulation of Electron Transport in Chloroplasts 1807, 999–1011. https://doi.org/10.1016/j.bbabio.2010.12.003

Park, S., Ruhlman, T.A., Weng, M.-L., Hajrah, N.H., Sabir, J.S.M., Jansen, R.K., 2017. Contrasting Patterns of Nucleotide Substitution Rates Provide Insight into Dynamic Evolution of Plastid and Mitochondrial Genomes of Geranium. Genome Biol. Evol. 9, 1766–1780. https://doi.org/10.1093/gbe/evx124

Parkinson, C.L., Mower, J.P., Qiu, Y.-L., Shirk, A.J., Song, K., Young, N.D., dePamphilis, C.W., Palmer, J.D., 2005. Multiple major increases and decreases in mitochondrial substitution rates in the plant family Geraniaceae. BMC Evol. Biol. 5, 73. https://doi.org/10.1186/1471-2148-5-73

Peltier, J.-B., Cai, Y., Sun, Q., Zabrouskov, V., Giacomelli, L., Rudella, A., Ytterberg, A.J., Rutschow, H., Wijk, K.J. van, 2006. The Oligomeric Stromal Proteome of Arabidopsis thaliana Chloroplasts. Mol. Cell. Proteomics 5, 114–133. https://doi.org/10.1074/mcp.M500180-MCP200

Peltier, J.-B., Ripoll, D.R., Friso, G., Rudella, A., Cai, Y., Ytterberg, J., Giacomelli, L., Pillardy, J., Wijk, K.J. van, 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. https://doi.org/10.1074/jbc.M309212200

Peltier, J.-B., Ytterberg, J., Liberles, D.A., Roepstorff, P., Wijk, K.J. van, 2001. Identification of a 350-kDa ClpP Protease Complex with 10 Different Clp Isoforms in Chloroplasts of Arabidopsis thaliana. J. Biol. Chem. 276, 16318–16327. https://doi.org/10.1074/jbc.M010503200

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E., 2004. UCSF Chimera—A visualization system for exploratory research and analysis - Pettersen - 2004 - Journal of Computational Chemistry - Wiley Online Library. J. Comput. Chem.

Preston, B.D., 1996. Error-prone retrotransposition: rime of the ancient mutators. Proc. Natl. Acad. Sci. 93, 7427–7431. https://doi.org/10.1073/pnas.93.15.7427

Raubeson, L.A., Jansen, R.K., 2005. Chloroplast genomes of plants., in: Plant Diversity and Evolution: Genotypic and Phenotypic Variation in Higher Plants. CABI, Wallingford (UK), pp. 45–68.

Rautenberg, A., Sloan, D.B., Aldén, V., Oxelman, B., 2012. Phylogenetic Relationships of Silene multinervia and Silene Section Conoimorpha (Caryophyllaceae). Syst. Bot. 37, 226–237. https://doi.org/10.1600/036364412X616792

Rockenbach, K., Havird, J.C., Monroe, J.G., Triant, D.A., Taylor, D.R., Sloan, D.B., 2016. Positive Selection in Rapidly Evolving Plastid–Nuclear Enzyme Complexes. Genetics 204, 1507–1522. https://doi.org/10.1534/genetics.116.188268

Rombel, I.T., Sykes, K.F., Rayner, S., Johnston, S.A., 2002. ORF-FINDER: a vector for high-throughput gene identification. Gene 282, 33–41. https://doi.org/10.1016/S0378-1119(01)00819-8

Sabot, F., Schulman, A.H., 2006. Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. Heredity 97, 381–388. https://doi.org/10.1038/sj.hdy.6800903

Sanitá Lima, M., Smith, D.R., 2017. Pervasive Transcription of Mitochondrial, Plastid, and Nucleomorph Genomes across Diverse Plastid-Bearing Species. Genome Biol. Evol. 9, 2650–2657. https://doi.org/10.1093/gbe/evx207

Schelin, J., Lindmark, F., Clarke, A.K., 2002. The clpP multigene family for the ATP-dependent Clp protease in the cyanobacterium Synechococcus. Microbiol. Read. Engl. 148, 2255–2265. https://doi.org/10.1099/00221287-148-7-2255

Shevchenko, A., Tomas, H., Havlis, J., Olsen, J.V., Mann, M., 2006. In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat. Protoc. 1, 2856–2860. https://doi.org/10.1038/nprot.2006.468

Sloan, D.B., Alverson, A.J., Chuckalovcak, J.P., Wu, M., McCauley, D.E., Palmer, J.D., Taylor, D.R., 2012a. Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria with Exceptionally High Mutation Rates. PLOS Biol. 10, e1001241. https://doi.org/10.1371/journal.pbio.1001241

Sloan, D.B., Alverson, A.J., Wu, M., Palmer, J.D., Taylor, D.R., 2012b. Recent Acceleration of Plastid Sequence and Structural Evolution Coincides with Extreme Mitochondrial Divergence in the Angiosperm Genus Silene. Genome Biol. Evol. 4, 294–306. https://doi.org/10.1093/gbe/evs006

Sloan, D.B., MacQueen, A.H., Alverson, A.J., Palmer, J.D., Taylor, D.R., 2010. Extensive Loss of RNA Editing Sites in Rapidly Evolving Silene Mitochondrial Genomes: Selection vs. Retroprocessing as the Driving Force. Genetics 185, 1369–1380. https://doi.org/10.1534/genetics.110.118000

Sloan, D.B., Oxelman, B., Rautenberg, A., Taylor, D.R., 2009. Phylogenetic analysis of mitochondrial substitution rate variation in the angiosperm tribe Sileneae. BMC Evol. Biol. 9, 260. https://doi.org/10.1186/1471-2148-9-260

Sloan, D.B., Triant, D.A., Forrester, N.J., Bergner, L.M., Wu, M., Taylor, D.R., 2014a. A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae (Caryophyllaceae). Mol. Phylogenet. Evol. 72, 82–89. https://doi.org/10.1016/j.ympev.2013.12.004

Sloan, D.B., Triant, D.A., Wu, M., Taylor, D.R., 2014b. Cytonuclear interactions and relaxed selection accelerate sequence evolution in organelle ribosomes. Mol. Biol. Evol. 31, 673–682. https://doi.org/10.1093/molbev/mst259

Smith, D.R., 2009. Unparalleled GC content in the plastid DNA of Selaginella. Plant Mol. Biol. 71, 627. https://doi.org/10.1007/s11103-009-9545-3

Smith, D.R., Keeling, P.J., 2015. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. Proc. Natl. Acad. Sci. 112, 10177–10184. https://doi.org/10.1073/pnas.1422049112

Smith, D.R., Lee, R.W., 2014. A Plastid without a Genome: Evidence from the Nonphotosynthetic Green Algal Genus Polytomella. Plant Physiol. 164, 1812–1819. https://doi.org/10.1104/pp.113.233718

Sobanski, J., Giavalisco, P., Fischer, A., Walther, D., Schoettler, M.A., Pellizzer, T., Golczyk, H., Obata, T., Bock, R., Sears, B.B., Greiner, S., 2018. Biparental inheritance of chloroplasts is controlled by lipid biosynthesis. bioRxiv 330100. https://doi.org/10.1101/330100

Stanne, T.M., Pojidaeva, E., Andersson, F.I., Clarke, A.K., 2007. Distinctive Types of ATP-dependent Clp Proteases in Cyanobacteria. J. Biol. Chem. 282, 14394–14402. https://doi.org/10.1074/jbc.M700275200

Stoike, L.L., Sears, B.B., 1998. Plastome Mutator–Induced Alterations Arise in Oenothera Chloroplast DNA Through Template Slippage. Genetics 149, 347–353.

Stoletzki, N., Eyre-Walker, A., 2011. The positive correlation between dN/dS and dS in mammals is due to runs of adjacent substitutions. Mol. Biol. Evol. 28, 1371–1380. https://doi.org/10.1093/molbev/msq320

Straub, S.C., Fishbein, M., Livshultz, T., Foster, Z., Parks, M., Weitemier, K., Cronn, R.C., Liston, A., 2011. Building a model: developing genomic resources for common milkweed (Asclepias syriaca) with low coverage genome sequencing. BMC Genomics 12, 211. https://doi.org/10.1186/1471-2164-12-211

Tillich, M., Funk, H.T., Schmitz-Linneweber, C., Poltnigg, P., Sabater, B., Martin, M., Maier, R.M., 2005. Editing of plastid RNA in Arabidopsis thaliana ecotypes. Plant J. 43, 708–715. https://doi.org/10.1111/j.1365-313X.2005.02484.x

Traverse, C.C., Ochman, H., 2016. Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. Proc. Natl. Acad. Sci. 113, 3311–3316. https://doi.org/10.1073/pnas.1525329113

Tsudzuki, T., Wakasugi, T., Sugiura, M., 2001. Comparative Analysis of RNA Editing Sites in Higher Plant Chloroplasts. J. Mol. Evol. 53, 327–332. https://doi.org/10.1007/s002390010222

van Wijk, K.J., Peltier, J.-B., Giacomelli, L., 2007. Isolation of Chloroplast Proteins from <Emphasis Type="Italic">Arabidopsis thaliana</Emphasis> for Proteome Analysis, in: Plant Proteomics, Methods in Molecular Biology. Humana Press, pp. 43–48. https://doi.org/10.1385/1-59745-227-0:43

Venkat, A., Hahn, M.W., Thornton, J.W., 2017. Multinucleotide mutations cause false inferences of positive selection. bioRxiv 165969. https://doi.org/10.1101/165969

Welsch, R., Zhou, X., Yuan, H., Álvarez, D., Sun, T., Schlossarek, D., Yang, Y., Shen, G., Zhang, H., Rodriguez-Concepcion, M., Thannhauser, T.W., Li, L., 2018. Clp Protease and OR Directly Control the Proteostasis of Phytoene Synthase, the Crucial Enzyme for Carotenoid Biosynthesis in Arabidopsis. Mol. Plant 11, 149–162. https://doi.org/10.1016/j.molp.2017.11.003

Weng, M.-L., Blazier, J.C., Govindu, M., Jansen, R.K., 2014. Reconstruction of the Ancestral Plastid Genome in Geraniaceae Reveals a Correlation between Genome Rearrangements, Repeats, and Nucleotide Substitution Rates. Mol. Biol. Evol. 31, 645–659. https://doi.org/10.1093/molbev/mst257

Weng, M.-L., Ruhlman, T.A., Jansen, R.K., 2016. Plastid–Nuclear Interaction and Accelerated Coevolution in Plastid Ribosomal Genes in Geraniaceae. Genome Biol. Evol. 8, 1824–1838. https://doi.org/10.1093/gbe/evw115

Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M.S., Burleigh, J.G., Gitzendanner, M.A., Ruhfel, B.R., Wafula, E., Der, J.P., Graham, S.W., Mathews, S., Melkonian, M., Soltis, D.E., Soltis, P.S., Miles, N.W., Rothfels, C.J., Pokorny, L., Shaw, A.J., DeGironimo, L., Stevenson, D.W., Surek, B., Villarreal, J.C., Roure, B.,

Philippe, H., dePamphilis, C.W., Chen, T., Deyholos, M.K., Baucom, R.S., Kutchan, T.M., Augustin, M.M., Wang, J., Zhang, Y., Tian, Z., Yan, Z., Wu, X., Sun, X., Wong, G.K.-S., Leebens-Mack, J., 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl. Acad. Sci. U. S. A. 111, E4859-4868. https://doi.org/10.1073/pnas.1323926111

Williams, A.V., Boykin, L.M., Howell, K.A., Nevill, P.G., Small, I., 2015. The Complete Sequence of the Acacia ligulata Chloroplast Genome Reveals a Highly Divergent clpP1 Gene. PLOS ONE 10, e0125768. https://doi.org/10.1371/journal.pone.0125768

Wolf, J.B.W., Künstner, A., Nam, K., Jakobsson, M., Ellegren, H., 2009. Nonlinear Dynamics of Nonsynonymous (dN) and Synonymous (dS) Substitution Rates Affects Inference of Selection. Genome Biol. Evol. 1, 308–319. https://doi.org/10.1093/gbe/evp030

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

Wolfe, K.H., Morden, C.W., Palmer, J.D., 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc. Natl. Acad. Sci. 89, 10648–10652.

Wu, Z., Cuthbert, J.M., Taylor, D.R., Sloan, D.B., 2015. The massive mitochondrial genome of the angiosperm Silene noctiflora is evolving by gain or loss of entire chromosomes. Proc. Natl. Acad. Sci. 112, 10185–10191. https://doi.org/10.1073/pnas.1421397112

Yan, Z., Ye, G., Werren, J., 2018. Evolutionary rate coevolution between mitochondria and mitochondria-associated nuclear-encoded proteins in insects. bioRxiv 288456. https://doi.org/10.1101/288456

Yang, Z., 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24, 1586–1591. https://doi.org/10.1093/molbev/msm088

Yao, X., Tang, P., Li, Z., Li, D., Liu, Y., Huang, H., 2015. The First Complete Chloroplast Genome Sequences in Actinidiaceae: Genome Structure and Comparative Analysis. PLOS ONE 10, e0129347. https://doi.org/10.1371/journal.pone.0129347

Yeang, C.-H., Haussler, D., 2007. Detecting Coevolution in and among Protein Domains. PLoS Comput. Biol. 3. https://doi.org/10.1371/journal.pcbi.0030211

Yu, A.Y.H., Houry, W.A., 2007. ClpP: A distinctive family of cylindrical energy-dependent serine proteases. FEBS Lett. 581, 3749–3757. https://doi.org/10.1016/j.febslet.2007.04.076

Zeiler, E., List, A., Alte, F., Gersch, M., Wachtel, R., Poreba, M., Drag, M., Groll, M., Sieber, S.A., 2013. Structural and functional insights into caseinolytic proteases reveal an unprecedented regulation principle of their catalytic triad. Proc. Natl. Acad. Sci. U. S. A. 110, 11302–11307. https://doi.org/10.1073/pnas.1219125110

Zhang, J., Nielsen, R., Yang, Z., 2005. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. Mol. Biol. Evol. 22, 2472–2479. https://doi.org/10.1093/molbev/msi237

Zhang, J., Ruhlman, T.A., Sabir, J., Blazier, J.C., Jansen, R.K., 2015. Coordinated Rates of Evolution between Interacting Plastid and Nuclear Genes in Geraniaceae. Plant Cell 27, 563–573. https://doi.org/10.1105/tpc.114.134353

Zhang, J., Ruhlman, T.A., Sabir, J.S.M., Blazier, J.C., Weng, M.-L., Park, S., Jansen, R.K., 2016. Coevolution between Nuclear-Encoded DNA Replication, Recombination, and Repair Genes and Plastid Genome Complexity. Genome Biol. Evol. 8, 622–634. https://doi.org/10.1093/gbe/evw033

Zhang, Y., Ma, J., Yang, B., Li, R., Zhu, W., Sun, L., Tian, J., Zhang, L., 2014. The complete chloroplast genome sequence of Taxus chinensis var. mairei (Taxaceae): loss of an inverted repeat region and comparative analysis with related species. Gene 540, 201–209. https://doi.org/10.1016/j.gene.2014.02.037

Zheng, B., MacDonald, T.M., Sutinen, S., Hurry, V., Clarke, A.K., 2006. A nuclear-encoded ClpP subunit of the chloroplast ATP-dependent Clp protease is essential for early development in <Emphasis Type="Italic">Arabidopsis thaliana</Emphasis>. Planta 224, 1103–1115. https://doi.org/10.1007/s00425-006-0292-2

Zhu, A., Guo, W., Jain, K., Mower, J.P., 2014. Unprecedented Heterogeneity in the Synonymous Substitution Rate within a Plant Genome. Mol. Biol. Evol. 31, 1228–1236. https://doi.org/10.1093/molbev/msu079

CHAPTER 3: LONG-READ TRANSCRIPTOME AND OTHER GENOMIC RESOURCES FOR THE ANGIOSPERM *SILENE NOCTIFLORA*[2]

Summary

The angiosperm genus *Silene* is a model system for several traits of ecological and evolutionary significance in plants, including breeding system and sex chromosome evolution, host-pathogen interactions, invasive species biology, heavy metal tolerance, and cytonuclear interactions. Despite its importance, genomic resources for this large genus of approximately 850 species are scarce, with only one published whole-genome sequence (from the dioecious species *S. latifolia*). Here, we provide genomic and transcriptomic resources for a hermaphroditic representative of this genus (*S. noctiflora*), including a PacBio Iso-Seq transcriptome, which uses long-read, single-molecule sequencing technology to analyze full-length mRNA transcripts. Using these data, we have assembled and annotated high-quality full-length cDNA sequences for approximately 14,126 *S. noctiflora* genes and 25,317 isoforms. We demonstrated the utility of these data to distinguish between recent and highly similar gene duplicates by identifying novel paralogous genes in an essential protease complex. Further, we provide a draft assembly for the approximately 2.7-Gb genome of this species, which is near the upper range of genome-size values reported for diploids in this genus and three-fold larger than the 0.9-Gb genome of *S. conica*, another species in the same subgenus. Karyotyping confirmed that *S. noctiflora* is a diploid, indicating that its large genome size is not due to polyploidization. These resources should facilitate further study and development of this genus as a model in plant ecology and evolution.

Introduction

Silene is the largest genus in the angiosperm family Caryophyllaceae and serves as a model system in many fields of ecology and evolutionary biology (Bernasconi et al., 2009; Jafari et al., 2020).

[2] Published in *G3: Genes|Genomes|Genetics*, Volume 11, August 2021.
Authors: Alissa M. Williams, Michael W. Itgen, Amanda K. Broz, Olivia G. Carter, Daniel B. Sloan

For instance, *Silene* is used to study breeding system evolution, as the genus includes hermaphroditic, gynodioecious, gynomonoecious, monoecious, and dioecious species (Charlesworth, 2006; Desfeux et al., 1996). Despite the diversity of *Silene* sexual systems, there is only one available whole genome sequence for the entire genus—from the dioecious species *S. latifolia*, which has heteromorphic XY sex chromosomes (Krasovec et al., 2018; Papadopulos et al., 2015). Whole genome resources are not available for any of the hermaphroditic species, which has limited comparative genomic studies into the evolution of dioecy within this genus.

Silene is also used as a model system for investigating organelle genome evolution and the coevolution between nuclear and cytoplasmic genomes (i.e., cytonuclear interactions) (Garraud et al., 2011; Klaas and Olson, 2006; Olson and Mccauley, 2002; Städler and Delph, 2002). *Silene conica* and *S. noctiflora* have two of the largest known plant mitochondrial genomes at 11 Mb and 7 Mb, respectively (Sloan et al., 2012a). In contrast, the mitochondrial genome of *S. latifolia* is only 0.25 Mb, about 45 times smaller than that of *S. conica* (Sloan et al., 2012a). Interestingly, the *Silene* species with expanded mitogenomes also display unusually high evolutionary rates and structural changes in mitochondrial and plastid DNA (Mower et al., 2007; Sloan et al., 2012a). The natural variation in organelle genome evolution found in this genus has been used to study how these differences affect cytonuclear interactions (Havird et al., 2015; Williams et al., 2019).

The ability to use *Silene* as a model for cytonuclear evolution is still limited by the lack of extensive nuclear genome resources. Previous work has characterized *Silene* nuclear genome size and chromosome number. Nuclear genome sizes in the genus vary considerably, although not as starkly as mitochondrial genome sizes, ranging roughly 4.5-fold among diploids (haploid sizes of 0.71 to 3.23 Gb) and 8-fold when the tetraploid *S. stellata* (5.77 Gb) is included (Bai et al., 2012; Dagher-Kharrat et al., 2013; Kruckeberg, 1960; Pellicer and Leitch, 2020; Siroký et al., 2001). Most of the available nuclear sequence data comes from short-read RNA sequencing, which has been conducted on multiple *Silene* species (Balounova et al., 2019; Bertrand et al., 2018; Blavet et al., 2011; Casimiro-Soriguer et al., 2016; Havird et al., 2017; Muyle et al., 2012; Sloan et al., 2012b). These datasets have provided an important

resource for molecular studies of *Silene*, but are limited because of the challenges associated with assembling short-read sequences, especially in distinguishing similar sequences arising from gene duplication, heterozygosity, and/or alternative splicing (Alkan et al., 2011; Hahn et al., 2014; Lan et al., 2017; Schatz et al., 2012).

We have generated genomic resources critical for investigations into *S. noctiflora*, a species of interest due to its extremely unusual organelle evolution and resultant use as a model for cytonuclear interactions, as well as its status as a hermaphrodite in a genus representing many types of breeding system. We include a high-quality transcriptome using long-read PacBio Iso-Seq technology, genome size estimates, and a draft nuclear genome assembly. These resources will expand opportunities for molecular and ecological studies within the genus.

Materials and Methods

Study system

Silene noctiflora (**Figure 3.1**) is largely hermaphroditic but can produce a mixture of hermaphroditic and male-sterile flowers on the same plant (gynomonoecy) (Davis and Delph, 2005). Also known as the night-flowering catchfly, this annual species is native to Eurasia and introduced throughout much of the world (Davis and Delph, 2005; McNeill, 1980).

Plant growth conditions, tissue sampling, and nucleic acid extractions

Plants used for genome sequencing, Iso-Seq, and flow cytometry estimates of genome size were grown under standard greenhouse conditions with 16-hr light/8-hr dark at Colorado State University (**Table 3.1**). DNA for short-insert paired-end Illumina libraries was extracted from leaf tissue of a 7-week-old *S. noctiflora* individual from an Opole, Poland (OPL) population using a Qiagen Plant DNeasy kit. To obtain sufficient DNA quantity for construction of Illumina mate-par libraries, additional DNA was extracted from the same individual 6 weeks later using a modified CTAB protocol (Doyle and Doyle,

1987) for construction of Illumina mate-pair libraries. For Iso-Seq library construction, RNA was extracted from a single 12-week-old *S. noctiflora* OPL individual (grown from seed of the plant used for DNA extraction), using a Qiagen Plant RNeasy kit. RNA extractions were performed for four different tissue samples: 1) a large flower bud with calyx removed, 2) an entire smaller flower bud including calyx, 3) the most recent (top-most) pair of cauline leaves, and 4) one leaf from the second most recent pair of cauline leaves. The four RNA extractions were quantified with Qubit RNA BR kit (Thermo Fisher Scientific). Purity and integrity were assessed with a NanoDrop 2000 (Thermo Fisher Scientific) and TapeStation 2200 (Agilent Technologies). Different tissues and developmental stages were sampled (and eventually pooled; see below) to capture a larger diversity of transcripts and thereby increase the number of genes represented.

PacBio Iso-Seq transcriptome sequencing and analysis

Iso-Seq is an application of Pacific Biosciences (PacBio) long-read sequencing technology that uses cDNA templates to generate high quality reads for full-length transcripts. The high error rate generally associated with PacBio sequencing is drastically reduced using circular consensus sequencing (CCS), which uses hairpin adapters on each end of a double-stranded molecule to create a circular, single-stranded topology (Au et al., 2012; Hestand et al., 2016; Rhoads and Au, 2015; Wenger et al., 2019). This topology allows the polymerase to read the same full-length molecule multiple times over, generating an accurate consensus sequence (Ono et al., 2013; Wang et al., 2019). PacBio Iso-Seq has been used to study the transcriptomes of many organisms, often in the context of identifying splice variants, or alternative transcripts (Abdel-Ghany et al., 2016; Gordon et al., 2015; Guo et al., 2016; Rhoads and Au, 2015; Wang et al., 2016; Weirather et al., 2017; Xu et al., 2015). Alternative transcripts can be identified using CCS because this technology obtains consensus sequences for full-length single transcripts (Zhao et al., 2019). In the same way, CCS can also be used to distinguish paralogs or gene duplicates.

To create an Iso-Seq library for *S. noctiflora*, the four RNA extractions (1.5 μg each) were pooled into a single sample and sent to the Arizona Genomics Institute for PacBio Iso-Seq library construction

and sequencing. The library was constructed on the pooled RNA sample using Poly(A) selection, following the standard PacBio Iso-Seq protocol ("Procedure & Checklist – Iso-Seq Template Preparation for Sequel Systems," Pacific Biosciences, PN-101-070-200 Version 06, September 2018), and then was sequenced with a PacBio Sequel (first generation) platform on two SMRT Cells.

Raw movie files of long-read, single-molecule sequences (one per SMRT Cell) were processed using the PacBio Iso-Seq v3.1 pipeline (Anvar *et al.* 2018; Pacific Biosciences 2020). Circular consensus sequence calling was performed on each movie file separately using the command *ccs* with the recommended parameters *--noPolish* and *--minPasses 1*. Next, primer removal was performed on each dataset by running the command *lima* with parameters *--isoseq* and *--no-pbi*. Poly(A) tails were trimmed and concatemers were removed using the *refine* command with the parameter *--require-polya*. Data from the two cells were merged at this point using the commands *dataset create --type TranscriptSet* and *dataset create --type SubreadSet*. Finally, the merged data were run through the *cluster* and *polish* commands. We also ran the *cluster* and *polish* commands on each dataset individually after skipping the merge step.

Trinotate v3.2.0 (Bryant et al., 2017) was used to annotate the final polished sequences produced by the Iso-Seq pipeline after merging the datasets. To complete this process, we used Transdecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder/wiki), SQLite v3 (Kreibich, 2010), NCBI BLAST + v2.2.29 (Camacho et al., 2009), HMMER v3.2.1 (including RNAMMER) (Lagesen et al., 2007; Potter et al., 2018), signalP v4 (Petersen et al., 2011), and tmhmm v2 (Krogh et al., 2001). The Pfam (Bateman et al., 2004) and UniProt ("UniProt," 2015) databases were included in the Trinotate installation. The transcripts and Transdecoder-predicted peptides were searched against the respective databases, following the standard Trinotate pipeline. All of these results were loaded into a Trinotate SQLite database.

Cogent v4.0.0 (https://github.com/Magdoll/Cogent/wiki) and minimap2 v2.17 (Li 2018) were used to conduct family finding on the final sequences by the Iso-Seq pipeline by partitioning sequences into groups based on similarity. While the Iso-Seq pipeline collapses reads into individual transcripts, it does not collapse alternative transcripts originating from the same gene. Cogent further collapses

alternative transcripts into groups, where each group is meant to represent a single gene. Next, coding genome reconstruction was performed on each group from the above step; thus, the Cogent output included both a file containing groups of alternative transcripts (final.partition.txt at https://github.com/alissawilliams/Silene_noctiflora_IsoSeq) and a transcript-based genome. Finally, this transcript-based genome was used to determine total gene and isoform (alternative transcript) counts via cDNA_Cupcake scripts (https://github.com/Magdoll/cDNA_Cupcake/wiki; (Jeffries et al., 2020; Wang et al., 2020). A modified form of the script *make_file_for_sampling_from_collapsed.py* was run with the parameter --*include_single_exons* in order to include all transcripts in the analysis. Gene and isoform counts were calculated using custom Python and R scripts on the resultant file. These Cogent, minimap2, and cDNA_Cupcake steps were performed on the merged dataset as well as individually on the datasets from each SMRT Cell.

We used genes from the plastid caseinolytic protease (Clp) as a case study to assess the ability of Iso-Seq dataset to distinguish paralogs (gene duplicates) of various levels of divergence. To identify nuclear-encoded plastid Clp core genes in our dataset, we used blastn in conjunction with the Cogent family finding output. There are eight nuclear-encoded plastid Clp core genes in *Arabidopsis thaliana*: *CLPP3-6* and *CLPR1-4* (Nishimura and van Wijk, 2015). Additionally, the genus *Silene* shares a duplication of *CLPP5*, denoted *CLPP5A* and *CLPP5B* (Rockenbach et al., 2016). We obtained the sequences of all nine of these genes from a previous study (Rockenbach et al., 2016) and used them as queries in blastn searches against the *S. noctiflora* Iso-Seq transcriptome. We then identified which groups of collapsed alternative transcripts (from the Cogent output) contained these BLAST hits. BLAST hits for eight of the nine nuclear-encoded Clp core subunits in *Silene* (including *CLPP5A* and *CLPP5B*) were found in a single Cogent group. The sequences within each group were confirmed to represent a single gene via alignment and manual inspection; thus, these eight core subunits are single copy in *S. noctiflora*. However, in the case of *CLPR2*, two different Cogent groups contained relevant transcripts, indicating a possible case of gene duplication. Sequence alignment and manual inspection of the transcripts within these two Cogent groups revealed that one group contained two unique sequences.

These data, along with sequencing results from a separate project in which we cloned two versions of *S. noctiflora CLPR2* using primers designed for *S. latifolia CLPR2*, suggested that there are actually three distinct *CLPR2* sequences in *S. noctiflora*. In the subsequent phylogenetic analysis of *CLPR2*, we used the longest sequences from each of the three identified groups.

A phylogenetic tree was constructed using sequences from the three different *S. noctiflora CLPR2* genes. In addition to the three *S. noctiflora* sequences, we also included *Agrostemma githago*, *S. conica, S. latifolia, S. paradoxa,* and *S. vulgaris CLPR2* sequences from a previous study (Rockenbach et al., 2016), as well as three *S. undulata CLPR2* sequences identified using blastn against the *S. undulata* TSA database (accession GEYX00000000). All 11 sequences were aligned using the *einsi* option in MAFFT v7.222 (Katoh and Standley, 2013), and trimmed at the 5′ end based on the trimming conducted in Rockenbach *et al*. (2016). The resultant sequence file was run through jModelTest v2.1.10 (Darriba et al., 2012) to choose a model of sequence evolution. We chose the top model based on the Bayesian Information Criterion (K80+I) and ran PhyML v3.3 (Guindon et al., 2010) with 1000 bootstrap replicates and 100 random starts.

Genome size estimates by flow cytometry

Leaf or seedling samples were collected from multiple individuals of varying age (between 2 and 14 weeks) for each of our target *Silene* species and shipped fresh to Plant Cytometry Services (Schijndel, Netherlands). Genome sizes were determined using the CyStain PI Absolute P reagent kit (05-5502). Samples were chopped with a razor blade in 500 µl of ice-cold Extraction Buffer in a plastic petri dish, along with *Pachysandra terminalis* tissue as an internal standard (3.5 pg/2C). After 30-60 sec of incubation, 2 ml of Staining Buffer was added. Each sample was then passed through a nylon filter of 50 µm mesh size, and then incubated for 30+ min at room temperature. The filtered solution was then sent through a CyFlow ML flow cytometer (Partec GmbH). The fluorescence of the stained nuclei, which passed through the focus of a light beam with a 50 mW, 532 nm green laser, was measured by a photomultiplier and converted into voltage pulses. The voltage pulses were processed using Flomax

version 2.4d (Partec) to yield integral and peak signals. Genome sizes were reported in units of pg/2C. The conversion used to report each size (x) in units of Gb was (x/2)*0.978 (Gregory et al., 2007).

Karyotyping

Silene noctiflora OPL seeds were germinated on wet filter paper and grown for 5 days. Radicles were trimmed off and transferred to ice water for 24 hrs. The radicles were then fixed in a 3:1 solution of absolute ethanol and glacial acetic acid and stored at -20°C. Chromosomes were visualized using a squash preparation with Feulgen staining. Fixed radicles were rinsed in distilled water for 5 min at 20°C. Radicles were then hydrolyzed in 5M HCl at 20°C for 60 min followed by three rinses in distilled water. The hydrolyzed radicles were transferred to Schiff's reagent to stain the DNA for 120 min at 20°C and were then destained by rinsing in SO_2 water at 20°C three times for 2 min, two times for 10 min, once for 20 min, and then transferred to distilled water. Squashes were prepared by placing a piece of tissue in 45% acetic acid for 10 min and then minced on glass. A coverslip was placed over the minced tissue and pressed with enough pressure to produce a monolayer of nuclei. Slides were placed on dry ice for 1 min, and the coverslip was removed. The slides were transferred to 96% ethanol for 2 min, air dried, and mounted with mounting medium. Chromosomes were observed using a compound light microscope at 100× magnification.

Genome sequencing and assembly

Extracted *S. noctiflora* OPL DNA samples were used for Illumina library construction and sequencing. A paired-end library with a target insert size of 275-bp was constructed at the Yale Center for Genome Analysis and sequenced on a 2×150-bp HiSeq 2500 run (three lanes). Two mate-pair libraries (with target insert sizes of 3-5 kb and 8-11 kb) were generated at GeneWiz and sequenced on a 2×150-bp HiSeq 2500 run (one lane each). Approximately 480M, 250M, and 230M read pairs were generated for the 275-bp, 3-5 kb, and 8-11 kb libraries, respectively. These reads are available via the NCBI SRA (accessions

SRR9591157-SRR9591159). Reads were trimmed for quality and to remove 3′ adapters, using cutadapt v1.3 (Martin, 2011) under the following paramters*: -n 3 -O 6 -q 20 -m 30 -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC --paired-output*. The trimmed reads were assembled with ALLPATHS-LG release 44837 (Gnerre et al., 2011). Estimates of mean insert size and standard deviation for each library were provided as input for the assembly by first mapping a sample of reads to the published *S. noctiflora* plastid genome (GenBank accession JF715056.1). These estimates were as follows: 274 bp (± 22 bp), 3752 bp (± 419 bp), and 9873 bp (± 1283 bp).

BUSCO analyses

Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis (Seppey et al., 2019) compares an assembly (transcriptomic or genomic) to a set of highly conserved orthologs from a particular clade in order to assess the completeness of the assembly. BUSCO (v4.1.4) analysis was performed on the Iso-Seq transcriptome and the genome assembly, as well as the output of the individual SMRT Cells. In each case, fasta files containing all genomic or transcriptomic sequences were run through BUSCO using the lineage eudicots_odb10 (2020-09-10) and default parameters. The graphical summary of results was produced using the script generate_plot.py included in the BUSCO installation.

Data availability

The original subread bam files and final transcript sequences longer than 199 bp from the PacBio Iso-Seq transcriptome are available at NCBI Sequence Read Archive (SRA accession SRR11784995) and NCBI Transcriptome Shotgun Assembly Sequence Database (TSA accession GIOF01000000), respectively. The genome assembly has been deposited in GenBank (accession VHZZ00000000.1). Additional data have been provided at GitHub (https://github.com/alissawilliams/Silene_noctiflora_IsoSeq): 1) the full transcriptome as outputted by the PacBio Iso-Seq pipeline, 2) the annotation report for the transcriptome, 3) a custom script used to create a gene_trans_map file for our data in order to use Trinotate on non-Trinity-derived data (i.e. transcripts derived from sources other than a Trinity assembly, in this case Iso-

Seq transcripts), 4) the Cogent output containing collapsed groups of transcripts, and 5) the set of trimmed, aligned sequences used in the *CLPR2* phylogenetic analysis.

Results and Discussion

Silene noctiflora Iso-Seq transcriptome: Gene content and duplication

Sequencing of the Iso-Seq library on two Sequel SMRT Cells produced 711,625 and 686,576 reads for the first and second cells, respectively, where each read was derived from a single molecule. The two SMRT Cells differed substantially in data yield, with totals of 12,765,109 and 21,844,543 subreads, corresponding to subread counts of 17.9 and 31.8 per read, respectively. These reads were merged into 65,642 distinct high-quality transcripts according to the thresholds of the Iso-Seq 3.1 *merge* and *polish* commands. Of these transcripts, only 14 were found to be non-plant sequences, all of which were derived from *Frankliniella occidentalis* (the western flower thrip), a common greenhouse pest that likely contaminated our tissue samples. We annotated these transcripts using Trinotate (Bryant et al., 2017); our dataset contains 69,846 total entries for the 65,642 transcripts (transcripts with multiple predicted proteins are represented by multiple entries). Of the 69,846 entries, 48,742 (74.3%) have an annotated PFAM domain, 47,504 (68.0%) have a KEGG annotation, and 55,993 (80.2%) have at least one predicted Gene Ontology term.

Each high-quality transcript represents collapsed reads, meaning that identical or nearly identical sequences are represented by the same final sequence. However, the Iso-Seq pipeline does not collapse alternatively spliced transcripts, or isoforms; thus, this final dataset includes multiple transcripts derived from the same genes. In addition to separately representing isoforms, the transcriptome data could also contain alleles of the same gene and transcripts from paralogs (gene duplicates). Given sufficiently divergent alleles or paralogs, pairs of these types of sequences will also be represented by separate final transcripts in this dataset. Due to the low levels of polymorphism and heterozygosity in *S. noctiflora* (Sloan *et al.* 2012a), we did not expect different alleles to comprise a major portion of this dataset.

Based on a BUSCO analysis (Seppey et al., 2019), the Iso-Seq transcriptome had a completeness of 74.9%. This estimate included a large number of duplicated BUSCOs (47.7%), but these do not necessarily represent true gene duplications for the reasons stated above (**Figure 3.2)**. The merged dataset had a higher completeness percentage than either of the individual SMRT Cells, where the second SMRT Cell was more complete than the first, consistent with the differential data yield between the two cells (**Figure 3.2**). The estimated BUSCO completeness of the transcriptome was lower than that of the assembled nuclear genome (see below), which suggests that some genes with low or tissue-specific expression were not captured. Future efforts to generate deeper sequencing across a wider sample of tissues and environments may be beneficial in this respect.

We used the Cogent (https://github.com/Magdoll/Cogent/wiki) family finding algorithm to further collapse the transcripts into groups of isoforms (alternative transcripts) originating from the same gene. Notably, if paralogs (gene duplicates) have high enough sequence similarity, this binning could include them in the same group. We then used the Cogent data along with Cupcake (https://github.com/Magdoll/cDNA_Cupcake/wiki) to calculate the number of genes and isoforms represented in the transcriptome. Based on this analysis, the Iso-Seq transcriptome contains 14,126 *S. noctiflora* genes and 25,317 isoforms. Of the 14,126 genes, 7,027 had a single isoform (49.7%). We also calculated gene and isoform counts for each individual SMRT Cell; the first SMRT Cell produced 6,790 genes and 10,568 isoforms, while the second SMRT Cell produced 10,283 genes and 17,000 isoforms.

We wanted to test the ability of Iso-Seq to detect and distinguish known paralogs of varying levels of divergence using the Cogent family finding output. To this end, we used a sample gene family—the core subunit genes of the plastid Clp complex, as they have a rich history of paralogy. In *E. coli* and most other bacteria, the core of the Clp complex, which is responsible for proteolysis, contains 14 identical subunits (Yu and Houry, 2007). In cyanobacteria, gene duplication has led to four different core subunit-encoding genes (Stanne et al., 2007). Continued gene duplication in the land plant lineage has further reshaped this complex in plastids; the 14 core subunits are encoded by nine different genes in *A. thaliana*, eight of which are nuclear encoded (*CLPP3-6, CLPR1-4*), and one of which is plastid encoded

(*clpP1*) (Nishimura and van Wijk, 2015). Further, we had previously identified a more recent duplication of *CLPP5* in *Silene*, as well as duplications of the plastid-encoded *clpP1* in a small number of angiosperm species (Erixon and Oxelman, 2008; Rockenbach et al., 2016; Williams et al., 2019). The Clp complex is one of the most highly expressed stromal proteases (Nishimura and van Wijk, 2015). It is expressed in most tissues throughout the life stages of the plant, including the tissues from which we extracted RNA (Zheng et al., 2002). Thus, we would expect a transcriptome generated from the tissues we used to yield sequences of the various components of the Clp complex.

We used the Cogent output to examine the nine nuclear-encoded Clp core genes in *S. noctiflora*. The core genes *CLPP3, CLPP4, CLPP5A, CLPP5B, CLPP6, CLPR1, CLPR3,* and *CLPR4* were each represented by a single group in the Cogent output, whereas *CLPR2* was represented by two groups. Upon further examination, one of these groups actually represented two different genes, yielding a total of three *CLPR2* genes in *S. noctiflora*. Thus, *CLPR2* was duplicated in this lineage, and then one paralog underwent a second gene duplication. Based on a phylogenetic analysis (**Figure 3.3**), these two duplications are shared with *S. undulata* but none of the other sampled *Silene* species. Thus, these duplications likely occurred after *Silene* section Elisanthe (including *S. noctiflora*, *S. undulata*, and *S. turkestanica*) diverged from the other members of the genus (Jafari et al., 2020; Moiloa et al., 2021).

The Iso-Seq data allowed us to identify transcripts from every known nuclear-encoded Clp core gene in *S. noctiflora*, including the closely related *CLPP5A* and *CLPP5B* subunits, as well as an additional, previously unreported triplication of *CLPR2*. To corroborate the triplication of *CLPR2* in *S. noctiflora* that was identified using the Iso-Seq transcriptome, we used the *CLPR2* sequence from Rockenbach et al (2016) as a query in a blastn search against the *S. noctiflora* genome assembly. This search returned four scaffold hits. Upon examination, each *CLPR2* gene identified in the Iso-Seq transcriptome was represented by one scaffold. The fourth scaffold represented all three gene copies in a short region of high sequence identity between them, suggesting collapsing of similar sequence content within the genome assembly. Thus, each *CLPR2* gene was fully represented by sequences on two

scaffolds—there was one unique scaffold per gene containing most of the sequence and one scaffold containing sequence shared by all three genes.

Silene genome size estimates and chromosome number

Genome sizes of *S. noctiflora, S. conica, S. vulgaris,* and *S. latifolia* were determined using flow cytometry. Our estimates for *S. vulgaris* and *S. latifolia* (1.07 and 2.67 Gb, respectively; **Table 3.1**) were concordant with previously published estimates for these two species of 1.11 and 2.64 Gb (Costich et al., 1991; Siroký et al., 2001). Interestingly, despite their similar and extreme patterns of organelle evolution (Sloan et al., 2014, 2012a), including large mitochondrial genomes, *S. noctiflora* and *S. conica* have very different nuclear genome sizes. We found their respective genome sizes to be approximately 2.74 and 0.93 Gb, respectively (**Table 3.1**), which are on opposite ends of the spectrum for *Silene* diploids (Pellicer and Leitch, 2020). The *S. noctiflora* nuclear genome is almost three-fold larger than that of *S. conica* suggesting that mitochondrial genome size is not necessarily correlated with nuclear genome size.

Most diploids in the genus, including *S. noctiflora,* have a chromosome number of 2n=24, which is likely the ancestral number (Bari, 1973; Ghasemi et al., 2015; Gholipour and Sheidai, 2010; Kemal et al., 2009; McNeill, 1980; Mirzadeh Vaghefi and Jalili, 2019; Yildiz et al., 2008). There are also numerous polyploid *Silene* species, including tetraploid, hexaploid, and octaploid forms (Bai et al., 2012; Kruckeberg, 1960; Popp et al., 2005; Popp and Oxelman, 2007, 2001). *Silene noctiflora* has been previously reported as a diploid (Ghasemi et al., 2015; McNeill, 1980; Yildiz et al., 2008). Given its relatively large genome size, we sought to confirm this result in our sampled population with a karyotype analysis (**Figure 3.4**), which indeed supported the conclusion that *S. noctiflora* OPL is diploid.

The *Silene noctiflora* nuclear genome

Illumina sequencing produced ~50× coverage of the *S. noctiflora* genome for a 275-bp paired-end library and ~15-20× for each of two mate-pair libraries. By performing a *de novo* assembly of these reads,

we obtained a total assembly length (including estimated scaffold gaps) of 2.58 Gb, which is generally consistent with our estimate based on flow cytometry for *S. noctiflora* OPL (2.71 Gb). Given that we relied entirely on short-read sequencing technology, it was not surprising that the resulting assembly of this large genome was highly fragmented (79,768 scaffolds with a scaffold N50 of 59 kb; 222,040 contigs [minimum length of 1 kb for reporting contigs] with a contig N50 of 4.8 kb). Moreover, assembly gaps made up 73% of the total scaffold length, presumably representing the highly repetitive content that is typical of plant nuclear genomes. As such, the assembled gap-free sequences amount to only about a quarter of the genome (702 Mb). Given the expected low levels of polymorphism and heterozygosity in *S. noctiflora* (Sloan et al., 2012a), the assembly was interpreted as a single haplotype and no attempt was made to phase the two distinct haplotypes within the diploid.

BUSCO analysis (Seppey et al., 2019) provided an estimate of 89.5% completeness for the *S. noctiflora* genome assembly (**Figure 3.2**). Only 3.0% of BUSCOs were reported to be duplicated, in great contrast to the transcriptome, where 47.7% of BUSCOs were duplicated. Given that the final Iso-Seq dataset includes alternatively spliced transcripts as separate entries, it is not surprising that the transcriptome had a higher percentage of duplicated BUSCOs than the genome assembly.

As a complement to the Iso-Seq transcriptome, this *S. noctiflora* genome assembly should provide a useful resource to query for sequences of interest, especially in genic regions, and to compare against *S. latifolia* and other members of this genus. However, a more complete assembly that includes repetitive regions of the genome will require additional data from long-read technologies such as PacBio or nanopore sequencing. The Iso-Seq data generated in this study may be helpful in combination with improved genomic sequencing data in the future, as a means to improve scaffolding (Zhu et al., 2018), resolve paralogs (e.g., the collapsed regions of the *CLPR2* paralogs in the genome assembly), and annotate gene models.

Table 3.1: Genome sizes determined by flow cytometry

				Mean Genome Size	
Species	**Population**	**Location**	**Samples, 2C (pg)**	**2C (pg)**	**1C (Gb)**
Silene noctiflora	OPL*	Opole, Poland	5.65, 5.61, 5.46, 5.44	5.54	2.71
	OSR	Giles County, VA	5.75, 5.61	5.68	2.78
	BRP	Nelson County, VA	5.63, 5.57	5.60	2.74
Silene conica	ABR	Abruzzo, Italy	1.92, 1.92, 1.88	1.91	0.93
Silene vulgaris	S9L	Giles County, VA	2.19, 2.16	2.18	1.07
Silene latifolia	UK2600	Bedford County, VA	5.46, 5.45	5.46	2.67

*The *S. noctiflora* OPL population was used for Iso-Seq, genome assembly, and karyotyping
Units: pg = picogram, Gb = gigabase, 1C = haploid amount, 2C = diploid amount

Figure 3.1: *Silene noctiflora*, also known as the night-flowering catchfly.

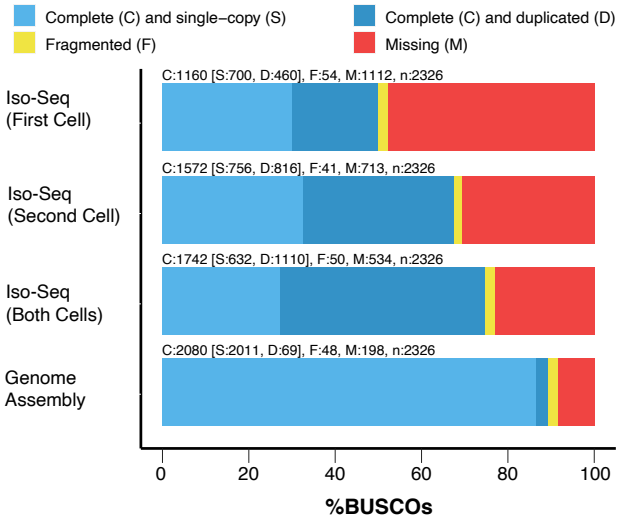


Figure 3.2: BUSCO analysis of the *S. noctiflora* genome assembly, Iso-Seq transcriptome (full dataset), and the individual SMRT Cells that were merged to create the Iso-Seq transcriptome.

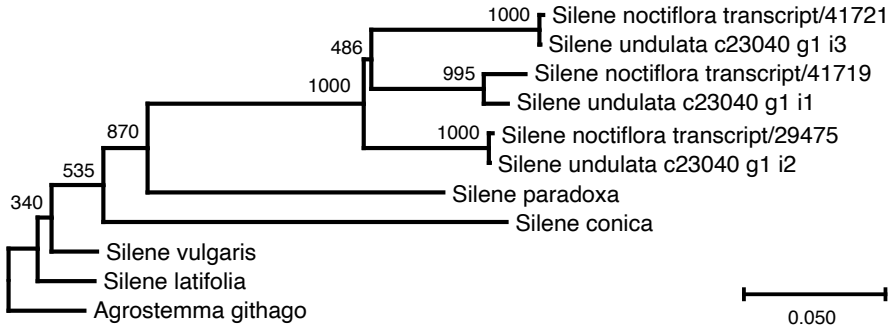


Figure 3.3: Phylogenetic analysis of *CLPR2* genes in *S. noctiflora* and related species. Branch lengths represent nucleotide sequence divergence. This tree was rooted on the *Agrostemma githago* sequence. The placement of *S. paradoxa* is in conflict with the species tree (Jafari et al., 2020), likely due to long branch attraction and the multiple independent evolutionary rate accelerations in this protein across *Silene* (Rockenbach et al., 2016).

Figure 3.4: Micrograph verifying the diploidy of *Silene noctiflora* at 100× magnification. Although an exact chromosome count is difficult to determine, this image suggests that *S. noctiflora* is a diploid with the typical number of 24 chromosomes previously documented in this species and the genus in general, rather than polyploid with 48 or more chromosomes (Bari, 1973; Ghasemi et al., 2015; Gholipour and Sheidai, 2010; Kemal et al., 2009; McNeill, 1980; Mirzadeh Vaghefi and Jalili, 2019; Yildiz et al., 2008).

LITERATURE CITED

Abdel-Ghany, S.E., Hamilton, M., Jacobi, J.L., Ngam, P., Devitt, N., Schilkey, F., Ben-Hur, A., Reddy, A.S.N., 2016. A survey of the sorghum transcriptome using single-molecule long reads. Nature Communications 7, 1–11. https://doi.org/10.1038/ncomms11706

Alkan, C., Sajjadian, S., Eichler, E.E., 2011. Limitations of next-generation genome sequence assembly. Nat Methods 8, 61–65. https://doi.org/10.1038/nmeth.1527

Anvar, S.Y., Allard, G., Tseng, E., Sheynkman, G.M., de Klerk, E., Vermaat, M., Yin, R.H., Johansson, H.E., Ariyurek, Y., den Dunnen, J.T., Turner, S.W., 't Hoen, P.A.C., 2018. Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. Genome Biology 19, 46. https://doi.org/10.1186/s13059-018-1418-0

Au, K.F., Underwood, J.G., Lee, L., Wong, W.H., 2012. Improving PacBio Long Read Accuracy by Short Read Alignment. PLoS One 7. https://doi.org/10.1371/journal.pone.0046679

Bai, C., Alverson, W.S., Follansbee, A., Waller, D.M., 2012. New reports of nuclear DNA content for 407 vascular plant taxa from the United States. Ann. Bot. 110, 1623–1629. https://doi.org/10.1093/aob/mcs222

Balounova, V., Gogela, R., Cegan, R., Cangren, P., Zluvova, J., Safar, J., Kovacova, V., Bergero, R., Hobza, R., Vyskot, B., Oxelman, B., Charlesworth, D., Janousek, B., 2019. Evolution of sex determination and heterogamety changes in section Otites of the genus Silene. Scientific Reports 9, 1–13. https://doi.org/10.1038/s41598-018-37412-x

Bari, E.A., 1973. Cytological Studies in the Genus Silene L. New Phytologist 72, 833–838. https://doi.org/10.1111/j.1469-8137.1973.tb02059.x

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats, C., Eddy, S.R., 2004. The Pfam protein families database. Nucleic Acids Res 32, D138–D141. https://doi.org/10.1093/nar/gkh121

Bernasconi, G., Antonovics, J., Biere, A., Charlesworth, D., Delph, L.F., Filatov, D., Giraud, T., Hood, M.E., Marais, G. a. B., McCauley, D., Pannell, J.R., Shykoff, J.A., Vyskot, B., Wolfe, L.M., Widmer, A., 2009. Silene as a model system in ecology and evolution. Heredity 103, 5–14. https://doi.org/10.1038/hdy.2009.34

Bertrand, Y.J.K., Petri, A., Scheen, A.-C., Töpel, M., Oxelman, B., 2018. De novo transcriptome assembly, annotation, and identification of low-copy number genes in the flowering plant genus Silene (Caryophyllaceae). bioRxiv 290510. https://doi.org/10.1101/290510

Blavet, N., Charif, D., Oger-Desfeux, C., Marais, G.A., Widmer, A., 2011. Comparative high-throughput transcriptome sequencing and development of SiESTa, the Silene EST annotation database. BMC Genomics 12, 376. https://doi.org/10.1186/1471-2164-12-376

Bryant, D.M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M.B., Payzin-Dogru, D., Lee, T.J., Leigh, N.D., Kuo, T.-H., Davis, F.G., Bateman, J., Bryant, S., Guzikowski, A.R., Tsai, S.L., Coyne, S., Ye, W.W., Freeman, R.M., Peshkin, L., Tabin, C.J., Regev, A., Haas, B.J., Whited, J.L., 2017. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. Cell Reports 18, 762–776. https://doi.org/10.1016/j.celrep.2016.12.063

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421. https://doi.org/10.1186/1471-2105-10-421

Casimiro-Soriguer, I., Narbona, E., Buide, M.L., del Valle, J.C., Whittall, J.B., 2016. Transcriptome and Biochemical Analysis of a Flower Color Polymorphism in Silene littorea (Caryophyllaceae). Front. Plant Sci. 7. https://doi.org/10.3389/fpls.2016.00204

Charlesworth, D., 2006. Evolution of Plant Breeding Systems. Current Biology 16, R726–R735. https://doi.org/10.1016/j.cub.2006.07.068

Costich, D.E., Meagher, T.R., Yurkow, E.J., 1991. A rapid means of sex identification inSilene latifolia by use of flow cytometry. Plant Mol Biol Rep 9, 359–370. https://doi.org/10.1007/BF02672012

Dagher-Kharrat, M.B., Abdel-Samad, N., Douaihy, B., Bourge, M., Fridlender, A., Siljak-Yakovlev, S., Brown, S.C., 2013. Nuclear DNA C-values for biodiversity screening: Case of the Lebanese flora. Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology 147, 1228–1237. https://doi.org/10.1080/11263504.2013.861530

Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2012. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9, 772–772. https://doi.org/10.1038/nmeth.2109

Davis, S.L., Delph, L.F., 2005. Prior Selfing and Gynomonoecy in Silene noctiflora L. (Caryophyllaceae): Opportunities for Enhanced Outcrossing and Reproductive Assurance. International Journal of Plant Sciences 166, 475–480. https://doi.org/10.1086/428630

Desfeux, C., Maurice, S., Henry, J.P., Lejeune, B., Gouyon, P.H., 1996. Evolution of reproductive systems in the genus Silene. Proc. Biol. Sci. 263, 409–414. https://doi.org/10.1098/rspb.1996.0062

Doyle, J.J., Doyle, J.L., 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue [WWW Document]. PHYTOCHEMICAL BULLETIN. URL https://worldveg.tind.io/record/33886 (accessed 5.13.20).

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. https://doi.org/10.1371/journal.pone.0001386

Garraud, C., Brachi, B., Dufay, M., Touzet, P., Shykoff, J.A., 2011. Genetic determination of male sterility in gynodioecious Silene nutans. Heredity 106, 757–764. https://doi.org/10.1038/hdy.2010.116

Ghasemi, F.S., Jalili, A., Mirzadeh Vaghefi, S.S., 2015. CHROMOSOME REPORT OF THREE SPECIES OF FLORA OF IRAN. The Iranian Journal of Botany 21, 165–168.

Gholipour, A., Sheidai, M., 2010. Karyotype analysis and new chromosome number reports in Silene species (sect. Auriculatae, Caryophyllaceae). Biologia 65, 23–27. https://doi.org/10.2478/s11756-009-0215-3

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E.S., Jaffe, D.B., 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. U.S.A. 108, 1513–1518. https://doi.org/10.1073/pnas.1017351108

Gordon, S.P., Tseng, E., Salamov, A., Zhang, J., Meng, X., Zhao, Z., Kang, D., Underwood, J., Grigoriev, I.V., Figueroa, M., Schilling, J.S., Chen, F., Wang, Z., 2015. Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. PLoS One 10. https://doi.org/10.1371/journal.pone.0132628

Gregory, T.R., Nicol, J.A., Tamm, H., Kullman, B., Kullman, K., Leitch, I.J., Murray, B.G., Kapraun, D.F., Greilhuber, J., Bennett, M.D., 2007. Eukaryotic genome size databases. Nucleic Acids Res 35, D332–D338. https://doi.org/10.1093/nar/gkl828

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol 59, 307–321. https://doi.org/10.1093/sysbio/syq010

Guo, W., Grewe, F., Fan, W., Young, G.J., Knoop, V., Palmer, J.D., Mower, J.P., 2016. Ginkgo and Welwitschia Mitogenomes Reveal Extreme Contrasts in Gymnosperm Mitochondrial Evolution. Mol Biol Evol 33, 1448–1460. https://doi.org/10.1093/molbev/msw024

Hahn, M.W., Zhang, S.V., Moyle, L.C., 2014. Sequencing, Assembling, and Correcting Draft Genomes Using Recombinant Populations. G3 (Bethesda) 4, 669–679. https://doi.org/10.1534/g3.114.010264

Havird, J.C., Trapp, P., Miller, C.M., Bazos, I., Sloan, D.B., 2017. Causes and Consequences of Rapidly Evolving mtDNA in a Plant Lineage. Genome Biol Evol 9, 323–336. https://doi.org/10.1093/gbe/evx010

Havird, J.C., Whitehill Nicholas S., Snow Christopher D., Sloan Daniel B., 2015. Conservative and compensatory evolution in oxidative phosphorylation complexes of angiosperms with highly

divergent rates of mitochondrial genome evolution. Evolution 69, 3069–3081. https://doi.org/10.1111/evo.12808

Hestand, M.S., Houdt, J.V., Cristofoli, F., Vermeesch, J.R., 2016. Polymerase specific error rates and profiles identified by single molecule sequencing. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 784–785, 39–45. https://doi.org/10.1016/j.mrfmmm.2016.01.003

Jafari, F., Zarre, S., Gholipour, A., Eggens, F., Rabeler, R.K., Oxelman, B., 2020. A new taxonomic backbone for the infrageneric classification of the species-rich genus Silene (Caryophyllaceae). TAXON 69, 337–368. https://doi.org/10.1002/tax.12230

Jeffries, A.R., Leung, S., Castanho, I., Moore, K., Davies, J.P., Dempster, E.L., Bray, N.J., O'Neill, P., Tseng, E., Ahmed, Z., Collier, D., Prabhakar, S., Schalkwyk, L., Gandal, M.J., Hannon, E., Mill, J., 2020. Full-length transcript sequencing of human and mouse identifies widespread isoform diversity and alternative splicing in the cerebral cortex. bioRxiv 2020.10.14.339200. https://doi.org/10.1101/2020.10.14.339200

Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30, 772–780. https://doi.org/10.1093/molbev/mst010

Kemal, Y., Minareci, E., Çirpici, A., 2009. Karyotypic study on Silene, section Lasiostemones species from Turkey. Caryologia 62, 134–141. https://doi.org/10.1080/00087114.2004.10589678

Klaas, A.L., Olson, M.S., 2006. Spatial Distributions of Cytoplasmic Types and Sex Expression in Alaskan Populations of Silene acaulis. International Journal of Plant Sciences 167, 179–189. https://doi.org/10.1086/498965

Krasovec, M., Chester, M., Ridout, K., Filatov, D.A., 2018. The Mutation Rate and the Age of the Sex Chromosomes in Silene latifolia. Curr. Biol. 28, 1832-1838.e4. https://doi.org/10.1016/j.cub.2018.04.069

Kreibich, J.A., 2010. Using SQLite. O'Reilly Media, Inc.

Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305, 567–580. https://doi.org/10.1006/jmbi.2000.4315

Kruckeberg, A.R., 1960. CHROMOSOME NUMBERS IN SILENE (CARYOPHYLLACEAE). II. Madroño 15, 205–215.

Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.-H., Rognes, T., Ussery, D.W., 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35, 3100–3108. https://doi.org/10.1093/nar/gkm160

Lan, T., Renner, T., Ibarra-Laclette, E., Farr, K.M., Chang, T.-H., Cervantes-Pérez, S.A., Zheng, C., Sankoff, D., Tang, H., Purbojati, R.W., Putra, A., Drautz-Moses, D.I., Schuster, S.C., Herrera-Estrella, L., Albert, V.A., 2017. Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome. Proc Natl Acad Sci U S A 114, E4435–E4441. https://doi.org/10.1073/pnas.1702072114

Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. https://doi.org/10.14806/ej.17.1.200

McNeill, J., 1980. THE BIOLOGY OF CANADIAN WEEDS.: 46. Silene noctiflora L. Can. J. Plant Sci. 60, 1243–1253. https://doi.org/10.4141/cjps80-177

Mirzadeh Vaghefi, S.S., Jalili, A., 2019. CHROMOSOME NUMBERS OF SOME VASCULAR PLANT SPECIES FROM IRAN. The Iranian Journal of Botany 25, 140–144. https://doi.org/10.22092/ijb.2019.126775.1248

Moiloa, N.A., Mesbah, M., Nylinder, S., Manning, J., Forest, F., de Boer, H.J., Bacon, C.D., Oxelman, B., 2021. Biogeographic origins of southern African Silene (Caryophyllaceae). Molecular Phylogenetics and Evolution 107199. https://doi.org/10.1016/j.ympev.2021.107199

Mower, J.P., Touzet, P., Gummow, J.S., Delph, L.F., Palmer, J.D., 2007. Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. BMC Evol. Biol. 7, 135. https://doi.org/10.1186/1471-2148-7-135

Muyle, A., Zemp, N., Deschamps, C., Mousset, S., Widmer, A., Marais, G.A.B., 2012. Rapid De Novo Evolution of X Chromosome Dosage Compensation in Silene latifolia, a Plant with Young Sex Chromosomes. PLOS Biology 10, e1001308. https://doi.org/10.1371/journal.pbio.1001308

Nishimura, K., van Wijk, K.J., 2015. Organization, function and substrates of the essential Clp protease system in plastids. Biochimica et Biophysica Acta (BBA) - Bioenergetics, SI: Chloroplast Biogenesis 1847, 915–930. https://doi.org/10.1016/j.bbabio.2014.11.012

Olson, M.S., Mccauley, D.E., 2002. Mitochondrial Dna Diversity, Population Structure, and Gender Association in the Gynodioecious Plant Silene Vulgaris. Evolution 56, 253–262. https://doi.org/10.1111/j.0014-3820.2002.tb01335.x

Ono, Y., Asai, K., Hamada, M., 2013. PBSIM: PacBio reads simulator—toward accurate genome assembly. Bioinformatics 29, 119–121. https://doi.org/10.1093/bioinformatics/bts649

PacificBiosciences, 2020. IsoSeq. Pacific Biosciences.

Papadopulos, A.S.T., Chester, M., Ridout, K., Filatov, D.A., 2015. Rapid Y degeneration and dosage compensation in plant sex chromosomes. PNAS 112, 13021–13026. https://doi.org/10.1073/pnas.1508454112

Pellicer, J., Leitch, I.J., 2020. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. New Phytologist 226, 301–305. https://doi.org/10.1111/nph.16261

Petersen, T.N., Brunak, S., von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods 8, 785–786. https://doi.org/10.1038/nmeth.1701

Popp, M., Erixon, P., Eggens, F., Oxelman, B., 2005. Origin and Evolution of a Circumpolar Polyploid Species Complex in Silene (Caryophyllaceae) Inferred from Low Copy Nuclear RNA Polymerase Introns, rDNA, and Chloroplast DNA. Systematic Botany 30, 302–313.

Popp, M., Oxelman, B., 2007. Origin and evolution of North American polyploid Silene (Caryophyllaceae). American Journal of Botany 94, 330–349. https://doi.org/10.3732/ajb.94.3.330

Popp, M., Oxelman, B., 2001. Inferring the History of the Polyploid Silene aegaea (Caryophyllaceae) Using Plastid and Homoeologous Nuclear DNA Sequences. Molecular Phylogenetics and Evolution 20, 474–481. https://doi.org/10.1006/mpev.2001.0977

Potter, S.C., Luciani, A., Eddy, S.R., Park, Y., Lopez, R., Finn, R.D., 2018. HMMER web server: 2018 update. Nucleic Acids Res 46, W200–W204. https://doi.org/10.1093/nar/gky448

Rhoads, A., Au, K.F., 2015. PacBio Sequencing and Its Applications. Genomics, Proteomics & Bioinformatics, SI: Metagenomics of Marine Environments 13, 278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Rockenbach, K., Havird, J.C., Monroe, J.G., Triant, D.A., Taylor, D.R., Sloan, D.B., 2016. Positive Selection in Rapidly Evolving Plastid–Nuclear Enzyme Complexes. Genetics 204, 1507–1522. https://doi.org/10.1534/genetics.116.188268

Schatz, M.C., Witkowski, J., McCombie, W.R., 2012. Current challenges in de novo plant genome sequencing and assembly. Genome Biol 13, 243. https://doi.org/10.1186/gb-2012-13-4-243

Seppey, M., Manni, M., Zdobnov, E.M., 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness, in: Kollmar, M. (Ed.), Gene Prediction: Methods and Protocols, Methods in Molecular Biology. Springer, New York, NY, pp. 227–245. https://doi.org/10.1007/978-1-4939-9173-0_14

Siroký, J., Lysák, M.A., Dolezel, J., Kejnovský, E., Vyskot, B., 2001. Heterogeneity of rDNA distribution and genome size in Silene spp. Chromosome Res. 9, 387–393. https://doi.org/10.1023/a:1016783501674

Sloan, D.B., Alverson, A.J., Chuckalovcak, J.P., Wu, M., McCauley, D.E., Palmer, J.D., Taylor, D.R., 2012a. Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant

Mitochondria with Exceptionally High Mutation Rates. PLOS Biology 10, e1001241. https://doi.org/10.1371/journal.pbio.1001241

Sloan, D.B., Keller, S.R., Berardi, A.E., Sanderson, B.J., Karpovich, J.F., Taylor, D.R., 2012b. De novo transcriptome assembly and polymorphism detection in the flowering plant Silene vulgaris (Caryophyllaceae). Molecular Ecology Resources 12, 333–343. https://doi.org/10.1111/j.1755-0998.2011.03079.x

Sloan, D.B., Triant, D.A., Forrester, N.J., Bergner, L.M., Wu, M., Taylor, D.R., 2014. A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae (Caryophyllaceae). Molecular Phylogenetics and Evolution 72, 82–89. https://doi.org/10.1016/j.ympev.2013.12.004

Städler, T., Delph, L.F., 2002. Ancient mitochondrial haplotypes and evidence for intragenic recombination in a gynodioecious plant. PNAS 99, 11730–11735. https://doi.org/10.1073/pnas.182267799

Stanne, T.M., Pojidaeva, E., Andersson, F.I., Clarke, A.K., 2007. Distinctive Types of ATP-dependent Clp Proteases in Cyanobacteria. J. Biol. Chem. 282, 14394–14402. https://doi.org/10.1074/jbc.M700275200

UniProt: a hub for protein information, 2015. . Nucleic Acids Res 43, D204–D212. https://doi.org/10.1093/nar/gku989

Wang, B., Kumar, V., Olson, A., Ware, D., 2019. Reviving the Transcriptome Studies: An Insight Into the Emergence of Single-Molecule Transcriptome Sequencing. Front. Genet. 10. https://doi.org/10.3389/fgene.2019.00384

Wang, B., Tseng, E., Baybayan, P., Eng, K., Regulski, M., Jiao, Y., Wang, L., Olson, A., Chougule, K., Buren, P.V., Ware, D., 2020. Variant phasing and haplotypic expression from long-read sequencing in maize. Communications Biology 3, 1–11. https://doi.org/10.1038/s42003-020-0805-8

Wang, B., Tseng, E., Regulski, M., Clark, T.A., Hon, T., Jiao, Y., Lu, Z., Olson, A., Stein, J.C., Ware, D., 2016. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nature Communications 7, 1–13. https://doi.org/10.1038/ncomms11708

Weirather, J.L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., Buck, D., Au, K.F., 2017. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Res 6. https://doi.org/10.12688/f1000research.10571.2

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A.M., Schatz, M.C., Myers, G., DePristo, M.A., Ruan, J., Marschall, T., Sedlazeck, F.J., Zook, J.M., Li, H., Koren, S., Carroll, A., Rank, D.R., Hunkapiller, M.W., 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology 37, 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Williams, A.M., Friso, G., Wijk, K.J. van, Sloan, D.B., 2019. Extreme variation in rates of evolution in the plastid Clp protease complex. The Plant Journal 98, 243–259. https://doi.org/10.1111/tpj.14208

Xu, Z., Peters, R.J., Weirather, J., Luo, H., Liao, B., Zhang, X., Zhu, Y., Ji, A., Zhang, B., Hu, S., Au, K.F., Song, J., Chen, S., 2015. Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of Salvia miltiorrhiza and tanshinone biosynthesis. The Plant Journal 82, 951–961. https://doi.org/10.1111/tpj.12865

Yildiz, K., Minareci, E., Çirpici, A., Dadandı, M.Y., 2008. A karyotypic study on Silene, section Siphonomorpha species of Turkey. Nordic Journal of Botany 26, 368–374. https://doi.org/10.1111/j.1756-1051.2008.00289.x

Yu, A.Y.H., Houry, W.A., 2007. ClpP: A distinctive family of cylindrical energy-dependent serine proteases. FEBS Letters 581, 3749–3757. https://doi.org/10.1016/j.febslet.2007.04.076

Zhao, L., Zhang, H., Kohnen, M.V., Prasad, K.V.S.K., Gu, L., Reddy, A.S.N., 2019. Analysis of Transcriptome and Epitranscriptome in Plants Using PacBio Iso-Seq and Nanopore-Based Direct RNA Sequencing. Front Genet 10, 253. https://doi.org/10.3389/fgene.2019.00253

Zheng, B., Halperin, T., Hruskova-Heidingsfeldova, O., Adam, Z., Clarke, A.K., 2002. Characterization of Chloroplast Clp proteins in Arabidopsis: Localization, tissue specificity and stress responses. Physiologia Plantarum 114, 92–101. https://doi.org/10.1034/j.1399-3054.2002.1140113.x

Zhu, B.-H., Xiao, J., Xue, W., Xu, G.-C., Sun, M.-Y., Li, J.-T., 2018. P_RNA_scaffolder: a fast and accurate genome scaffolder using paired-end RNA-sequencing reads. BMC Genomics 19, 175. https://doi.org/10.1186/s12864-018-4567-3

CHAPTER 4: GENE DUPLICATION AND RATE VARIATION IN THE EVOLUTION OF NON-PHOTOSYNTHETIC PATHWAYS IN PLASTIDS[3]

Summary

While the chloroplast (plastid) is known for its role in photosynthesis, it is also involved in many other metabolic pathways essential for plant survival. As such, plastids contain an extensive suite of enzymes required for non-photosynthetic processes. The evolution of the associated genes has been especially dynamic in flowering plants (angiosperms), including examples of gene duplication and extensive rate variation. We examined the role of ongoing gene duplication in two key plastid enzymes, the acetyl-CoA carboxylase (ACCase) and the caseinolytic protease (Clp), responsible for fatty acid biosynthesis and protein turnover, respectively. In plants, there are two ACCase complexes—a homomeric version present in the cytosol and a heteromeric version present in the plastid. Duplications of the nuclear-encoded homomeric ACCase gene and retargeting of one resultant protein to the plastid have been previously reported in multiple species. We find that these genes encoding retargeted homomeric ACCase proteins exhibit elevated rates of sequence evolution, consistent with neofunctionalization and/or relaxation of selection. The plastid Clp complex catalytic core is composed of nine paralogous proteins that arose via ancient gene duplication in the cyanobacterial/plastid lineage. We show that further gene duplication occurred more recently in the nuclear-encoded core subunits of this complex, yielding additional paralogs in many species of angiosperms. Moreover, in six of eight cases, subunits that have undergone recent duplication display increased rates of sequence evolution relative to those that have remained single copy. We also compared rate patterns between pairs of Clp core paralogs to gain insight into post-duplication evolutionary routes. These results show that gene duplication and rate variation continue to shape the plastid proteome.

[3] Submitted to *Molecular Phylogenetics and Evolution* in September 2021.
Authors: Alissa M. Williams, Olivia G. Carter, Evan S. Forsythe, Hannah K. Mendoza, Daniel B. Sloan

Introduction:

The plastid is a dynamic proteomic environment in which key photosynthetic and non-photosynthetic biochemical reactions occur. Major non-photosynthetic functions of plastids include the reaction catalyzed by the acetyl-CoA carboxylase (ACCase) enzyme and protein degradation performed by the caseinolytic protease (Clp) complex (Caroca et al., 2021; Green, 2011; Konishi et al., 1996; Nishimura et al., 2017; Nishimura and van Wijk, 2015). Both of these functions are essential in plants and thus the genes involved are generally highly conserved; however, these genes have undergone rapid evolution in multiple angiosperm species (Barnard-Kubow et al., 2014; Erixon and Oxelman, 2008; Jansen et al., 2007; Park et al., 2017; Sloan et al., 2014, 2014; Wicke et al., 2011; Williams et al., 2019, 2015; Zhang et al., 2014). While many hypotheses about these patterns of accelerated evolution have been posited, the underlying evolutionary mechanisms, causes, and consequences remain largely unknown.

The ACCase enzyme catalyzes the first committed step of fatty acid biosynthesis, the carboxylation of acetyl-CoA to malonyl-CoA (Salie and Thelen, 2016; Sasaki and Nagano, 2004). This step requires four different enzyme functions—one biotin carboxylase, one biotin carboxyl carrier, and two (α and β) carboxyltransferases (Salie and Thelen, 2016; Sasaki and Nagano, 2004; Schulte et al., 1997). In plants, there are two forms of the ACCase enzyme. The homomeric version, present in the cytosol, is encoded by a single nuclear gene (Konishi et al., 1996; Konishi and Sasaki, 1994). The heteromeric version, present in the plastid, is encoded by five genes in *Arabidopsis thaliana*; each function is encoded by a single gene except for the biotin carboxyl carrier, which is encoded by two genes in *Arabidopsis thaliana* (Konishi et al., 1996; Konishi and Sasaki, 1994; Salie and Thelen, 2016). Four of these genes are in the nuclear genome while the fifth (*accD*) is in the plastid genome (Caroca et al., 2021; Sasaki and Nagano, 2004). In a few angiosperm lineages, including the Brassicaceae, Caryophyllaceae, Geraniaceae, and Poaceae, there have been duplications of the homomeric ACCase gene with subsequent retargeting of one resultant protein to the plastid (Babiychuk et al., 2011; Konishi and Sasaki, 1994; Park et al., 2017; Parker et al., 2014; Rockenbach et al., 2016; Schulte et al., 1997).

The Clp complex is one of the most abundant stromal proteases and degrades a variety of targets (Apitz et al., 2016; Bouchnak and van Wijk, 2021; Majeran et al., 2000; Montandon et al., 2019; Nishimura et al., 2017; Nishimura and van Wijk, 2015; Welsch et al., 2018). This complex consists of many types of subunits. Adapters bind proteins targeted for degradation and deliver them to chaperones, which use ATP to unfold the targeted proteins into the proteolytic core of the complex (Nishimura and van Wijk, 2015). The core consists of 14 subunits that are encoded by nine different paralogous genes (Olinares et al., 2011a; Peltier et al., 2004; Sjögren et al., 2006; Stanne et al., 2007). Eight of these genes reside in the nuclear genome (*CLPP3-6, CLPR1-4*), while the ninth is encoded in the plastid genome (*clpP1*) (Nishimura et al., 2017; Olinares et al., 2011b). The ClpP subunits contain a catalytically active Ser-His-Asp triad, whereas the ClpR subunits do not (Nishimura and van Wijk, 2015; Porankiewicz et al., 1999). These nine paralogs are the results of gene duplications throughout cyanobacterial and plastid evolution and are shared by all land plants (Olinares et al., 2011a). Ongoing gene duplication of individual subunits has been noted in a handful of angiosperm lineages (Rockenbach et al., 2016; Williams et al., 2021, 2019).

Thus, the evolutionary trajectory of both the plastid ACCase and the plastid Clp complex is characterized by gene duplication at both ancient and recent timescales. Gene duplication is common in land plants, in part due to the frequency with which whole genome duplication (polyploidization) occurs in this lineage (Clark and Donoghue, 2018; De Bodt et al., 2005; del Pozo and Ramirez-Parra, 2015; Flagel and Wendel, 2009; Panchy et al., 2016; Wendel et al., 2018). Nearly all species of land plants have polyploidization events in their evolutionary histories (Clark and Donoghue, 2018; Leebens-Mack et al., 2019; Panchy et al., 2016). Angiosperms in particular seem to have a propensity for whole genome duplication; the entire clade shares an ancient polyploidization event and many lineages have undergone subsequent rounds of whole genome duplication (Clark and Donoghue, 2018; Panchy et al., 2016; Renny-Byfield and Wendel, 2014; Soltis et al., 2009). While every gene is initially affected by whole genome duplication, only 10-30% of those duplicates are maintained in the genome longer-term (Hahn, 2009; Maere et al., 2005; Paterson et al., 2006). Though polyploidy is likely a main contributor to gene

duplication in plants, other forms of gene duplication are also common (Flagel and Wendel, 2009). For instance, tandem duplication has been shown to be common in both *Arabidopsis thaliana* and *Oryza sativa*, where tandemly arrayed gene clusters make up 15-20% of genic content. Additionally, multiple studies have shown that transposon-mediated gene duplication occurs frequently in plants (Flagel and Wendel, 2009; Freeling et al., 2008; Rizzon et al., 2006; Wang et al., 2006).

Gene duplication is an important evolutionary process and is thought to be a major source of evolutionary novelty (Hahn, 2009; Ohno, 1970; Taylor and Raes, 2004; Zhang, 2003). The most common evolutionary fate of paralogs is retention of one copy and pseudogenization and loss of the other copy (Lynch and Conery, 2000; Zhang, 2003; Zhang et al., 2003). However, several evolutionary mechanisms have been described in which retention of both gene duplicates is favored. The increased gene-dosage advantage model describes a scenario in which increased amount of gene product produced by the two essentially identical gene copies is beneficial and thus both copies retain ancestral function (Hahn, 2009; Ohno, 1970; Pegueroles et al., 2013; Zhang, 2003). The neofunctionalization model posits that one paralog acquires new functions while the other retains ancestral functionality (Hahn, 2009; Ohno, 1970; Pegueroles et al., 2013; Zhang, 2003). In the subfunctionalization model, an ancestral function is split between the two duplicates (Hahn, 2009; Ohno, 1970; Pegueroles et al., 2013; Zhang, 2003), in some cases creating the possibility for each paralog to optimize a subset of the ancestral function in a process known as escape from adaptive conflict (Des Marais and Rausher, 2008; Huang et al., 2015; Sikosek et al., 2012).

To distinguish between these evolutionary fates, many studies have employed evolutionary rate comparisons (Hahn, 2009; Pegueroles et al., 2013). These comparisons involve both paralogs as well as their common ancestor (Pegueroles et al., 2013). Under both the gene-dosage advantage and subfunctionalization models, gene duplicates are expected to evolve at approximately the same rate as each other (Pegueroles et al., 2013). The difference in evolutionary rates predicted by these two models is found in comparisons to the common ancestor; with a gene-dosage advantage, the expectation is that the paralogs will evolve at the same rate as the common ancestor, while with subfunctionalization, the

expectation is that the paralogs will evolve at an increased rate relative to the common ancestor (though this assumption has been challenged) (Force et al., 1999; Hahn, 2009; He and Zhang, 2005; Lynch and Force, 2000; Pegueroles et al., 2013; Zhang, 2003). By contrast, under the neofunctionalization model, asymmetry between evolutionary rates of paralogs is expected, where one paralog retains the ancestral evolutionary rate while the other experiences rate acceleration after being freed from selective constraints (Hahn, 2009; Pegueroles et al., 2013; Zhang, 2003). The proportion of paralogs with asymmetric rates of evolution has been estimated at anywhere from 5% to 65% in a variety of studies (Conant and Wagner, 2003; Dermitzakis and Clark, 2001; Kondrashov et al., 2002; Panchin et al., 2010; Pegueroles et al., 2013; Van de Peer et al., 2001). This wide range of estimates is likely due to differences in study systems, definitions and identifications of paralogs, gene types, and time since duplication. Despite the varying estimates of evolutionary rate asymmetry, it is clear that paralogs evolve under a mixture of evolutionary regimes.

Here, we characterize recent gene duplication events and subsequent changes in evolutionary rate in ACCase and Clp core subunits. We show that ACCase genes exhibit patterns of duplication, protein retargeting, and accelerated evolution consistent with neofunctionalization and/or relaxed selection. Additionally, we examine duplications of nuclear-encoded plastid Clp core subunits and demonstrate that duplication leads to significant changes in the rate of evolution in most cases but that patterns differ across Clp subunits, meaning multiple post-duplication evolutionary routes are represented across pairs of paralogs. This work provides additional insights into the interplay between gene duplication and evolutionary rate in the molecular evolution of plastid proteins.

Materials and Methods:

Compilation and curation of *ACC* nucleotide sequences

Previous work identified duplications of the homomeric ACCase gene *ACC* and subsequent retargeting of one resultant protein to the plastid in the angiosperm families Poaceae (Konishi and Sasaki,

1994; Park et al., 2017; Rockenbach et al., 2016), Brassicaceae (Babiychuk et al., 2011; Park et al., 2017; Parker et al., 2014; Schulte et al., 1997), Caryophyllaceae (Rockenbach et al., 2016), and Geraniaceae (Park et al., 2017). *ACC* sequences were obtained for multiple species in each of these families. All cytosol-targeted *ACC* genes were designated *ACC1* while all plastid-targeted *ACC* genes were designated *ACC2* per established conventions (Babiychuk et al., 2011; Sasaki and Nagano, 2004); thus, sharing the same identifier does not necessarily indicate orthology because of the multiple independent origins of plastid-targeted *ACC2* genes. *Amborella trichopoda*, which has a single *ACC* gene that we designated *ACC1*, was used as an outgroup.

Trimmed *ACC1* and *ACC2* coding sequences (CDSs) were obtained from Rockenbach et al. (2016) for *Amborella trichopoda, Arabidopsis thaliana, Agrostemma githago*, *Silene noctiflora*, *Silene paradoxa,* and *Triticum aestivum*. The trimming in Rockenbach et al (2016) was codon-guided and included removal of the target peptide. *ACC1* and *ACC2* CDSs from the following species were compiled using gene identifiers from Table S4 in Park et al. (2017): Geraniaceae: *California macrophylla, Erodium texanum, Geranium incanum, Geranium maderense, Geranium phaeum, Monsonia emarginata, Pelargonium cotyledonis*; Brassicaceae: *Capsella rubella*; Poaceae: *Oryza sativa*, *Sorghum bicolor*. Duplications of *ACC* were additionally identified in two Poaceae species—*Aegilops tauschii* and *Zea mays*—by performing BLAST searches against the genomes of these organisms on NCBI and Phytozome v13, respectively (Camacho et al., 2009; Goodstein et al., 2012).

All *ACC1* and *ACC2* sequences were included in a single file and aligned using the MAFFT *einsi* option (Katoh and Standley, 2013) in codon space using the *align_fasta_with_mafft_codon* subroutine in the sloan.pm Perl module (https://github.com/dbsloan/perl_modules). 5′ trimming was conducted according to the trimming performed in Rockenbach et al. (2016). Additional trimming of poorly aligned regions was performed manually in a codon-based manner.

Compilation and curation of Clp core subunit amino acid and nucleotide sequences

To identify Clp core subunit amino acid sequences, a custom Python script (https://github.com/alissawilliams/Gene_duplication_ACCase_Clp/scripts /local_blast5.py) was used to reciprocally blast (blastp v2.2.29) *Arabidopsis thaliana* amino acid sequences against predicted protein sequences from each of 22 other angiosperm species in the dataset. These 22 species were the same set used in Williams et al. (2019) with the exclusion of *Silene latifolia* and *Silene noctiflora*, since Clp core subunit duplications have been previously studied in *Sileneae* (Rockenbach et al., 2016; Williams et al., 2021, 2019). This sampling was chosen to represent both the diversity of angiosperms and the range of rate variation in Clp complex evolution (Williams et al., 2019; see Table S3).

Compiled amino acid sequences for each subunit were aligned using the *einsi* option in MAFFT v7.222 (Katoh and Standley, 2013) and trimmed using GBLOCKS v0.91b (Castresana, 2000) with parameter *-b1* set to the default value of *-b2* and parameter *-b5* set to *h*. All alignments were examined manually to confirm homology. Sequences were also screened to prevent inclusion of multiple splice variants from a single gene. In cases where genomic data were used, only one transcript per gene was used. In cases where transcriptomic data were used, sequences were eliminated when alternative splicing was obvious (i.e. inclusion of an intron where the other sequence had a gap or variation only in one short piece of the transcript at either end). Catalytic site status and length were determined using the amino acid sequence data.

Nucleotide sequences for each identified Clp core subunit protein sequence were compiled from the corresponding CDS or transcript sequence file. For non-CDS sequences, ORFfinder (Wheeler et al., 2003) was used to identify the coding sequence. Compiled CDS sequences for each subunit were aligned with the MAFFT *einsi* option (Katoh and Standley, 2013) in codon space as above. 5′ and 3′ end trimming was performed manually in a codon-based manner.

Generating constraint trees for the *ACC* and Clp subunit alignments for use in PAML

A constraint tree stipulates a fixed topology (branching order) that is used by a phylogenetic program (in this case, PAML) when calculating branch lengths. To generate a constraint tree for the *ACC*

alignment, RAxML v8.2.12 (Stamatakis, 2014) was used on the trimmed nucleotide alignment with parameters *-m = GTRGAMMA, -p = 12345, -f = a, -x = 12345,* and *-# = 100*. The resultant topology confirmed that there were independent *ACC* duplications at the base of each family (Park et al., 2017; Rockenbach et al., 2016).

To construct constraint trees for Clp core subunits, each trimmed amino acid alignment was analyzed with ProtTest v3.4.2 (Darriba et al., 2011) to choose a model of sequence evolution. The top model based on the Bayesian Information Criterion was chosen for use in PhyML v3.3 (Guindon et al., 2010), which was run with 1000 bootstrap replicates and 100 random starts. The resultant phylogenetic trees were used to determine whether duplication events were lineage-specific or shared among species in the dataset. In almost all cases, paralogs from a single species were sister to one another in the trees, indicating lineage-specific duplications. There were a few cases in which paralogs from a single species were not sister to one another. However, given low bootstrap support and the difficulty of resolving species relationships using a single gene with highly variable rates of evolution, we proceeded under the assumption that these duplications were lineage-specific as well. Thus, the constraint trees for each individual Clp core subunit were constructed using the known species tree (The Angiosperm Phylogeny Group et al., 2016), with duplications encoded as species-specific (mapped to terminal branches of the species tree).

Running PAML for ACCase and Clp core subunit genes

For each alignment, PAML v4.9j (Yang, 2007) was used to infer d_N/d_S values for all branches using the free ratios model (model = 1) and parameters *CodonFreq = 2* and *cleandata = 0*. Additionally, *model = 0* and *model = 2* runs were conducted for all alignments, again using *CodonFreq = 2* and *cleandata = 0*. The *model = 0* runs forced all branches to have the same d_N/d_S ratio, while the *model = 2* runs allowed different d_N/d_S values for specified groups of branches.

For the *ACC* alignment, one *model = 2* run was conducted with plastid-targeted branches as the foreground. The resultant tree had one d_N/d_S value for plastid-targeted (*ACC2*) branches and a second

d_N/d_S value for cytosol-targeted (*ACC1*) branches (including all internal pre-duplication branches). This output was compared with the *model = 0* run to determine whether allowing two d_N/d_S ratios (one for each of those groups) was a better fit to the data than allowing just a single d_N/d_S value. For the Clp subunit alignments, *model = 2* was used twice. In the first run, all terminal branches (and in the case of two subunits, internal post-duplication branches) were designated as the foreground. In the second run, there were three classes of branches, where all branches were categorized the same as in the first run except that post-duplication branches (internal or terminal) were placed in a third category. The three-partition and two-partition models were compared to determine whether allowing an additional d_N/d_S ratio for post-duplication branches was a better fit to the data than just separating terminal from internal branches. The models were compared using likelihood ratio tests.

For the *ACC* alignment, a branch-site test (Yang, 2007; Yang and Nielsen, 2002) was also conducted to test for evidence of positive selection on branches for plastid-targeted genes, which were set as the foreground branches for this analysis. A null model and an alternative model both used the parameters *model = 2, NSsites = 2, CodonFreq = 2*, and *cleandata = 0*. The alternative model otherwise used all default values, while the null model additionally used *fix_omega = 1* and *omega = 1*. The models were compared using a likelihood ratio test.

Running HyPhy for *ACC*

In addition to running a PAML branch-site test on the *ACC* alignment (Yang, 2007; Yang and Nielsen, 2002), tests for positive and relaxed selection were implemented in HyPhy v2.5.32 (Kosakovsky Pond et al., 2020). Positive selection was tested for using the aBSREL and BUSTED methods (Murrell et al., 2015; Smith et al., 2015). The RELAX method was used to test for relaxed vs. intensified selection (Wertheim et al., 2015). As with the PAML runs, the constraint tree used for HyPhy methods had the branches separated into two categories (*ACC1* and *ACC2*).

Comparisons between *ACC1* and *ACC2* genes

To compare d_N and d_S between cytosolic-targeted and plastid-targeted *ACC* genes (*ACC1* and *ACC2*, respectively), a mean root-to-tip distance was calculated for each family in the tree. The base of each duplication event was used as the root for each family. For both d_N and d_S, the four mean distances for *ACC1* were compared to those of *ACC2* using a paired t-test in R. Because of the *a priori* prediction that retargeting to the plastid would be associated with accelerated protein sequence evolution, a one-sided test (*ACC2* > *ACC1*) was used for d_N, while a two-sided test was used for d_S.

Fisher's exact test on Clp subunit paralogs

Using the output from the free ratios (model = 1) PAML runs, Fisher's exact test was used to test for asymmetry in the ratio of the estimated numbers of nonsynonymous and synonymous substitutions (Pegueroles et al., 2013). Nonsynonymous and synonymous substitution estimates were entered into the *fisher.test()* function in R with default parameters. For each pair of duplicates, a test between paralog 1 and paralog 2 was performed (**Figure 4.1**). If the paralogs were found to be evolving symmetrically, their combined numbers of substitutions were compared to those of the ancestral branch (**Figure 4.1**). If the paralogs were found to be evolving asymmetrically, each one was compared individually against the ancestral branch (**Figure 4.1**). The four cases in which there were more than two species-specific paralogs (*Soja max* and *Gossypium raimondii CLPP5*; *Musa acuminata* and *Vitis vinifera CLPR4*) were excluded from this analysis.

Data availability

Scripts, untrimmed and trimmed alignments, PAML output, and HyPhy output are provided for both *ACC* and Clp subunits at https://github.com/alissawilliams/Gene_duplication_ACCase_Clp.

Results:

Plastid-targeted ACCases evolve more rapidly than cytosol-targeted ACCases across angiosperms

Across the sampled clades (Geraniaceae, Caryophyllaceae, Brassicaceae, and Poaceae), nearly all plastid-targeted *ACC2* genes have higher d_N/d_S values than their cytosol-targeted *ACC1* counterparts (**Figure 4.2, Figure S4.1**). The single-partition model assigned all branches a d_N/d_S value of 0.1266, while the two-partition model assigned *ACC1* branches a value of 0.0883 and *ACC2* branches a value of 0.1936 ($\chi^2 = 466.84$, $p << 0.0001$). This pattern is true for both terminal and internal branches. The increase in in d_N/d_S ratios in *ACC2* branches is generally driven by increases in d_N rather than reductions in d_S ($t = 4.48$, $p = 0.01$ for d_N; $t = 0.72$, $p = 0.5249$ for d_S; **Figure 4.2, Figure S4.2**), suggesting changes in selective pressure.

Using a branch-sites test in PAML (Yang, 2007), we did not find a significant signature of positive selection spanning the alignment ($\chi^2 = 0$, $p = 1$), although there were multiple individual sites found to be under positive selection (**Table S4.1**). Two HyPhy methods found limited, though significant, evidence for positive selection—the aBSREL run (Smith et al., 2015) detected one branch under positive selection ($p = 0.04$) and the BUSTED run (Murrell et al., 2015) assigned 0.12% of sites in foreground (*ACC2*) branches to the positive selection class relative to 0.05% of sites in background (*ACC1*) branches ($p = 0.0026$). The HyPhy RELAX method (Wertheim et al., 2015) found significant evidence for relaxed selection in the *ACC2* branches relative to the rest of the tree ($K = 0.09$, $p << 0.001$).

Characterizing ongoing duplication of nuclear-encoded Clp core subunit genes in angiosperms

Of the 23 angiosperm species in our dataset, 11 had one or more duplications of nuclear genes encoding Clp core subunits, and all eight of these genes were duplicated in at least one species (**Figure 4.3**). Most of these duplications were represented by two paralogs, but in four cases, we identified more than two paralogs for a particular subunit in a particular species. For *CLPP5*, *Soja max* and *Gossypium raimondii* have five and seven copies, respectively, and for *CLPR4*, both *Musa acuminata* and *Vitis vinifera* have four copies.

Soja max had duplications of the largest number of subunits (six of eight), followed by *Plantago maritima* and *Populus trichocarpa* with duplications of five subunits. Of the 11 species with duplications,

Eucalyptus grandis and *Oenothera biennis* were the only species that had duplications of just one subunit. Across subunits, *CLPP5* had the highest number of paralogs (37 in 23 species) and *CLPR2* had the lowest (24 in 23 species).

In total, we identified 72 gene copies of Clp core subunits resulting from duplication events, including 40 catalytic subunits (*CLPP3*-*CLPP6*) (**Figure 4.3**). Of the 40 paralogs of catalytic subunits, we found evidence of loss of one or more catalytic sites in multiple genes (**Table S4.2**). Across all 72 paralogs, we also found evidence of truncation of multiple different gene copies (including some with catalytic site loss) (**Table S4.2, Table S4.3**).

Recent paralogs of Clp core subunits tend to have higher rates of protein sequence evolution than their single-copy counterparts

Out of the eight nuclear-encoded Clp core subunit trees (**Figures S4.3-S4.10**), seven showed statistically significant differences between a model that allowed for different d_N/d_S rates in gene duplicates vs. single-copy genes (the three-partition model) and one that forces the same d_N/d_S rate on these two types of branches (the two-partition model) based on an uncorrected significance threshold of $p = 0.05$. (**Figure 4.4, Table 4.1**). In six of those cases, duplicated terminal branches had a higher d_N/d_S rate than non-duplicated terminal branches, while in the remaining case, the reverse was true. We separated internal branches from terminal branches to account for the fact that terminal branches will, on average, have higher d_N/d_S estimates than internal branches because selection has had more time to act on older deleterious mutations (Hasegawa et al., 1998; Ho et al., 2005). Further, terminal branches represent both interspecific divergence and intraspecific polymorphism, which is important because the latter inflates evolutionary rate calculations (Ho et al., 2005; Moilanen and Majamaa, 2003; Nielsen and Weinreich, 1999).

We also compared the evolutionary rates of paralogs to one another as well as to their common ancestor, again using an uncorrected significance threshold of $p = 0.05$ (**Table 4.2**). Of the 26 pairs of paralogs, 13 (50%) showed statistically significant rate asymmetry relative to each another. In 10 (77%)

of those cases, only one paralog had a significantly different evolutionary rate than the common ancestor (and in all 10 of those cases, that paralog was evolving at a faster rate than the common ancestor). Of the 13 pairs with symmetric evolutionary rates, five (38%) were asymmetric relative to the common ancestor. In three of those cases, the combined paralog evolutionary rate was significantly faster than that of the ancestor.

Discussion:

Neofunctionalization and accelerated evolution of duplicated *ACC* genes in multiple clades of flowering plants

Independent duplications of *ACC* and subsequent retargeting events have been previously reported in multiple angiosperm clades (Babiychuk et al., 2011; Konishi and Sasaki, 1994; Park et al., 2017; Parker et al., 2014; Rockenbach et al., 2016; Schulte et al., 1997). The process of retargeting of a paralog is inherently a form of neofunctionalization because the newly retargeted protein functions in a different cellular compartment than it did ancestrally. A hallmark of neofunctionalization is evolutionary rate asymmetry between paralogs due to selection associated with gaining a new function (Hahn, 2009; Pegueroles et al., 2013; Zhang, 2003). We found that branches of our *ACC* tree representing proteins targeted to the plastid had statistically significantly higher d_N/d_S values than branches representing paralogs targeted to the cytosol (**Figure 4.2, Figure S4.1**), consistent with the predictions under neofunctionalization. These results were based on a trimmed alignment lacking the target peptide, which we excluded because target peptides exhibit fast rates of evolution and reduced constraints on primary amino acid sequence (Bruce, 2001, 2000; Jarvis, 2008). Thus, our results show that *ACC* genes encoding proteins retargeted to the plastid are undergoing evolutionary rate increases unrelated to the target peptide, suggesting that other functional domains are also evolving rapidly.

Retargeting of the cytosolic, homomeric ACCase protein to the plastid is somewhat unexpected given that a heteromeric ACCase complex already exists in plastids. Whether the retargeted homomeric

ACCases functionally replaces or coexists with the heteromeric version appears to vary across clades. In some angiosperm groups, the two complexes coexist, including in *Arabidopsis thaliana* and likely in other members of the Brassicaceae (Babiychuk et al., 2011; Rousseau-Gueutin et al., 2013). In other clades, the homomeric ACCase has replaced the heteromeric version, as was reported in the Poaceae (Konishi and Sasaki, 1994). The duplication found in *Silene noctiflora* and *Silene paradoxa* may also represent a replacement event given that both species lack at least one heteromeric ACCase gene each, where *S. noctiflora* lacks all of them (Rockenbach et al., 2016). In some cases, including *Monsonia emarginata* in the Geraniaceae, the plastid-encoded *accD* gene of the heteromeric complex has been transferred to the nuclear genome, again suggesting that the heteromeric version is still functional (Park et al., 2017; Rousseau-Gueutin et al., 2013). These contrasting histories of replacement vs. coexistence may mean that duplicates in different clades are evolving under different selection regimes.

Variation in post-duplication fates could confound tests of selection conducted across the entire *ACC* tree. Using PAML and HyPhy (Murrell et al., 2015; Smith et al., 2015; Wertheim et al., 2015; Yang, 2007), we tested for positive selection and relaxed selection in *ACC2* genes relative to *ACC1* genes, both of which can contribute to increased rates of protein sequence evolution. The results were mixed; there is some evidence for relaxed selection across all *ACC2* branches as well as for positive selection in a small number of branches and sites (**Table S4.1**). Across the four families in our sample, the smallest ratio between mean *ACC2* d_N and mean *ACC1* d_N was found in the Poaceae (1.5 vs. 2.2-2.6 for the other three families). Since the heteromeric ACCase is completely absent in the Poaceae (Konishi and Sasaki, 1994), we would expect stronger purifying selection on the plastid homomeric ACCase in this clade compared to clades in which the two versions coexist. Thus, these results are consistent with the hypothesis that relaxed selection is contributing to rate accelerations and that there is greater relaxation of selection when homomeric and heteromeric ACCases functions redundantly in the plastid, though the evidence is still limited. The potential for positive selection on retargeted ACCases is intriguing given that these proteins are thought to perform the same function as the ancestral protein; it is possible that retargeted proteins are adapting to specific biochemical and/or osmotic conditions within the new destination. Increased

evolutionary rates after subcellular retargeting have been previously noted, though we do not fully understand their underlying causes (Byun-McKay and Geeta, 2007; Marques et al., 2008).

Ongoing duplication of nuclear-encoded Clp core subunit genes is common in angiosperms

Across green plants, duplication of the plastid-encoded Clp core subunit gene *clpP1* has only been found in a handful of lineages (Williams et al., 2019). While other studies have identified recent duplications of nuclear-encoded Clp core subunit genes (Rockenbach et al., 2016; Williams et al., 2021), our work shows that duplications of these nuclear-encoded subunits are pervasive across angiosperms (**Figure 4.3**). Because we used a mix of transcriptomic and genomic data, we took into consideration the possibility of misidentifying transcript variants as paralogs but our use of primary transcripts only and manual curation to remove hits that appeared to be splice variants (see Materials and Methods) minimizes the risk of this type of error.

The prevalence of whole genome duplication in plants may partially explain the prevalence of Clp core subunit duplication (Clark and Donoghue, 2018; De Bodt et al., 2005; del Pozo and Ramirez-Parra, 2015; Flagel and Wendel, 2009; Panchy et al., 2016; Wendel et al., 2018). For instance, *Soja max* is a partially diploidized tetraploid, meaning that this lineage underwent a polyploidization event very recently and has only just started the subsequent process of genome reduction (Shultz et al., 2006). *Soja max* had the largest number of duplicated subunits across our sample, which is consistent with this history of whole genome duplication. Similarly, *Populus trichopoda*, which tied for the second largest number of duplicated subunits, only recently underwent genome reduction after whole genome duplication (Tuskan et al., 2006). In these cases, we may simply be observing the short-term effects of polyploidization prior to returning to a single copy of each of these genes.

Subunit stoichiometry and subfunctionalization in the evolution of the plastid Clp complex

Clp core subunit ratios have been studied in *Arabidopsis thaliana* (Olinares et al., 2011a). The core consists of two rings—a ClpP1/ClpR1-4 ring with a 3:1:1:1:1 subunit ratio, respectively, and a

ClpP3-6 ring with a 1:2:3:1 subunit ratio, respectively (Olinares et al., 2011a). Despite the high degree of structural similarity amongst the plastid Clp core subunits, core composition (i.e. the number of each type of core subunit) does not appear to vary in *A. thaliana* (Olinares et al., 2011a; Peltier et al., 2004). Due to the stability of subunit interactions in *A. thaliana*, Clp complexes in other angiosperms are typically assumed to have the same ratios of core subunits, but our results suggest that varied numbers of core subunit paralogs may lead to varied stoichiometry across species. Additional work has shown that loss of catalytic activity in ClpP5 (present in three copies in *A. thaliana*) is lethal while loss of catalytic activity in ClpP3 (present in one copy in *A. thaliana*) is tolerated, suggesting that core subunit composition may be flexible if there is simply a required or threshold number of catalytic subunits (Liao et al., 2018).

In fact, core subunit composition has been dynamic throughout the evolutionary history of the green lineage. The Clp complex is widely conserved across bacteria; in most bacteria, including *E. coli*, the Clp core consists of 14 identical subunits (Nishimura and van Wijk, 2015; Yu and Houry, 2007). However, in cyanobacteria, several gene duplications have produced four different core subunits—three catalytic ClpP subunits and one catalytically inactive ClpR subunit (Andersson et al., 2009; Stanne et al., 2007). In green lineage (Viridiplantae) plastids, which are descended from ancient cyanobacteria, gene duplication has continued to expand the number of core subunit genes, yielding nine different types of proteins incorporated into the Clp core (ClpP1,3-6, ClpR1-4) (Nishimura and van Wijk, 2015). Interestingly, ClpR subunits are incorporated into the core despite their lack of catalytic activity; they are thought to play a structural role in the complex, including chaperone docking onto the proteolytic core (Nishimura and van Wijk, 2015; Olinares et al., 2011b, 2011a; Sjögren and Clarke, 2011). In *A. thaliana*, the Clp chaperone is believed to bind only to the ClpP1/ClpR1-4 ring, whereas chaperone proteins bind to both rings of the Clp core in bacteria (Peltier et al., 2004; Yu and Houry, 2007). This ClpP/ClpR division of function (catalytic activity vs chaperone binding) is indicative of subfunctionalization. Further, though the plastid Clp core subunit genes share common ancestry and are structurally similar, knockouts of individual subunits tend to produce severe phenotypes, including lethality in several cases (Kim et al., 2009; Koussevitzky et al., 2007; Rudella et al., 2006).

Possible subfunctionalization in recent paralogs of Clp core subunits

Given that subfunctionalization has likely played a major role in plastid Clp complex evolution, we were particularly interested in whether we could identify subfunctionalization after more recent duplication events. Taken to an extreme, subfunctionalization would involve having one gene for each of the 14 core subunits, which would lead to further expansion of the typical nine core subunit genes. The total number of core subunits after including recent paralogs and the plastid-encoded ClpP1 was less than 14 in most species. *Musa acuminata*, *Plantago maritima*, and *Populus trichocarpa* had 14 each, *Soja max* had 17, and *Gossypium raimondii* had 15 (**Figure 4.3)**. The numbers larger than 14 were driven in both cases by multiple paralogs of *CLPP5*, with five and seven copies, respectively. ClpP5 has the largest number of subunits stoichiometrically among the eight nuclear-encoded subunits, so the fact that the two largest numbers of paralogs were both found in *CLPP5* could potentially suggest that some species are moving toward a 1:1 relationship between genes and core subunits. However, this explanation is not supported by other evidence. For example, the other cases of >2 paralogs were found for *CLPR4*, which encodes a protein that is present in just a single copy in the core in *Arabidopsis*. Further, it is not clear that all of these paralogs are capable of producing functional proteins given truncations and loss of catalytic sites (**Table S4.2, Table S4.3**).

We tested for signatures of subfunctionalization by looking at evolutionary rate asymmetry. Under subfunctionalization, we would expect paralogs to evolve at symmetric rates relative to one another but asymmetrically relative to their common ancestor (Pegueroles et al., 2013). We found five of these cases in our dataset: *Plantago maritima CLPP3*, *Geranium maderense CLPP4, Medicago truncatula CLPP5, Populus trichocarpa CLPP5,* and *Soja max CLPR1* (**Table 4.2**). In cases of subfunctionalization, we would expect the paralogs to evolve more quickly than the common ancestor because of relaxed selection due to their more limited functional roles, which was only the case for the former three. In those three cases, the evidence is consistent with subfunctionalization, particularly given that all six involved paralogs are full length. Further, the *P. maritima CLPP3* paralogs share the same

substitutions in all three catalytic sites, which indicates duplication after the loss of catalytic activity, and the *G. maderense CLPP4* and *M. truncatula CLPP5* paralogs all have fully retained catalytic triads (**Table S4.2**).

Possible pseudogenization or neofunctionalization in recent paralogs of Clp core subunits

Predictions about evolutionary rates under neofunctionalization are similar to predictions under the degeneration/gene loss model—one paralog will maintain the ancestral evolutionary rate while the other undergoes evolutionary rate acceleration (Hahn, 2009; Pegueroles et al., 2013; Zhang, 2003). Previous work in the Clp complex has shown that even ClpP1 subunits demonstrating massive accelerations in evolutionary rate can still be functional, meaning that high evolutionary rates alone do not necessarily indicate pseudogenization (Barnard-Kubow et al., 2014; Williams et al., 2019, 2015). Other sequence features can help us differentiate between pseudogenization and neofunctionalization. For example, truncation of a sequence can be evidence that it is no longer producing a functional protein; additionally, for ClpP subunits, loss of catalytic sites may also be an indication of degeneration/pseudogenization (though there may be exceptions, including the *P. maritima CLPP3* paralogs mentioned above). In our dataset, the paralogs of *Musa acuminata CLPP4* and *P. trichocarpa CLPP4* follow these patterns (**Table 4.2**). In each of these pairs, the paralogs are evolving asymmetrically, and the paralog with a faster rate of evolution is truncated and lacking all three catalytic sites, suggesting loss of function (**Table S4.2**). Another example of probable pseudogenization is found for the second copy of *M. acuminata CLPR1*. This paralog was annotated as two separate genes due to an internal stop codon, which would lead to a truncation in the resultant protein.

As for neofunctionalization, there are other cases in our dataset where *CLPP* paralogs evolving asymmetrically both encode proteins that are full length with retained catalytic sites (for instance, *Geranium maderense CLPP5* and *Oenothera biennis CLPP5*). There are no known instances of retargeting of plastid Clp core subunits; thus, evolutionary drivers of neofunctionalization of duplicated subunits are unknown. It is possible that neofunctionalization in this complex could involve recruiting

additional interacting proteins—the ClpT proteins, for instance, are involved in assembly of the core and are a recent evolutionary innovation specific to green plants (Colombo et al., 2014; Kim et al., 2015; Nishimura and van Wijk, 2015; Sjögren and Clarke, 2011). Additionally, ongoing work has identified potential new adapter proteins in the plastid Clp complex (Montandon et al., 2019; Nishimura et al., 2015). Another possibility is tissue-specific expression of paralogs, which has not been documented in the Clp complex but has been identified in mitochondrial complexes (Boss et al., 1997; Guerrero-Castillo et al., 2017; Sinkler et al., 2017).

Possible retention of Clp core paralogs under the gene dosage advantage hypothesis

We also identified eight cases of symmetrically evolving paralogs that are also evolving symmetrically relative to the common ancestor (**Table 4.2**). Under our initial predictions, these would represent paralogs retained under the gene dosage advantage hypothesis (Hahn, 2009; Ohno, 1970; Pegueroles et al., 2013; Zhang, 2003). Of these eight paralog pairs, four are from *Soja max* (which had six total pairs of paralogs), and three are from *Populus trichocarpa* (which had five total pairs of paralogs). As described above, both of these species are in the process of rediploidization after a recent whole genome duplication (Shultz et al., 2006; Tuskan et al., 2006). It is possible that these results reflect the fact that the gene duplications happened so recently that the paralogs have not had time to diverge. This possibility is further supported given that the estimates of numbers of substitutions for many of these paralogs were so low that there was virtually no power to detect significant asymmetry.

Alternative hypotheses and future directions

While we based our analyses on established expectations for evolutionary rates under different post-duplication fates (gene dosage advantage, neofunctionalization, and subfunctionalization), other work has challenged the universality of these predictions. He and Zhang (2005) outline the subneofunctionalization model, in which gene duplicates undergo rapid subfunctionalization followed by prolonged neofunctionalization. Asymmetric evolutionary rates are often assumed to be the result of

either neofunctionalization or degeneration, but subfunctionalization can also occur in an asymmetric fashion (He and Zhang, 2005). This hypothesis could relate to some of our results; cases of asymmetric evolutionary rates could be due to subfunctionalization rather than neofunctionalization. Additionally, functional constraint can also exist under neofunctionalization, leading to lower substitution rates and possibly symmetric rates of evolution, meaning that symmetrically evolving paralogs could represent cases of neofunctionalization rather than subfunctionalization or gene dosage advantage (He and Zhang, 2005).

Regardless, our results demonstrate that post-duplication evolutionary fates of paralogs vary widely across clades, even when the same genes are involved. Duplications of the homomeric ACCase complex gene (*ACC*) and subsequent retargeting of one protein to the plastid have been previously reported (Babiychuk et al., 2011; Konishi and Sasaki, 1994; Park et al., 2017; Parker et al., 2014; Rockenbach et al., 2016; Schulte et al., 1997). Our results show that the retargeted duplicates almost universally have increased d_N/d_S rates (**Figure 4.2, Figure S4.1**). As for plastid Clp core subunit duplications, duplication has clearly shaped this complex over the course of Viridiplantae evolution. We provide evidence of all possible post-duplication routes of recent paralogs amongst the different subunits and different species in our dataset. Overall, our results provide compelling evidence that subunit ratios and stoichiometry may be dynamic across angiosperm lineages. Isolation of plastid Clp complexes and analyses of subunit composition have been performed in a handful of species (Moreno et al., 2017; Olinares et al., 2011a; Williams et al., 2019); future work could determine these compositions in other angiosperms, including those that have undergone recent gene duplications. Our work demonstrates that gene duplication has been and continues to be an important force in plastid evolution.

Table 4.1: Differences in evolutionary rates between duplicated and non-duplicated plastid Clp core subunits. Reported *p*-values are based on likelihood ratio tests for 2-partition vs. 3-partition PAML models (see Materials and Methods). Log-likelihood (lnL) values are reported for each model.

Subunit	lnL 2-partition model	lnL 3-partition model	*p*-value	Class with higher d_N/d_S
CLPP3	-10058.01	-10047.93	7.12e-06	Duplicated
CLPP4	-9606.41	-9552.26	<1.00e-10	Duplicated
CLPP5	-8121.80	-8070.90	<1.00e-10	Duplicated
CLPP6	-7697.62	-7691.31	3.82e-04	Non-duplicated
CLPR1	-14534.75	-14517.85	6.07e-09	Duplicated
CLPR2	-12018.69	-12002.74	1.63e-08	Duplicated
CLPR3	-10556.37	-10556.05	0.42	n/a
CLPR4	-10395.60	-10337.35	<1.00e-10	Duplicated

Table 4.2: *p*-values for asymmetries between paralogs and between paralogs and their common ancestor.

Species	Gene	Paralog 1 vs. paralog 2	Paralog 1 vs. ancestor	Paralog 2 vs. ancestor	Paralogs 1+2 vs. ancestor
P. maritima	*CLPP3*	0.83			1.25e-04*
S. max	*CLPP3*	1			0.20
P. maritima	*CLPP4*	1.28e-07	1.07e-09*	0.71	
M. acuminata	*CLPP4*	0.04	0.15	2.12e-04*	
S. max	*CLPP4*	0.57			1
P. trichocarpa	*CLPP4*	3.31e-04	3.44e-12*	1.76e-13*	
G. maderense	*CLPP4*	0.08			1.72e-03*
M. truncatula	*CLPP5*	1			3.61e-05*
P. trichocarpa	*CLPP5*	1			0.04^
O. biennis	*CLPP5*	9.12e-3	0.10	5.37e-4*	
G. maderense	*CLPP5*	1.71e-04	0.42	6.23e-4*	
S. max	*CLPP6*	0.27			1
P. trichocarpa	*CLPP6*	0.46			1
G. raimondii	*CLPP6*	0.71			1
M. acuminata	*CLPR1*	2.20e-16	0.82	2.20e-16*	
S. max	*CLPR1*	0.47			0.05^
P. trichocarpa	*CLPR1*	0.77			0.42
V. vinifera	*CLPR1*	4.45e-3	0.17	1.38e-09*	
M. guttatus	*CLPR1*	0.05	0.59	4.70e-04*	
P. maritima	*CLPR1*	4.70e-05	0.53	2.54e-05*	
P. maritima	*CLPR2*	0.01	0.18	1.63e-06*	
S. max	*CLPR3*	1			0.55
P. trichocarpa	*CLPR3*	0.34			0.35
M. truncatula	*CLPR4*	2.29e-03	0.21	0.11	
E. grandis	*CLPR4*	2.51e-05	0.68	5.42e-11*	
P. maritima	*CLPR4*	9.73e-05	9.53e-04	4.17e-14*	

* denotes that paralog(s) has/have significantly higher evolutionary rate than ancestor branch
^ denotes that paralog(s) has/have significantly lower evolutionary rate than ancestor branch

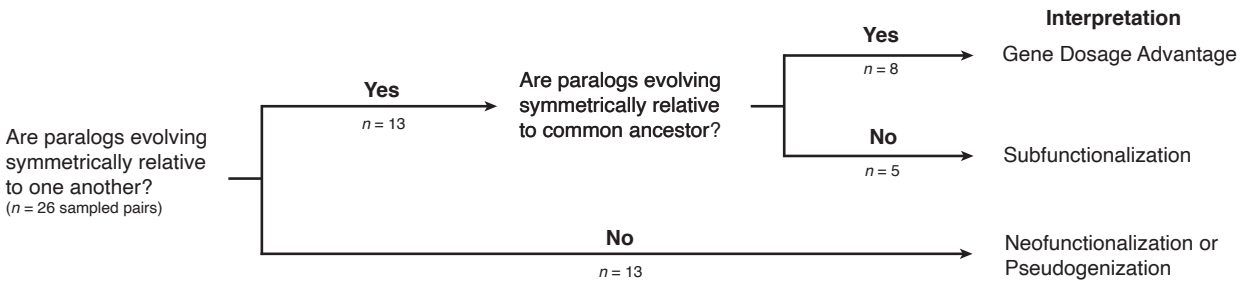


Figure 4.1: Expectations under different post-duplication models. For each pair of paralogs (n = 26), we first determined whether they were evolving symmetrically relative to one another using N and S estimates from PAML output. Paralogs evolving asymmetrically are predicted to represent neofunctionalization or pseudogenization events. For paralogs evolving symmetrically (n = 8), combined N and S values were compared to those of the immediate ancestor branch. Pairs evolving symmetrically relative to the common ancestor (n = 8) are predicted to represent gene dosage advantage while those evolving asymmetrically relative to the common ancestor (n = 5) are predicted to represent subfunctionalization.

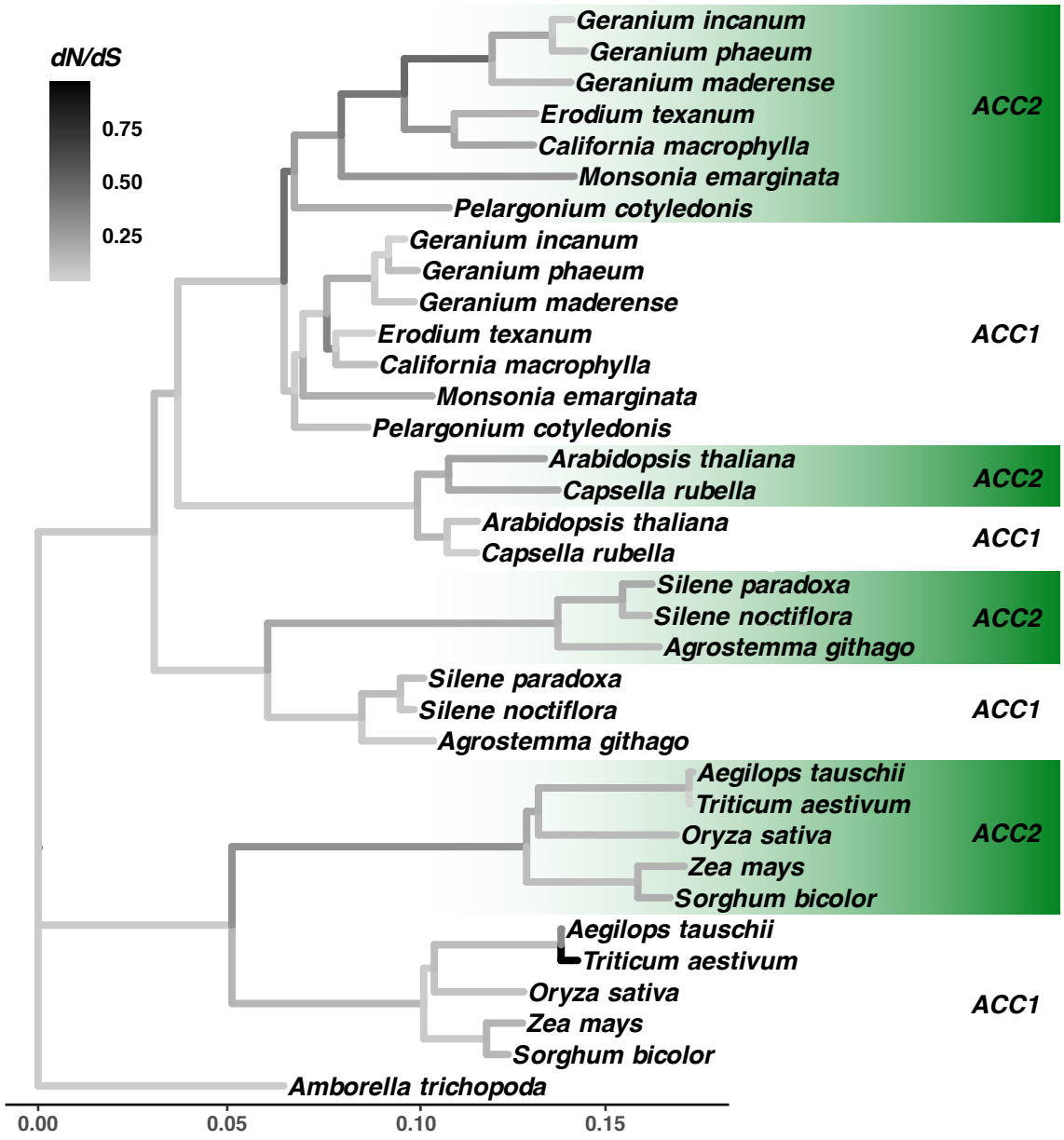


Figure 4.2: *ACC* genes across the Brassicaceae, Caryophyllaceae, Geraniaceae, and Poaceae, with the single copy of *ACC* in *Amborella trichopoda* as an outgroup. Branch lengths represent d_N values and branch colors represent d_N/d_S ratios.

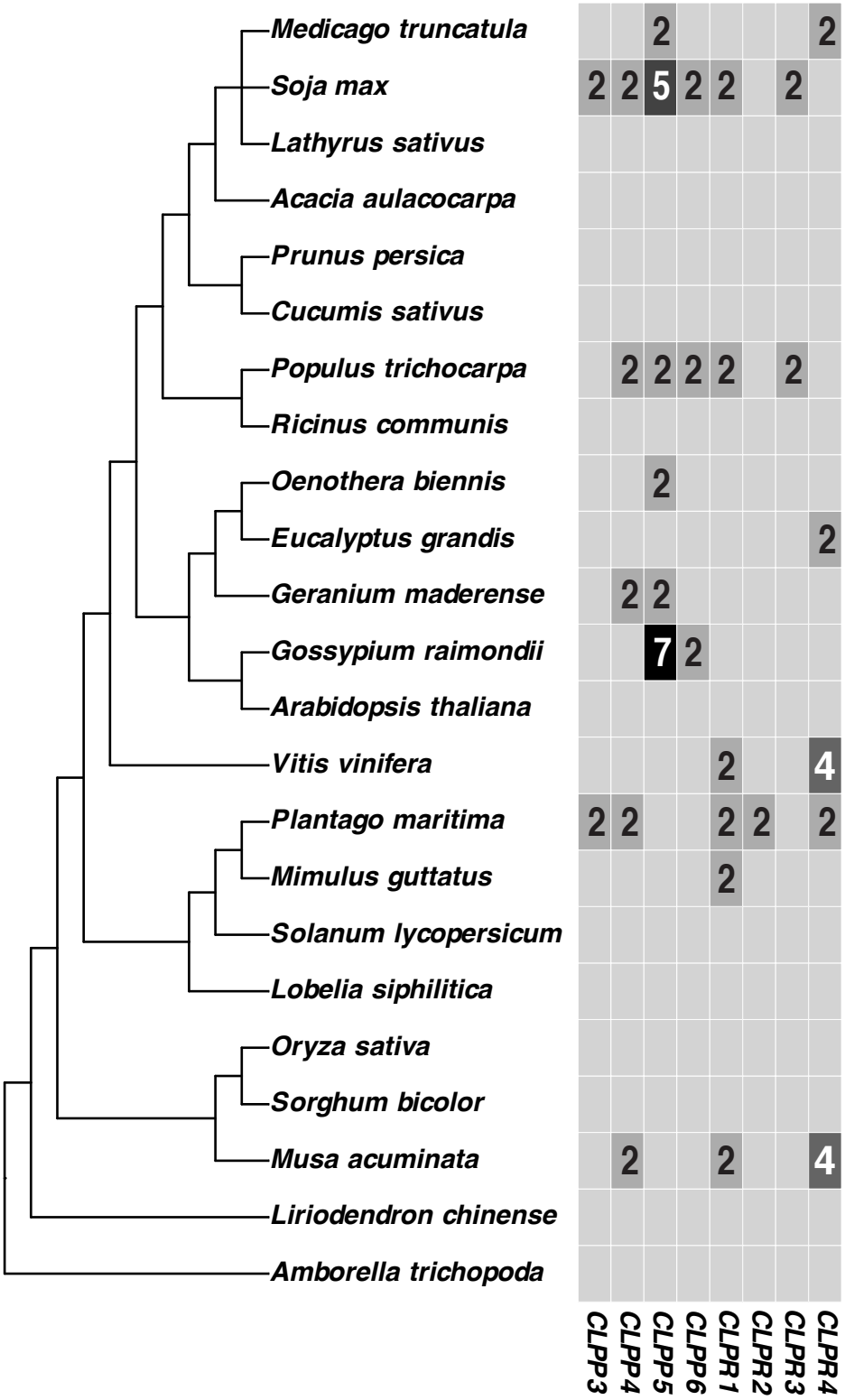


Figure 4.3: Copy numbers of the nuclear-encoded subunits of the plastid Clp core across angiosperms. Boxes without numbers indicate single-copy genes.

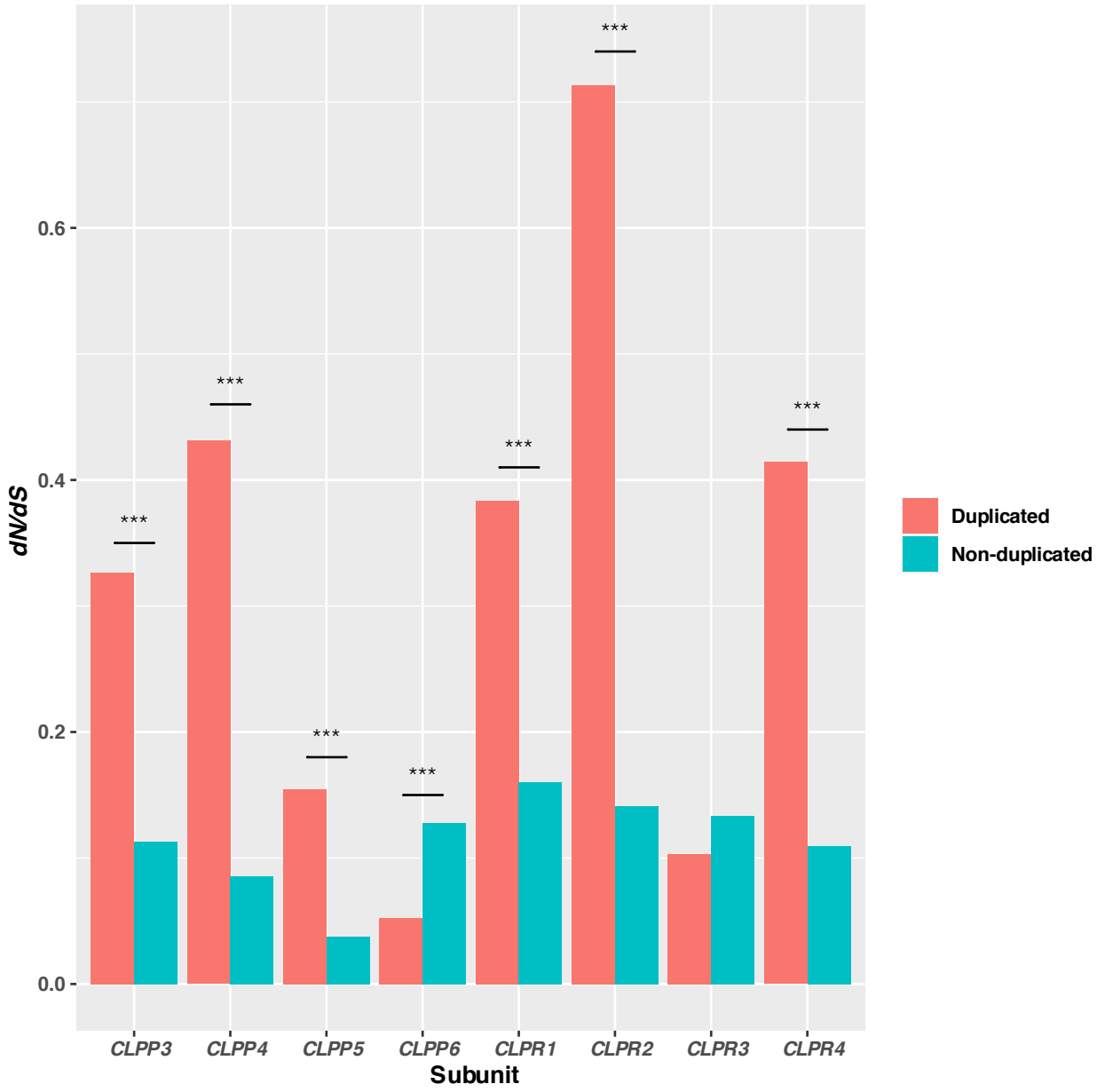


Figure 4.4: d_N/d_S ratios of duplicated and non-duplicated plastid Clp core subunits across angiosperms. The values were calculated using a PAML branch test with three groups, where each group was assigned its own d_N/d_S value: non-duplicated terminal branches, duplicated terminal branches (and in the cases of *CLPP5* and *CLPR4*, internal post-duplication branches), and internal branches. Significant differences ($p<<0.001$) are indicated with ***.

LITERATURE CITED

Andersson, F.I., Tryggvesson, A., Sharon, M., Diemand, A.V., Classen, M., Best, C., Schmidt, R., Schelin, J., Stanne, T.M., Bukau, B., Robinson, C.V., Witt, S., Mogk, A., Clarke, A.K., 2009. Structure and function of a novel type of ATP-dependent Clp protease. J. Biol. Chem. 284, 13519–13532. https://doi.org/10.1074/jbc.M809588200

Apitz, J., Nishimura, K., Schmied, J., Wolf, A., Hedtke, B., Wijk, K.J. van, Grimm, B., 2016. Posttranslational Control of ALA Synthesis Includes GluTR Degradation by Clp Protease and Stabilization by GluTR-Binding Protein. Plant Physiology 170, 2040–2051. https://doi.org/10.1104/pp.15.01945

Babiychuk, E., Vandepoele, K., Wissing, J., Garcia-Diaz, M., Rycke, R.D., Akbari, H., Joubès, J., Beeckman, T., Jänsch, L., Frentzen, M., Montagu, M.C.E.V., Kushnir, S., 2011. Plastid gene expression and plant development require a plastidic protein of the mitochondrial transcription termination factor family. PNAS 108, 6674–6679. https://doi.org/10.1073/pnas.1103442108

Barnard-Kubow, K.B., Sloan, D.B., Galloway, L.F., 2014. Correlation between sequence divergence and polymorphism reveals similar evolutionary mechanisms acting across multiple timescales in a rapidly evolving plastid genome. BMC Evol Biol 14. https://doi.org/10.1186/s12862-014-0268-y

Boss, O., Samec, S., Paoloni-Giacobino, A., Rossier, C., Dulloo, A., Seydoux, J., Muzzin, P., Giacobino, J.P., 1997. Uncoupling protein-3: a new member of the mitochondrial carrier family with tissue-specific expression. FEBS Lett 408, 39–42. https://doi.org/10.1016/s0014-5793(97)00384-0

Bouchnak, I., van Wijk, K.J., 2021. Structure, Function and Substrates of Clp AAA+ protease systems in cyanobacteria, plastids and apicoplasts; a comparative analysis. Journal of Biological Chemistry 100338. https://doi.org/10.1016/j.jbc.2021.100338

Bruce, B.D., 2001. The paradox of plastid transit peptides: conservation of function despite divergence in primary structure. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research 1541, 2–21. https://doi.org/10.1016/S0167-4889(01)00149-5

Bruce, B.D., 2000. Chloroplast transit peptides: structure, function and evolution. Trends in Cell Biology 10, 440–447. https://doi.org/10.1016/S0962-8924(00)01833-X

Byun-McKay, S.A., Geeta, R., 2007. Protein subcellular relocalization: a new perspective on the origin of novel genes. Trends in Ecology & Evolution 22, 338–344. https://doi.org/10.1016/j.tree.2007.05.002

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421. https://doi.org/10.1186/1471-2105-10-421

Caroca, R., Howell, K.A., Malinova, I., Burgos, A., Tiller, N., Pellizzer, T., Annunziata, M.G., Hasse, C., Ruf, S., Karcher, D., Bock, R., 2021. Knock-down of the plastid-encoded acetyl-CoA carboxylase gene uncovers functions in metabolism and development. Plant Physiology. https://doi.org/10.1093/plphys/kiaa106

Castresana, J., 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Molecular Biology and Evolution 17, 540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334

Clark, J.W., Donoghue, P.C.J., 2018. Whole-Genome Duplication and Plant Macroevolution. Trends in Plant Science 23, 933–945. https://doi.org/10.1016/j.tplants.2018.07.006

Colombo, C.V., Ceccarelli, E.A., Rosano, G.L., 2014. Characterization of the accessory protein ClpT1 from Arabidopsis thaliana: oligomerization status and interaction with Hsp100 chaperones. BMC Plant Biology 14, 228. https://doi.org/10.1186/s12870-014-0228-0

Conant, G.C., Wagner, A., 2003. Asymmetric Sequence Divergence of Duplicate Genes. Genome Res. 13, 2052–2058. https://doi.org/10.1101/gr.1252603

Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27, 1164–1165. https://doi.org/10.1093/bioinformatics/btr088

De Bodt, S., Maere, S., Van de Peer, Y., 2005. Genome duplication and the origin of angiosperms. Trends in Ecology & Evolution 20, 591–597. https://doi.org/10.1016/j.tree.2005.07.008

del Pozo, J.C., Ramirez-Parra, E., 2015. Whole genome duplications in plants: an overview from Arabidopsis. Journal of Experimental Botany 66, 6991–7003. https://doi.org/10.1093/jxb/erv432

Dermitzakis, E.T., Clark, A.G., 2001. Differential selection after duplication in mammalian developmental genes. Mol Biol Evol 18, 557–562. https://doi.org/10.1093/oxfordjournals.molbev.a003835

Des Marais, D.L., Rausher, M.D., 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature 454, 762–765. https://doi.org/10.1038/nature07092

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. https://doi.org/10.1371/journal.pone.0001386

Flagel, L.E., Wendel, J.F., 2009. Gene duplication and evolutionary novelty in plants. New Phytologist 183, 557–564. https://doi.org/10.1111/j.1469-8137.2009.02923.x

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y., Postlethwait, J., 1999. Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetics 151, 1531–1545.

Freeling, M., Lyons, E., Pedersen, B., Alam, M., Ming, R., Lisch, D., 2008. Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. Genome Res 18, 1924–1937. https://doi.org/10.1101/gr.081026.108

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., Rokhsar, D.S., 2012. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Research 40, D1178–D1186. https://doi.org/10.1093/nar/gkr944

Green, B.R., 2011. Chloroplast genomes of photosynthetic eukaryotes. The Plant Journal 66, 34–44. https://doi.org/10.1111/j.1365-313X.2011.04541.x

Guerrero-Castillo, S., Cabrera-Orefice, A., Huynen, M.A., Arnold, S., 2017. Identification and evolutionary analysis of tissue-specific isoforms of mitochondrial complex I subunit NDUFV3. Biochim Biophys Acta Bioenerg 1858, 208–217. https://doi.org/10.1016/j.bbabio.2016.12.004

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol 59, 307–321. https://doi.org/10.1093/sysbio/syq010

Hahn, M.W., 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. J Hered 100, 605–617. https://doi.org/10.1093/jhered/esp047

Hasegawa, M., Cao, Y., Yang, Z., 1998. Preponderance of slightly deleterious polymorphism in mitochondrial DNA: nonsynonymous/synonymous rate ratio is much higher within species than between species. Mol Biol Evol 15, 1499–1505. https://doi.org/10.1093/oxfordjournals.molbev.a025877

He, X., Zhang, J., 2005. Rapid Subfunctionalization Accompanied by Prolonged and Substantial Neofunctionalization in Duplicate Gene Evolution. Genetics 169, 1157–1164. https://doi.org/10.1534/genetics.104.037051

Ho, S.Y.W., Phillips, M.J., Cooper, A., Drummond, A.J., 2005. Time Dependency of Molecular Rate Estimates and Systematic Overestimation of Recent Divergence Times. Molecular Biology and Evolution 22, 1561–1568. https://doi.org/10.1093/molbev/msi145

Huang, Y., Kendall, T., Forsythe, E.S., Dorantes-Acosta, A., Li, S., Caballero-Pérez, J., Chen, X., Arteaga-Vázquez, M., Beilstein, M.A., Mosher, R.A., 2015. Ancient Origin and Recent Innovations of RNA Polymerase IV and V. Mol Biol Evol 32, 1788–1799. https://doi.org/10.1093/molbev/msv060

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., dePamphilis, C.W., Leebens-Mack, J., Müller, K.F., Guisinger-Bellian, M., Haberle, R.C., Hansen, A.K., Chumley, T.W., Lee, S.-B., Peery, R., McNeal, J.R., Kuehl, J.V., Boore, J.L., 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. PNAS 104, 19369–19374. https://doi.org/10.1073/pnas.0709121104

Jarvis, P., 2008. Targeting of nucleus-encoded proteins to chloroplasts in plants. New Phytologist 179, 257–285. https://doi.org/10.1111/j.1469-8137.2008.02452.x

Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30, 772–780. https://doi.org/10.1093/molbev/mst010

Kim, J., Kimber, M.S., Nishimura, K., Friso, G., Schultz, L., Ponnala, L., Wijk, K.J. van, 2015. Structures, Functions, and Interactions of ClpT1 and ClpT2 in the Clp Protease System of Arabidopsis Chloroplasts. The Plant Cell 27, 1477–1496. https://doi.org/10.1105/tpc.15.00106

Kim, J., Rudella, A., Rodriguez, V.R., Zybailov, B., Olinares, P.D.B., Wijk, K.J. van, 2009. Subunits of the Plastid ClpPR Protease Complex Have Differential Contributions to Embryogenesis, Plastid Biogenesis, and Plant Development in Arabidopsis. The Plant Cell 21, 1669–1692. https://doi.org/10.1105/tpc.108.063784

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I., Koonin, E.V., 2002. Selection in the evolution of gene duplications. Genome Biol 3, RESEARCH0008. https://doi.org/10.1186/gb-2002-3-2-research0008

Konishi, T., Sasaki, Y., 1994. Compartmentalization of two forms of acetyl-CoA carboxylase in plants and the origin of their tolerance toward herbicides. PNAS 91, 3598–3601. https://doi.org/10.1073/pnas.91.9.3598

Konishi, T., Shinohara, K., Yamada, K., Sasaki, Y., 1996. Acetyl-CoA Carboxylase in Higher Plants: Most Plants Other Than Gramineae Have Both the Prokaryotic and the Eukaryotic Forms of This Enzyme. Plant and Cell Physiology 37, 117–122. https://doi.org/10.1093/oxfordjournals.pcp.a028920

Kosakovsky Pond, S.L., Poon, A.F.Y., Velazquez, R., Weaver, S., Hepler, N.L., Murrell, B., Shank, S.D., Magalis, B.R., Bouvier, D., Nekrutenko, A., Wisotsky, S., Spielman, S.J., Frost, S.D.W., Muse, S.V., 2020. HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. Molecular Biology and Evolution 37, 295–299. https://doi.org/10.1093/molbev/msz197

Koussevitzky, S., Stanne, T.M., Peto, C.A., Giap, T., Sjögren, L.L.E., Zhao, Y., Clarke, A.K., Chory, J., 2007. An <Emphasis Type="Italic">Arabidopsis thaliana</Emphasis> virescent mutant reveals a role for ClpR1 in plastid development. Plant Mol Biol 63, 85–96. https://doi.org/10.1007/s11103-006-9074-2

Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A., Graham, S.W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., Porsch, M., Quint, M., Rensing, S.A., Soltis, D.E., Soltis, P.S., Stevenson, D.W., Ullrich, K.K., Wickett, N.J., DeGironimo, L., Edger, P.P., Jordon-Thaden, I.E., Joya, S., Liu, T., Melkonian, B., Miles, N.W., Pokorny, L., Quigley, C., Thomas, P., Villarreal, J.C., Augustin, M.M., Barrett, M.D., Baucom, R.S., Beerling, D.J., Benstein, R.M., Biffin, E., Brockington, S.F., Burge, D.O., Burris, J.N., Burris, K.P., Burtet-Sarramegna, V., Caicedo, A.L., Cannon, S.B., Çebi, Z., Chang, Y., Chater, C., Cheeseman, J.M., Chen, T., Clarke, N.D., Clayton, H., Covshoff, S., Crandall-Stotler, B.J., Cross, H., dePamphilis, C.W., Der, J.P., Determann, R., Dickson, R.C., Di Stilio, V.S., Ellis, S., Fast, E., Feja, N., Field, K.J., Filatov, D.A., Finnegan, P.M., Floyd, S.K., Fogliani, B., García, N., Gâteblé, G., Godden, G.T., Goh, F. (Qi Y., Greiner, S., Harkess, A., Heaney, J.M., Helliwell, K.E., Heyduk, K., Hibberd, J.M., Hodel, R.G.J., Hollingsworth, P.M., Johnson, M.T.J., Jost, R., Joyce, B., Kapralov, M.V., Kazamia, E., Kellogg, E.A., Koch, M.A., Von Konrat, M., Könyves, K., Kutchan, T.M., Lam, V., Larsson, A., Leitch, A.R., Lentz, R., Li, F.-W., Lowe, A.J., Ludwig, M., Manos, P.S., Mavrodiev, E., McCormick, M.K., McKain, M., McLellan, T., McNeal, J.R., Miller, R.E., Nelson, M.N., Peng, Y., Ralph, P., Real, D., Riggins, C.W., Ruhsam, M., Sage, R.F., Sakai, A.K., Scascitella, M., Schilling, E.E., Schlösser, E.-M., Sederoff, H., Servick, S., Sessa, E.B., Shaw, A.J., Shaw, S.W., Sigel, E.M., Skema, C., Smith, A.G., Smithson, A., Stewart, C.N., Stinchcombe, J.R., Szövényi, P., Tate, J.A., Tiebel, H., Trapnell, D., Villegente, M., Wang, C.-N., Weller, S.G., Wenzel, M., Weststrand, S., Westwood, J.H., Whigham, D.F., Wu, S., Wulff, A.S., Yang, Y., Zhu, D., Zhuang, C., Zuidof, J., Chase, M.W., Pires, J.C., Rothfels, C.J., Yu, J., Chen, C., Chen, L., Cheng, S., Li, J., Li, R., Li, X., Lu, H., Ou, Y., Sun, X., Tan, X., Tang, J., Tian, Z., Wang, F.,

Wang, J., Wei, X., Xu, X., Yan, Z., Yang, F., Zhong, X., Zhou, F., Zhu, Y., Zhang, Y., Ayyampalayam, S., Barkman, T.J., Nguyen, N., Matasci, N., Nelson, D.R., Sayyari, E., Wafula, E.K., Walls, R.L., Warnow, T., An, H., Arrigo, N., Baniaga, A.E., Galuska, S., Jorgensen, S.A., Kidder, T.I., Kong, H., Lu-Irving, P., Marx, H.E., Qi, X., Reardon, C.R., Sutherland, B.L., Tiley, G.P., Welles, S.R., Yu, R., Zhan, S., Gramzow, L., Theißen, G., Wong, G.K.-S., One Thousand Plant Transcriptomes Initiative, 2019. One thousand plant transcriptomes and the phylogenomics of green plants. Nature 574, 679–685. https://doi.org/10.1038/s41586-019-1693-2

Liao, J.-Y.R., Friso, G., Kim, J., Wijk, K.J. van, 2018. Consequences of the loss of catalytic triads in chloroplast CLPPR protease core complexes in vivo. Plant Direct 2, e00086. https://doi.org/10.1002/pld3.86

Lynch, M., Conery, J.S., 2000. The Evolutionary Fate and Consequences of Duplicate Genes. Science 290, 1151–1155. https://doi.org/10.1126/science.290.5494.1151

Lynch, M., Force, A., 2000. The Probability of Duplicate Gene Preservation by Subfunctionalization. Genetics 154, 459–473.

Maere, S., Bodt, S.D., Raes, J., Casneuf, T., Montagu, M.V., Kuiper, M., Peer, Y.V. de, 2005. Modeling gene and genome duplications in eukaryotes. PNAS 102, 5454–5459. https://doi.org/10.1073/pnas.0501102102

Majeran, W., Wollman, F.-A., Vallon, O., 2000. Evidence for a Role of ClpP in the Degradation of the Chloroplast Cytochrome b6f Complex. The Plant Cell 12, 137–149. https://doi.org/10.1105/tpc.12.1.137

Marques, A.C., Vinckenbosch, N., Brawand, D., Kaessmann, H., 2008. Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. Genome Biol 9, R54. https://doi.org/10.1186/gb-2008-9-3-r54

Moilanen, J.S., Majamaa, K., 2003. Phylogenetic network and physicochemical properties of nonsynonymous mutations in the protein-coding genes of human mitochondrial DNA. Mol Biol Evol 20, 1195–1210. https://doi.org/10.1093/molbev/msg121

Montandon, C., Friso, G., Liao, J.-Y.R., Choi, J., van Wijk, K.J., 2019. In Vivo Trapping of Proteins Interacting with the Chloroplast CLPC1 Chaperone: Potential Substrates and Adaptors. J. Proteome Res. 18, 2585–2600. https://doi.org/10.1021/acs.jproteome.9b00112

Moreno, J.C., Tiller, N., Diez, M., Karcher, D., Tillich, M., Schöttler, M.A., Bock, R., 2017. Generation and characterization of a collection of knock-down lines for the chloroplast Clp protease complex in tobacco. J Exp Bot 68, 2199–2218. https://doi.org/10.1093/jxb/erx066

Murrell, B., Weaver, S., Smith, M.D., Wertheim, J.O., Murrell, S., Aylward, A., Eren, K., Pollner, T., Martin, D.P., Smith, D.M., Scheffler, K., Kosakovsky Pond, S.L., 2015. Gene-Wide Identification of Episodic Selection. Molecular Biology and Evolution 32, 1365–1371. https://doi.org/10.1093/molbev/msv035

Nielsen, R., Weinreich, D.M., 1999. The age of nonsynonymous and synonymous mutations in animal mtDNA and implications for the mildly deleterious theory. Genetics 153, 497–506.

Nishimura, K., Apitz, J., Friso, G., Kim, J., Ponnala, L., Grimm, B., van Wijk, K.J., 2015. Discovery of a Unique Clp Component, ClpF, in Chloroplasts: A Proposed Binary ClpF-ClpS1 Adaptor Complex Functions in Substrate Recognition and Delivery. Plant Cell 27, 2677–2691. https://doi.org/10.1105/tpc.15.00574

Nishimura, K., Kato, Y., Sakamoto, W., 2017. Essentials of Proteolytic Machineries in Chloroplasts. Molecular Plant 10, 4–19. https://doi.org/10.1016/j.molp.2016.08.005

Nishimura, K., van Wijk, K.J., 2015. Organization, function and substrates of the essential Clp protease system in plastids. Biochimica et Biophysica Acta (BBA) - Bioenergetics, SI: Chloroplast Biogenesis 1847, 915–930. https://doi.org/10.1016/j.bbabio.2014.11.012

Ohno, S., 1970. Evolution by Gene Duplication. Springer Science & Business Media.

Olinares, P.D.B., Kim, J., Davis, J.I., van Wijk, K.J., 2011a. Subunit stoichiometry, evolution, and functional implications of an asymmetric plant plastid ClpP/R protease complex in Arabidopsis. Plant Cell 23, 2348–2361. https://doi.org/10.1105/tpc.111.086454

Olinares, P.D.B., Kim, J., van Wijk, K.J., 2011b. The Clp protease system; a central component of the chloroplast protease network. Biochimica et Biophysica Acta (BBA) - Bioenergetics, Regulation of Electron Transport in Chloroplasts 1807, 999–1011. https://doi.org/10.1016/j.bbabio.2010.12.003

Panchin, A.Y., Gelfand, M.S., Ramensky, V.E., Artamonova, I.I., 2010. Asymmetric and non-uniform evolution of recently duplicated human genes. Biol Direct 5, 54. https://doi.org/10.1186/1745-6150-5-54

Panchy, N., Lehti-Shiu, M., Shiu, S.-H., 2016. Evolution of Gene Duplication in Plants. Plant Physiology 171, 2294–2316. https://doi.org/10.1104/pp.16.00523

Park, S., Ruhlman, T.A., Weng, M.-L., Hajrah, N.H., Sabir, J.S.M., Jansen, R.K., 2017. Contrasting Patterns of Nucleotide Substitution Rates Provide Insight into Dynamic Evolution of Plastid and Mitochondrial Genomes of Geranium. Genome Biol Evol 9, 1766–1780. https://doi.org/10.1093/gbe/evx124

Parker, N., Wang, Y., Meinke, D., 2014. Natural Variation in Sensitivity to a Loss of Chloroplast Translation in Arabidopsis. Plant Physiology 166, 2013–2027. https://doi.org/10.1104/pp.114.249052

Paterson, A.H., Chapman, B.A., Kissinger, J.C., Bowers, J.E., Feltus, F.A., Estill, J.C., 2006. Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon. Trends Genet 22, 597–602. https://doi.org/10.1016/j.tig.2006.09.003

Pegueroles, C., Laurie, S., Albà, M.M., 2013. Accelerated Evolution after Gene Duplication: A Time-Dependent Process Affecting Just One Copy. Molecular Biology and Evolution 30, 1830–1842. https://doi.org/10.1093/molbev/mst083

Peltier, J.-B., Ripoll, D.R., Friso, G., Rudella, A., Cai, Y., Ytterberg, J., Giacomelli, L., Pillardy, J., Wijk, K.J. van, 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. https://doi.org/10.1074/jbc.M309212200

Porankiewicz, J., Wang, J., Clarke, A.K., 1999. New insights into the ATP-dependent Clp protease: Escherichia coli and beyond. Molecular Microbiology 32, 449–458. https://doi.org/10.1046/j.1365-2958.1999.01357.x

Renny-Byfield, S., Wendel, J.F., 2014. Doubling down on genomes: polyploidy and crop plants. Am J Bot 101, 1711–1725. https://doi.org/10.3732/ajb.1400119

Rizzon, C., Ponger, L., Gaut, B.S., 2006. Striking Similarities in the Genomic Distribution of Tandemly Arrayed Genes in Arabidopsis and Rice. PLOS Computational Biology 2, e115. https://doi.org/10.1371/journal.pcbi.0020115

Rockenbach, K., Havird, J.C., Monroe, J.G., Triant, D.A., Taylor, D.R., Sloan, D.B., 2016. Positive Selection in Rapidly Evolving Plastid–Nuclear Enzyme Complexes. Genetics 204, 1507–1522. https://doi.org/10.1534/genetics.116.188268

Rousseau-Gueutin, M., Huang, X., Higginson, E., Ayliffe, M., Day, A., Timmis, J.N., 2013. Potential Functional Replacement of the Plastidic Acetyl-CoA Carboxylase Subunit (accD) Gene by Recent Transfers to the Nucleus in Some Angiosperm Lineages. Plant Physiology 161, 1918–1929. https://doi.org/10.1104/pp.113.214528

Rudella, A., Friso, G., Alonso, J.M., Ecker, J.R., Wijk, K.J. van, 2006. Downregulation of ClpR2 Leads to Reduced Accumulation of the ClpPRS Protease Complex and Defects in Chloroplast Biogenesis in Arabidopsis. The Plant Cell 18, 1704–1721. https://doi.org/10.1105/tpc.106.042861

Salie, M.J., Thelen, J.J., 2016. Regulation and structure of the heteromeric acetyl-CoA carboxylase. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids, Plant Lipid Biology 1861, 1207–1213. https://doi.org/10.1016/j.bbalip.2016.04.004

Sasaki, Y., Nagano, Y., 2004. Plant Acetyl-CoA Carboxylase: Structure, Biosynthesis, Regulation, and Gene Manipulation for Plant Breeding. Bioscience, Biotechnology, and Biochemistry 68, 1175–1184. https://doi.org/10.1271/bbb.68.1175

Schulte, W., Töpfer, R., Stracke, R., Schell, J., Martini, N., 1997. Multi-functional acetyl-CoA carboxylase from Brassica napus is encoded by a multi-gene family: Indication for plastidic localization of at least one isoform. PNAS 94, 3465–3470. https://doi.org/10.1073/pnas.94.7.3465

Shultz, J.L., Kurunam, D., Shopinski, K., Iqbal, M.J., Kazi, S., Zobrist, K., Bashir, R., Yaegashi, S., Lavu, N., Afzal, A.J., Yesudas, C.R., Kassem, M.A., Wu, C., Zhang, H.B., Town, C.D., Meksem, K., Lightfoot, D.A., 2006. The Soybean Genome Database (SoyGD): a browser for display of duplicated, polyploid, regions and sequence tagged sites on the integrated physical and genetic maps of Glycine max. Nucleic Acids Res 34, D758-765. https://doi.org/10.1093/nar/gkj050

Sikosek, T., Chan, H.S., Bornberg-Bauer, E., 2012. Escape from Adaptive Conflict follows from weak functional trade-offs and mutational robustness. Proc Natl Acad Sci U S A 109, 14888–14893. https://doi.org/10.1073/pnas.1115620109

Sinkler, C.A., Kalpage, H., Shay, J., Lee, I., Malek, M.H., Grossman, L.I., Hüttemann, M., 2017. Tissue- and Condition-Specific Isoforms of Mammalian Cytochrome c Oxidase Subunits: From Function to Human Disease [WWW Document]. Oxidative Medicine and Cellular Longevity. https://doi.org/10.1155/2017/1534056

Sjögren, L.L.E., Clarke, A.K., 2011. Assembly of the chloroplast ATP-dependent Clp protease in Arabidopsis is regulated by the ClpT accessory proteins. Plant Cell 23, 322–332. https://doi.org/10.1105/tpc.110.082321

Sjögren, L.L.E., Stanne, T.M., Zheng, B., Sutinen, S., Clarke, A.K., 2006. Structural and Functional Insights into the Chloroplast ATP-Dependent Clp Protease in Arabidopsis. The Plant Cell 18, 2635–2649. https://doi.org/10.1105/tpc.106.044594

Sloan, D.B., Triant, D.A., Forrester, N.J., Bergner, L.M., Wu, M., Taylor, D.R., 2014. A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae (Caryophyllaceae). Molecular Phylogenetics and Evolution 72, 82–89. https://doi.org/10.1016/j.ympev.2013.12.004

Smith, M.D., Wertheim, J.O., Weaver, S., Murrell, B., Scheffler, K., Kosakovsky Pond, S.L., 2015. Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. Molecular Biology and Evolution 32, 1342–1353. https://doi.org/10.1093/molbev/msv022

Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D., Depamphilis, C.W., Wall, P.K., Soltis, P.S., 2009. Polyploidy and angiosperm diversification. Am J Bot 96, 336–348. https://doi.org/10.3732/ajb.0800079

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stanne, T.M., Pojidaeva, E., Andersson, F.I., Clarke, A.K., 2007. Distinctive Types of ATP-dependent Clp Proteases in Cyanobacteria. J. Biol. Chem. 282, 14394–14402. https://doi.org/10.1074/jbc.M700275200

Taylor, J.S., Raes, J., 2004. Duplication and Divergence: The Evolution of New Genes and Old Ideas. Annual Review of Genetics 38, 615–643. https://doi.org/10.1146/annurev.genet.38.072902.092831

The Angiosperm Phylogeny Group, Chase, M.W., Christenhusz, M.J.M., Fay, M.F., Byng, J.W., Judd, W.S., Soltis, D.E., Mabberley, D.J., Sennikov, A.N., Soltis, P.S., Stevens, P.F., 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Botanical Journal of the Linnean Society 181, 1–20. https://doi.org/10.1111/boj.12385

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R.R., Bhalerao, R.P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.-L., Cooper, D., Coutinho, P.M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Déjardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M.,

Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjärvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leplé, J.-C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D.R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouzé, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.-J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y., Rokhsar, D., 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313, 1596–1604. https://doi.org/10.1126/science.1128691

Van de Peer, Y., Taylor, J.S., Braasch, I., Meyer, A., 2001. The Ghost of Selection Past: Rates of Evolution and Functional Divergence of Anciently Duplicated Genes. J Mol Evol 53, 436–446. https://doi.org/10.1007/s002390010233

Wang, W., Zheng, H., Fan, C., Li, J., Shi, J., Cai, Z., Zhang, G., Liu, D., Zhang, J., Vang, S., Lu, Z., Wong, G.K.-S., Long, M., Wang, J., 2006. High Rate of Chimeric Gene Origination by Retroposition in Plant Genomes. The Plant Cell 18, 1791–1802. https://doi.org/10.1105/tpc.106.041905

Welsch, R., Zhou, X., Yuan, H., Álvarez, D., Sun, T., Schlossarek, D., Yang, Y., Shen, G., Zhang, H., Rodriguez-Concepcion, M., Thannhauser, T.W., Li, L., 2018. Clp Protease and OR Directly Control the Proteostasis of Phytoene Synthase, the Crucial Enzyme for Carotenoid Biosynthesis in Arabidopsis. Mol Plant 11, 149–162. https://doi.org/10.1016/j.molp.2017.11.003

Wendel, J.F., Lisch, D., Hu, G., Mason, A.S., 2018. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. Curr Opin Genet Dev 49, 1–7. https://doi.org/10.1016/j.gde.2018.01.004

Wertheim, J.O., Murrell, B., Smith, M.D., Kosakovsky Pond, S.L., Scheffler, K., 2015. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. Molecular Biology and Evolution 32, 820–832. https://doi.org/10.1093/molbev/msu400

Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A., Wagner, L., 2003. Database resources of the National Center for Biotechnology. Nucleic Acids Res 31, 28–33.

Wicke, S., Schneeweiss, G.M., dePamphilis, C.W., Müller, K.F., Quandt, D., 2011. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol 76, 273–297. https://doi.org/10.1007/s11103-011-9762-4

Williams, A.M., Friso, G., Wijk, K.J. van, Sloan, D.B., 2019. Extreme variation in rates of evolution in the plastid Clp protease complex. The Plant Journal 98, 243–259. https://doi.org/10.1111/tpj.14208

Williams, A.M., Itgen, M.W., Broz, A.K., Carter, O.G., Sloan, D.B., 2021. Long-read transcriptome and other genomic resources for the angiosperm Silene noctiflora. G3 Genes|Genomes|Genetics. https://doi.org/10.1093/g3journal/jkab189

Williams, A.V., Boykin, L.M., Howell, K.A., Nevill, P.G., Small, I., 2015. The Complete Sequence of the Acacia ligulata Chloroplast Genome Reveals a Highly Divergent clpP1 Gene. PLOS ONE 10, e0125768. https://doi.org/10.1371/journal.pone.0125768

Yang, Z., 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24, 1586–1591. https://doi.org/10.1093/molbev/msm088

Yang, Z., Nielsen, R., 2002. Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. Molecular Biology and Evolution 19, 908–917. https://doi.org/10.1093/oxfordjournals.molbev.a004148

Yu, A.Y.H., Houry, W.A., 2007. ClpP: A distinctive family of cylindrical energy-dependent serine proteases. FEBS Letters 581, 3749–3757. https://doi.org/10.1016/j.febslet.2007.04.076

Zhang, J., 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution 18, 292–298. https://doi.org/10.1016/S0169-5347(03)00033-8

Zhang, P., Gu, Z., Li, W.-H., 2003. Different evolutionary patterns between young duplicate genes in the human genome. Genome Biology 4, R56. https://doi.org/10.1186/gb-2003-4-9-r56

Zhang, Y., Ma, J., Yang, B., Li, R., Zhu, W., Sun, L., Tian, J., Zhang, L., 2014. The complete chloroplast genome sequence of Taxus chinensis var. mairei (Taxaceae): loss of an inverted repeat region and comparative analysis with related species. Gene 540, 201–209. https://doi.org/10.1016/j.gene.2014.02.037

APPENDIX: SUPPLEMENTARY TABLES AND FIGURES

Table S2.1. *clpP1* and *psaA* d_N, d_S, and d_N/d_S values in 25 angiosperms.

Species	*clpP1* d_N	*clpP1* d_S	*clpP1* d_N/d_S	*psaA* d_N	*psaA* d_S	*psaA* d_N/d_S
Ancestor of *S. bicolor* and *O. sativa*	0.149	0.292	0.511	0.009	0.147	0.058
Sorghum bicolor	0	0.061	0	0.003	0.058	0.046
Oryza sativa	0.012	0.038	0.323	0.007	0.073	0.096
Musa acuminata	0.019	0.077	0.252	0.004	0.084	0.047
Cucumis sativus	0.073	0.086	0.855	0.009	0.107	0.084
Prunus persica	0.031	0.087	0.357	0.001	0.085	0.015
Lathyrus sativus	0.211	0.285	0.742	0.002	0.07	0.031
Soja max	0.021	0.184	0.111	0.011	0.199	0.055
Medicago truncatula	0.126	0.2	0.632	0.003	0.04	0.069
Acacia ligulata	0.386	0.263	1.468	0.002	0.033	0.054
Ricinus communis	0.009	0.156	0.057	0.004	0.076	0.048
Populus trichocarpa	0.019	0.064	0.301	0.001	0.074	0.016
Arabidopsis thaliana	0.044	0.158	0.277	0.006	0.147	0.041
Gossypium raimondii	0.015	0.069	0.214	0.005	0.1	0.048
Eucalyptus grandis	0.004	0.046	0.087	0.003	0.033	0.09
Oenothera biennis	0.365	0.478	0.764	0.002	0.108	0.022
Geranium maderense	0.493	0.424	1.161	0.009	0.169	0.05
Vitis vinifera	0.063	0.042	1.494	0.003	0.051	0.059
Erythranthe lutea	0.001	0.053	0.023	0.002	0.069	0.026
Plantago maritima	0.679	0.638	1.065	0.001	0.108	0.011
Solanum lycopersicum	0.13	0.13	0.999	0.002	0.093	0.026
Lobelia siphilitica	0.297	0.378	0.785	0.008	0.197	0.038
Silene noctiflora	0.578	0.532	1.085	0.003	0.019	0.134
Silene latifolia	0	0.047	0	0.001	0.004	0.136
Liriodendron chinense	0.008	0.039	0.199	0.003	0.047	0.061
Amborella trichopoda	0.034	0.135	0.249	0.006	0.135	0.046

*** Values are reported for terminal branches only, with exception of the ancestor of *S. bicolor* and *O. sativa*.**

Table S2.2: Large dataset sampling

Species	Accession
Abies koreana	NC_026892
Acacia ligulata	NC_026134
Acer morrisonense	NC_029371
Acidosasa purpurea	NC_015820
Acorus americanus	NC_010093
Acorus calamus	NC_007407
Acorus gramineus	NC_026299
Actinidia chinensis	NC_026690
Actinidia deliciosa	NC_026691
Acutodesmus obliquus	NC_008101
Adenophora remotiflora	NC_026999
Adiantum capillus-veneris	NC_004766
Aegilops bicornis	NC_024831
Aegilops cylindrica	NC_023096
Aegilops geniculata	NC_023097
Aegilops kotschyi	NC_024832
Aegilops longissima	NC_024830
Aegilops searsii	NC_024815
Aegilops sharonensis	NC_024816
Aegilops speltoides	NC_022135
Aegilops tauschii	NC_022133
Aethionema cordifolium	NC_009265
Aethionema grandiflorum	NC_009266
Agathis dammara	NC_023119
Ageratina adenophora	NC_015621
Agrostemma githago	NC_023357
Agrostis stolonifera	NC_008591
Ajuga reptans	NC_023102
Akebia trifoliata	NC_029427
Allium cepa	NC_024813
Allosyncarpia ternata	NC_022413
Alloteropsis angusta	NC_027951
Alloteropsis cimicina	NC_027952
Alloteropsis semialata	NC_027824
Alsophila spinulosa	NC_012818
Amborella trichopoda	NC_005086

Amentotaxus argotaenia	NC_027581
Amentotaxus formosana	NC_024945
Ammophila breviligulata	NC_027465
Ampelocalamus calcareus	NC_024731
Ampelodesmos mauritanicus	NC_027466
Ananas comosus	NC_026220
Andrographis paniculata	NC_022451
Anethum graveolens	NC_029470
Aneura mirabilis	NC_010359
Angelica acutiloba	NC_029391
Angelica dahurica	NC_029392
Angelica gigas	NC_029393
Angiopteris angustifolia	NC_026300
Angiopteris evecta	NC_008829
Angophora costata	NC_022412
Angophora floribunda	NC_022411
Anomochloa marantoidea	NC_014062
Anthoceros angustus	NC_004543
Anthoxanthum nitens	NC_027475
Anthoxanthum odoratum	NC_027467
Anthriscus cerefolium	NC_015113
Aquilaria sinensis	NC_029243
Arabidopsis arenosa	NC_029334
Arabidopsis cebennensis	NC_029335
Arabidopsis pedemontana	NC_029336
Arabidopsis thaliana	NC_000932
Arabis alpina	NC_023367
Arabis hirsuta	NC_009268
Aralia undulata	NC_022810
Araucaria heterophylla	NC_026450
Ardisia polysticta	NC_021121
Aristida purpurea	NC_025228
Artemisia frigida	NC_020607
Artemisia montana	NC_025910
Arundinaria appalachiana	NC_023934
Arundinaria fargesii	NC_024712
Arundinaria gigantea	NC_020341
Arundinaria tecta	NC_023935
Asclepias nivea	NC_022431

Asclepias syriaca	NC_022432
Aster spathulifolius	NC_027434
Astragalus nakaianus	NC_028171
Atropa belladonna	NC_004561
Auxenochlorella protothecoides	NC_023775
Avena sativa	NC_027468
Azadirachta indica	NC_023792
Bambusa arnhemica	NC_026958
Bambusa bambos	NC_026957
Bambusa emeiensis	NC_015830
Bambusa multiplex	NC_024668
Bambusa oldhamii	NC_012927
Barbarea verna	NC_009269
Bathycoccus prasinos	NC_024811
Berberis bealei	NC_022457
Bismarckia nobilis	NC_020366
Bletilla ochracea	NC_029483
Bletilla striata	NC_028422
Bomarea edulis	NC_025306
Boswellia sacra	NC_029420
Botryococcus braunii	NC_025545
Boulardia latisquama	NC_025641
Bouteloua curtipendula	NC_029414
Bowenia serrulata	NC_026036
Brachyelytrum aristosum	NC_027470
Brachypodium distachyon	NC_011032
Bracteacoccus giganteus	NC_028586
Brassaiopsis hainla	NC_022811
Brassica juncea	NC_028272
Brassica napus	NC_016734
Brassica rapa subsp.	NC_015139
Brighamia insignis	NC_028633
Briza maxima	NC_027471
Bromus vulgaris	NC_027472
Bryopsis hypnoides	NC_013359
Bryopsis plumosa	NC_026795
Buergersiochloa bambusoides	NC_026968
Bupleurum falcatum	NC_027834
Buxus microphylla	NC_009599

Calamus caryotoides	NC_020365
Calanthe triplicata	NC_024544
Callitropsis nootkatensis	NC_026295
Callitropsis vietnamensis	NC_026298
Calocedrus formosana	NC_023121
Calycanthus floridus var.	NC_004993
Camelina sativa	NC_029337
Camellia crapnelliana	NC_024541
Camellia cuspidata	NC_022459
Camellia danzaiensis	NC_022460
Camellia grandibracteata	NC_024659
Camellia impressinervis	NC_022461
Camellia leptophylla	NC_024660
Camellia oleifera	NC_023084
Camellia petelotii	NC_024661
Camellia pitardii	NC_022462
Camellia pubicosta	NC_024662
Camellia reticulata	NC_024663
Camellia sinensis	NC_020019
Camellia taliensis	NC_022264
Camellia yunnanensis	NC_022463
Campanula takesimana	NC_026203
Campynema lineare	NC_026785
Cannabis sativa	NC_026562
Cannabis sativa	NC_027223
Capsella bursa-pastoris	NC_009270
Capsella grandiflora	NC_028517
Capsella rubella	NC_027693
Capsicum annuum	NC_018552
Capsicum frutescens	NC_028007
Capsicum lycianthoides	NC_026551
Cardamine impatiens	NC_026445
Cardamine resedifolia	NC_026446
Carex siderosticta	NC_027250
Carica papaya	NC_010323
Carludovica palmata	NC_026786
Carnegiea gigantea	NC_027618
Carteria cerasiformis	NC_028585
Castanea mollissima	NC_014674

Castanopsis echinocarpa	NC_023801
Catharanthus roseus	NC_021423
Cathaya argyrophylla	NC_014589
Cattleya crispata	NC_026568
Cedrus deodara	NC_014575
Cenchrus americanus	NC_024171
Centaurea diffusa	NC_024286
Centotheca lappacea	NC_025229
Centropodia glauca	NC_029411
Cephalotaxus oliveri	NC_021110
Cephalotaxus wilsoniana	NC_016063
Ceratophyllum demersum	NC_009962
Ceratozamia hildae	NC_026037
Chaetosphaeridium globosum	NC_004115
Chara vulgaris	NC_008097
Characiochloris acuminata	NC_028584
Chikusichloa aquatica	NC_027184
Chimonocalamus longiusculus	NC_024714
Chionochloa macra	NC_025230
Chlamydomonas reinhardtii	NC_005353
Chloranthus japonicus	NC_026565
Chloranthus spicatus	NC_009598
Chlorella 'Chlorella' mirabilis	NC_025528
Chlorella sorokiniana	NC_023835
Chlorella variabilis	NC_015359
Chlorella vulgaris	NC_001865
Chlorokybus atmophyticus	NC_008822
Choricystis parasitica	NC_025539
Chrysanthemum indicum	NC_020320
Chrysanthemum x	NC_020092
Chrysobalanus icaco	NC_024061
Chusquea circinata	NC_027490
Chusquea liebmannii	NC_026969
Chusquea spectabilis	NC_026959
Cicer arietinum	NC_011163
Cistanche deserticola	NC_021111
Cistanche phelypaea	NC_025642
Citrus aurantiifolia	NC_024929
Citrus sinensis	NC_008334

Clematis terniflora	NC_028000
Coccomyxa subellipsoidea C-169	NC_015084
Cochlearia borzaeana	NC_029253
Cochlearia islandica	NC_029254
Cochlearia pyrenaica	NC_029331
Cochlearia tridactylites	NC_029332
Cocos nucifera	NC_022417
Coffea arabica	NC_008535
Coix lacryma-jobi	NC_013273
Coleataenia prionitis	NC_025231
Colobanthus quitensis	NC_028080
Colocasia esculenta	NC_016753
Colpothrinax cookii	NC_028026
Conopholis americana	NC_023131
Corallorhiza bulbosa	NC_025659
Corallorhiza macrantha	NC_025660
Corallorhiza mertensiana	NC_025661
Corallorhiza odontorhiza	NC_025664
Corallorhiza trifida	NC_025662
Corallorhiza wisteriana	NC_025663
Corymbia eximia	NC_022409
Corymbia gummifera	NC_022407
Corymbia henryi	NC_028409
Corymbia maculata	NC_022408
Corymbia tessellaris	NC_022410
Corymbia torelliana	NC_028410
Corynocarpus laevigata	NC_014807
Couepia guianensis	NC_024063
Crucihimalaya wallichii	NC_009271
Cryophytum crystallinum	NC_029049
Cryptomeria japonica	NC_010548
Cucumis hystrix	NC_023544
Cucumis melo subsp.	NC_015983
Cucumis sativus	NC_007144
Cunninghamia lanceolata	NC_021437
Cupressus gigantea	NC_028155
Cupressus sempervirens	NC_026296
Curcuma flaviflora	NC_028729
Curcuma roscoeana	NC_022928

Cuscuta exaltata	NC_009963
Cuscuta gronovii	NC_009765
Cuscuta obtusiflora	NC_009949
Cuscuta reflexa	NC_009766
Cycas revoluta	NC_020319
Cycas taitungensis	NC_009618
Cymbidium aloifolium	NC_021429
Cymbidium ensifolium	NC_028525
Cymbidium faberi	NC_027743
Cymbidium goeringii	NC_028524
Cymbidium mannii	NC_021433
Cymbidium sinense	NC_021430
Cymbidium tortisepalum	NC_021431
Cymbidium tracyanum	NC_021432
Cynanchum auriculatum	NC_029460
Cynanchum wilfordii	NC_029459
Cynara baetica	NC_028005
Cynara cornigera	NC_028006
Cynara humilis	NC_027113
Cypripedium formosanum	NC_026772
Cypripedium japonicum	NC_027227
Cypripedium macranthos	NC_024421
Cyrtomium devexiscapulae	NC_028542
Cyrtomium falcatum	NC_028705
Dactylis glomerata	NC_027473
Danthonia californica	NC_025232
Dasypogon bromeliifolius	NC_020367
Datura stramonium	NC_018117
Daucus carota	NC_008325
Dendrobium catenatum	NC_024019
Dendrobium chrysotoxum	NC_028549
Dendrobium huoshanense	NC_028430
Dendrobium nobile	NC_029456
Dendrobium strongylanthum	NC_027691
Dendrocalamus latiflorus	NC_013088
Dendropanax dentiger	NC_026546
Dendropanax morbifer	NC_027607
Deschampsia antarctica	NC_023533
Diandrolyra sp.	NC_026960

Diarrhena obovata	NC_027474
Dicloster acuatus	NC_025546
Dictyochloropsis reticulata	NC_025524
Dieffenbachia seguine	NC_027272
Digitaria exilis	NC_024176
Dioon spinulosum	NC_027512
Dioscorea elephantipes	NC_009601
Dioscorea rotundata	NC_024170
Dioscorea zingiberensis	NC_027090
Diplopterygium glaucum	NC_024158
Dipteronia sinensis	NC_029338
Dorcoceras hygrometricum	NC_016468
Draba nemorosa	NC_009272
Drimys granadensis	NC_008456
Dunalia brachyacantha	NC_026906
Dunalia obovata	NC_026563
Dunalia solanacea	NC_027099
Dunaliella salina	NC_016732
Echinochloa crus-galli	NC_028719
Echinochloa oryzicola	NC_024643
Echites umbellatus	NC_025655
Elaeagnus macrophylla	NC_028066
Elaeis guineensis	NC_017602
Eleutherococcus senticosus	NC_016430
Elleanthus sodiroi	NC_027266
Elliptochloris bilobata	NC_025548
Elodea canadensis	NC_018541
Elytrophorus spicatus	NC_025233
Encephalartos lehmannii	NC_027514
Ephedra equisetina	NC_011954
Ephedra foeminea	NC_029347
Epifagus virginiana	NC_001568
Epimedium sagittatum	NC_029428
Epipogium aphyllum	NC_026449
Epipogium roseum	NC_026448
Epipremnum aureum	NC_027954
Equisetum arvense	NC_014699
Equisetum hyemale	NC_020146
Eragrostis minor	NC_029412

Eragrostis tef	NC_029413
Eriachne stipacea	NC_025234
Erodium absinthoides	NC_026847
Erodium carvifolium	NC_015083
Erodium chrysanthum	NC_027065
Erodium crassifolium	NC_025906
Erodium gruinum	NC_025907
Erodium texanum	NC_014569
Erodium trifolium	NC_024635
Erycina pusilla	NC_018114
Ettlia pseudoalveolaris	NC_025532
Eucalyptus aromaphloia	NC_022396
Eucalyptus baxteri	NC_022382
Eucalyptus camaldulensis	NC_022398
Eucalyptus cladocalyx	NC_022394
Eucalyptus cloeziana	NC_022388
Eucalyptus curtisii	NC_022391
Eucalyptus deglupta	NC_022399
Eucalyptus delegatensis	NC_022380
Eucalyptus diversicolor	NC_022402
Eucalyptus diversifolia	NC_022383
Eucalyptus elata	NC_022385
Eucalyptus erythrocorys	NC_022406
Eucalyptus globulus subsp.	NC_008115
Eucalyptus grandis	NC_014570
Eucalyptus guilfoylei	NC_022405
Eucalyptus marginata	NC_022390
Eucalyptus melliodora	NC_022392
Eucalyptus microcorys	NC_022404
Eucalyptus nitens	NC_022395
Eucalyptus obliqua	NC_022378
Eucalyptus patens	NC_022389
Eucalyptus polybractea	NC_022393
Eucalyptus radiata	NC_022379
Eucalyptus regnans	NC_022386
Eucalyptus saligna	NC_022397
Eucalyptus salmonophloia	NC_022403
Eucalyptus sieberi	NC_022384
Eucalyptus spathulata	NC_022400

Eucalyptus torquata	NC_022401
Eucalyptus umbra	NC_022387
Eucalyptus verrucata	NC_022381
Eugenia uniflora	NC_027744
Euonymus japonicus	NC_028067
Euptelea pleiosperma	NC_029429
Eustrephus latifolius	NC_025305
Eutrema botschantzevii	NC_029379
Eutrema halophilum	NC_029378
Eutrema heterophyllum	NC_028728
Eutrema salsugineum	NC_028170
Eutrema yunnanense	NC_028727
Fagopyrum esculentum subsp.	NC_010776
Fagopyrum tataricum	NC_027161
Fargesia nitida	NC_024715
Fargesia spathacea	NC_024716
Fargesia yunnanensis	NC_024717
Fatsia japonica	NC_027685
Ferrocalamus rimosivaginus	NC_015831
Festuca altissima	NC_019648
Festuca arundinacea	NC_011713
Festuca ovina	NC_019649
Festuca pratensis	NC_019650
Ficus racemosa	NC_028185
Floydiella terrestris	NC_014346
Foeniculum vulgare	NC_029469
Fragaria chiloensis	NC_019601
Fragaria iinumae	NC_024258
Fragaria mandshurica	NC_018767
Fragaria vesca subsp.	NC_015206
Fragaria vesca subsp.	NC_018766
Fragaria virginiana	NC_019602
Francoa sonchifolia	NC_021101
Fritillaria cirrhosa	NC_024728
Fritillaria hupehensis	NC_024736
Fritillaria taipaiensis	NC_023247
Fusochloris perforata	NC_025543
Gaoligongshania megalothyrsa	NC_024718
Gelidocalamus tessellatus	NC_024719

Geminella minor	NC_025544
Genlisea margaretae	NC_025652
Gentiana crassicaulis	NC_027442
Gentiana straminea	NC_027441
Geranium palmatum	NC_014573
Ginkgo biloba	NC_016986
Gloeotilopsis sterilis	NC_025538
Glycine canescens	NC_021647
Glycine cyrtoloba	NC_021645
Glycine dolichocarpa	NC_021648
Glycine falcata	NC_021649
Glycine stenophita	NC_021646
Glycine syndetika	NC_021650
Glycine tomentella	NC_021636
Glycyrrhiza glabra	NC_024038
Gnetum gnemon	NC_026301
Gnetum montanum	NC_021438
Gnetum parvifolium	NC_011942
Gnetum ula	NC_028734
Gonium pectorale	NC_020438
Goodyera fumata	NC_026773
Goodyera procera	NC_029363
Goodyera schlechtendaliana	NC_029364
Goodyera velutina	NC_029365
Gossypium anomalum	NC_023213
Gossypium arboreum	NC_016712
Gossypium areysianum	NC_018112
Gossypium barbadense	NC_008641
Gossypium bickii	NC_023214
Gossypium capitis-viridis	NC_018111
Gossypium darwinii	NC_016670
Gossypium gossypioides	NC_017894
Gossypium herbaceum	NC_023215
Gossypium herbaceum subsp.	NC_016692
Gossypium hirsutum	NC_007944
Gossypium incanum	NC_018109
Gossypium longicalyx	NC_023216
Gossypium mustelinum	NC_016711
Gossypium raimondii	NC_016668

Gossypium robinsonii	NC_018113
Gossypium somalense	NC_018110
Gossypium stocksii	NC_023217
Gossypium sturtianum	NC_023218
Gossypium thurberi	NC_015204
Gossypium tomentosum	NC_016690
Gossypium turneri	NC_026835
Greslania sp.	NC_026961
Guadua chacoensis	NC_029232
Guadua weberbaueri	NC_026991
Guizotia abyssinica	NC_010601
Gynochthodes nanlingensis	NC_028614
Gynochthodes officinalis	NC_028009
Gynostemma pentaphyllum	NC_029484
Habenaria pantlingiana	NC_026775
Hafniomonas laevis	NC_028583
Hakonechloa macra	NC_025235
Haloxylon ammodendron	NC_027668
Haloxylon persicum	NC_027669
Hanabusaya asiatica	NC_024732
Helianthus annuus	NC_007977
Helianthus decapetalus	NC_023110
Helianthus divaricatus	NC_023109
Helianthus giganteus	NC_023107
Helianthus grosseserratus	NC_023108
Helianthus hirsutus	NC_023111
Helianthus maximiliani	NC_023114
Helianthus strumosus	NC_023113
Helianthus tuberosus	NC_023112
Heliconia collinsiana	NC_020362
Helicosporidium sp.	NC_008100
Helictochloa hookeri	NC_027469
Heloniopsis tubiflora	NC_027159
Hesperelaea palmeri	NC_025787
Hesperocyparis glabra	NC_026297
Hevea brasiliensis	NC_015308
Hibiscus syriacus	NC_026909
Hickelia madagascariensis	NC_026962
Hilaria cenchroides	NC_029415

Hirtella physophora	NC_024066
Hirtella racemosa	NC_024060
Hordeum jubatum	NC_027476
Hordeum vulgare subsp.	NC_008590
Humulus lupulus	NC_028032
Huperzia lucidula	NC_006861
Hydnora visseri	NC_029358
Hyoscyamus niger	NC_024261
Hypseocharis bilobata	NC_023260
Illicium oligandrum	NC_009600
Indocalamus longiauritus	NC_015803
Indocalamus wilsonii	NC_024720
Indosasa sinica	NC_024721
Inga leiocalycina	NC_028732
Interfilum terricola	NC_025542
Iochroma loxense	NC_026726
Iochroma nitidum	NC_026567
Iochroma stenanthum	NC_026574
Iochroma tingoanum	NC_027177
Ionopsidium acaule	NC_029333
Ipomoea batatas	NC_026703
Ipomoea purpurea	NC_009808
Iris gatesii	NC_024936
Iris sanguinea	NC_029227
Isachne distichophylla	NC_025236
Isatis tinctoria	NC_028415
Isoetes flaccida	NC_014675
Jacobaea vulgaris	NC_015543
Jasminum nudiflorum	NC_008407
Jatropha curcas	NC_012224
Jenufa minuta	NC_028582
Jenufa perforata	NC_028581
Juglans regia	NC_028617
Juniperus bermudiana	NC_024021
Juniperus cedrus	NC_028190
Juniperus monosperma	NC_024022
Juniperus scopulorum	NC_024023
Juniperus virginiana	NC_024024
Kalopanax septemlobus	NC_022814

Keteleeria davidiana	NC_011930
Klebsormidium flaccidum	NC_024167
Koliella corcontica	NC_025536
Koliella longiseta	NC_025531
Lactuca sativa	NC_007578
Larix decidua	NC_016058
Larrea tridentata	NC_028023
Lathraea squamaria	NC_027838
Lathyrus clymenum	NC_027148
Lathyrus davidii	NC_027073
Lathyrus graminifolius	NC_027074
Lathyrus inconspicuus	NC_027149
Lathyrus japonicus	NC_027075
Lathyrus littoralis	NC_027076
Lathyrus ochroleucus	NC_027077
Lathyrus odoratus	NC_027150
Lathyrus palustris	NC_027078
Lathyrus pubescens	NC_027079
Lathyrus sativus	NC_014063
Lathyrus tingitanus	NC_027151
Lathyrus venosus	NC_027080
Lavandula angustifolia	NC_029370
Lecomtella madagascariensis	NC_024106
Leersia tisserantii	NC_016677
Lemna minor	NC_010109
Lens culinaris	NC_027152
Leontopodium leiolepis	NC_027835
Lepidium virginicum	NC_009273
Lepidozamia peroffskyana	NC_027513
Leptosira terrestris	NC_009681
Leucaena trichandra	NC_028733
Licania alba	NC_024064
Licania heteromorpha	NC_024062
Licania sprucei	NC_024065
Ligusticum tenuissimum	NC_029394
Lilium hansonii	NC_027674
Lilium sp.	NC_027679
Lilium superbum	NC_026787
Lilium tsingtauense	NC_027675

Lindenbergia philippensis	NC_022859
Liquidambar formosana	NC_023092
Liriodendron tulipifera	NC_008326
Lithachne pauciflora	NC_026970
Lithocarpus balansae	NC_026577
Lobosphaera incisa	NC_025533
Lobularia maritima	NC_009274
Lolium multiflorum	NC_019651
Lolium perenne	NC_009950
Lonicera japonica	NC_026839
Lotus japonicus	NC_002694
Lupinus luteus	NC_023090
Luzuriaga radicans	NC_025333
Lycopersicon cheesmaniae	NC_026876
Lycopersicon chilense	NC_026877
Lycopersicon galapagense	NC_026878
Lycopersicon habrochaites	NC_026879
Lycopersicon lycopersicum	NC_007898
Lycopersicon lycopersicum	AC_000188
Lycopersicon neorickii	NC_026880
Lycopersicon peruvianum	NC_026881
Lycopersicon pimpinellifolium	NC_026882
Lygodium japonicum	NC_022136
Lysimachia coreana	NC_026197
Macadamia integrifolia	NC_025288
Machilus balansae	NC_028074
Machilus yunnanensis	NC_028073
Macrozamia mountperriensis	NC_027511
Magnolia denudata	NC_018357
Magnolia grandiflora	NC_020318
Magnolia kwangsiensis	NC_015892
Magnolia officinalis	NC_020316
Magnolia officinalis subsp.	NC_020317
Magnolia tripetala	NC_024027
Magnolia yunnanensis	NC_024545
Manihot esculenta	NC_010433
Mankyua chejuensis	NC_017006
Marchantia polymorpha	NC_001319
Marsilea crenata	NC_022137

Marvania geminata	NC_025549
Masdevallia coccinea	NC_026541
Masdevallia picturata	NC_026777
Medicago hybrida	NC_027153
Medicago papillosa	NC_027154
Medicago truncatula	NC_003119
Megaleranthis saniculifolia	NC_012615
Melianthus villosus	NC_023256
Melica mutica	NC_027477
Melica subulata	NC_027478
Meliosma aff. cuneifolia Moore 333	NC_029430
Mesostigma viride	NC_002186
Mesotaenium endlicherianum	NC_024169
Metanarthecium luteoviride	NC_029214
Metapanax delavayi	NC_022812
Metasequoia glyptostroboides	NC_027423
Micromonas sp.	NC_012575
Microthamnion kuetzingianum	NC_025537
Millettia pinnata	NC_016708
Miscanthus sacchariflorus	NC_028720
Miscanthus sinensis	NC_028721
Monachather paradoxus	NC_025237
Monomastix sp.	NC_012101
Monsonia speciosa	NC_014582
Morus indica	NC_008359
Morus mongolica	NC_025772
Morus notabilis	NC_027110
Musa balbisiana	NC_028439
Musa textilis	NC_022926
Mychonastes jurisii	NC_028579
Myriopteris lindheimeri	NC_014592
Myrmecia israelensis	NC_025525
Nageia nagi	NC_023120
Najas flexilis	NC_021936
Nandina domestica	NC_008336
Nasturtium officinale	NC_009275
Nelumbo lutea	NC_015605
Nelumbo nucifera	NC_025339
Neocystis brevis	NC_025535

Neohouzeaua sp.	NC_026963
Neololeba atra	NC_026964
Neottia nidus-avis	NC_016471
Nephroselmis astigmatica	NC_024829
Nephroselmis olivacea	NC_000927
Nerium oleander	NC_025656
Neyraudia reynaudiana	NC_024262
Nicotiana sylvestris	NC_007500
Nicotiana tabacum	NC_001879
Nicotiana tomentosiformis	NC_007602
Nicotiana undulata	NC_016068
Nothoceros aenigmaticus	NC_020259
Nuphar advena	NC_008788
Nyholmiella obtusifolia	NC_026979
Nymphaea alba	NC_006050
Nymphaea mexicana	NC_024542
Oedogonium cardiacum	NC_011031
Oenothera argillicola	NC_010358
Oenothera biennis	NC_010361
Oenothera elata subsp.	NC_002693
Oenothera glazioviana	NC_010360
Oenothera grandiflora	NC_029211
Oenothera oakesiana	NC_029212
Oenothera parviflora	NC_010362
Olea europaea	NC_013707
Olea europaea subsp.	NC_015401
Olea europaea subsp.	NC_015604
Olea europaea subsp.	NC_015623
Olea woodiana subsp.	NC_015608
Oligostachyum shiuyingianum	NC_024722
Olimarabidopsis pumila	NC_009267
Olmeca reflexa	NC_026965
Oltmannsiellopsis viridis	NC_008099
Olyra latifolia	NC_024165
Oncidium hybrid	NC_014056
Oncidium sphacelatum	NC_028148
Oncinotis tenuiloba	NC_025657
Oogamochlamys gigantea	NC_028580
Ophioglossum californicum	NC_020147

Orobanche californica	NC_025651
Orobanche crenata	NC_024845
Orobanche gracilis	NC_023464
Orthotrichum rogeri	NC_026212
Oryza australiensis	NC_024608
Oryza barthii	NC_027460
Oryza glaberrima	NC_024175
Oryza glumipatula	NC_027461
Oryza longistaminata	NC_027462
Oryza meridionalis	NC_016927
Oryza nivara	NC_005973
Oryza officinalis	NC_027463
Oryza punctata	NC_027676
Oryza rufipogon	NC_017835
Oryza sativa Indica Group	NC_027678
Oryza sativa Indica Group	NC_008155
Oryza sativa Japonica Group	NC_001320
Oryzopsis asperifolia	NC_027479
Osmundastrum cinnamomeum	NC_024157
Ostericum grosseserratum	NC_028618
Ostreococcus tauri	NC_008289
Ostrya rehderiana	NC_028349
Osyris alba	NC_027960
Otatea acuminata	NC_026971
Otatea glauca	NC_028631
Pabia signiensis	NC_025529
Pachycladon cheesemanii	NC_021102
Pachycladon enysii	NC_018565
Pachysandra terminalis	NC_029433
Paeonia obovata	NC_026076
Panax ginseng	NC_006290
Panax japonicus	NC_028703
Panax notoginseng	NC_026447
Panax quinquefolius	NC_027456
Panax vietnamensis	NC_028704
Panicum virgatum	NC_015990
Papaver somniferum	NC_029434
Paphiopedilum armeniacum	NC_026779
Paphiopedilum niveum	NC_026776

Parachlorella kessleri	NC_012978
Paradoxia multiseta	NC_025540
Pariana campestris	NC_027491
Pariana radiciflora	NC_026972
Parinari campestris	NC_024067
Paris verticillata	NC_024560
Parthenium argentatum	NC_013553
Pastinaca pimpinellifolia	NC_027450
Pedinomonas minor	NC_016733
Pedinomonas tuberculata	NC_025530
Pelargonium alternans	NC_023261
Pelargonium australe	NC_028053
Pelargonium cotyledonis	NC_028052
Pelargonium dichondrifolium	NC_028051
Pelargonium x	NC_008454
Pellia endiviifolia	NC_019628
Pentactina rupicola	NC_016921
Pentalinon luteum	NC_025658
Penthorum chinense	NC_023086
Petrosavia stellaris	NC_023356
Phacotus lenticularis	NC_028587
Phaenosperma globosum	NC_027480
Phalaenopsis aphrodite subsp.	NC_007499
Phalaenopsis equestris	NC_017609
Phalaenopsis hybrid	NC_025593
Phalaris arundinacea	NC_027481
Pharus lappulaceus	NC_023245
Pharus latifolius	NC_021372
Phaseolus vulgaris	NC_009259
Phelipanche purpurea	NC_023132
Phelipanche ramosa	NC_023465
Phleum alpinum	NC_027482
Phoenix dactylifera	NC_013991
Phragmipedium longifolium	NC_028149
Phragmites australis	NC_022958
Phyllostachys edulis	NC_015817
Phyllostachys nigra var.	NC_015826
Phyllostachys propinqua	NC_016699
Phyllostachys sulphurea	NC_024669

Physalis peruviana	NC_026570
Physcomitrella patens	NC_005087
Picea abies	NC_021456
Picea glauca	NC_028594
Picea jezoensis	NC_029374
Picea morrisonicola	NC_016069
Picea sitchensis	NC_011152
Picocystis salinarum	NC_024828
Pilostyles aethiopica	NC_029235
Pilostyles hamiltonii	NC_029236
Pinellia ternata	NC_027681
Pinguicula ehlersiae	NC_023463
Pinus contorta	NC_011153
Pinus massoniana	NC_021439
Pinus tabuliformis	NC_028531
Pinus taeda	NC_021440
Pinus taiwanensis	NC_027415
Pinus thunbergii	NC_001631
Piper cenocladum	NC_008457
Piper kadsura	NC_027941
Piptochaetium avenaceum	NC_027483
Pisum sativum	NC_014057
Planctonema lauterbornii	NC_025541
Plantago maritima	NC_028519
Plantago media	NC_028520
Platanus occidentalis	NC_008335
Pleioblastus maculatus	NC_024723
Pleodorina starrii	NC_021109
Poa palustris	NC_027484
Podocarpus lambertii	NC_023805
Podocarpus totara	NC_020361
Podococcus barteri	NC_027276
Polygonatum cyrtonema	NC_028429
Polygonatum sibiricum	NC_029485
Polygonatum verticillatum	NC_028523
Populus alba	NC_008235
Populus balsamifera	NC_024735
Populus euphratica	NC_024747
Populus fremontii	NC_024734

Populus tremula	NC_027425
Populus tremula x Populus alba	NC_028504
Populus trichocarpa	NC_009143
Prasinoderma coloniale	NC_024817
Praxelis clematidea	NC_023833
Premna microphylla	NC_026291
Primula poissonii	NC_024543
Prinsepia utilis	NC_021455
Prunus kansuensis	NC_023956
Prunus maximowiczii	NC_026981
Prunus mume	NC_023798
Prunus padus	NC_026982
Prunus persica	NC_014697
Prunus yedoensis	NC_026980
Pseudendoclonium akinetum	NC_008114
Pseudochloris wilhelmii	NC_025547
Pseudophoenix vinifera	NC_020364
Pseudosasa japonica	NC_028328
Pseudotsuga sinensis var.	NC_016064
Psilotum nudum	NC_003386
Pteridium aquilinum subsp.	NC_014348
Ptilidium pulcherrimum	NC_015402
Puccinellia nuttalliana	NC_027485
Puelia olyriformis	NC_023449
Pycnococcus provasolii	NC_012097
Pyramimonas parkeae	NC_012099
Pyrus pyrifolia	NC_015996
Pyrus spinosa	NC_023130
Quercus aliena	NC_026790
Quercus aquifolioides	NC_026913
Quercus baronii	NC_029490
Quercus rubra	NC_020152
Quercus spinosa	NC_026907
Raddia brasiliensis	NC_026966
Ranunculus macranthus	NC_008796
Raphanus sativus	NC_024469
Ravenala madagascariensis	NC_022927
Retrophyllum piresii	NC_024827
Rhazya stricta	NC_024292

Rheum palmatum	NC_027728
Rhizanthella gardneri	NC_014874
Rhynchoryza subulata	NC_016718
Ricinus communis	NC_016736
Rosmarinus officinalis	NC_027259
Roya anglica	NC_024168
Sabal domingensis	NC_026444
Sabia yunnanensis	NC_029431
Saccharum officinarum complex hybrid	NC_029221
Saccharum officinarum complex hybrid cultivar NCo 310	NC_006084
Saccharum officinarum complex hybrid cultivar SP80-3280	NC_005878
Salicornia bigelovii	NC_027226
Salicornia brachiata	NC_027224
Salicornia europaea	NC_027225
Salix babylonica	NC_028350
Salix interior	NC_024681
Salix purpurea	NC_026722
Salix suchowensis	NC_026462
Salvia miltiorrhiza	NC_020431
Sanionia uncinata	NC_025668
Sapindus mukorossi	NC_025554
Saracha punctata	NC_026694
Sarocalamus faberi	NC_024713
Sartidia dewinteri	NC_027147
Sartidia perrieri	NC_027146
Saussurea involucrata	NC_029465
Schefflera delavayi	NC_022813
Schizomeris leibleinii	NC_015645
Schrenkiella parvula	NC_028726
Schwalbea americana	NC_023115
Sciaphila densiflora	NC_027659
Scrophularia takesimensis	NC_026202
Scutellaria baicalensis	NC_027262
Scutellaria insignis	NC_028533
Secale cereale	NC_021761
Sedum oryzifolium	NC_027837
Sedum sarmentosum	NC_023085
Sedum takesimense	NC_026065
Selaginella moellendorffii	NC_013086

Sesamum indicum	NC_016433
Seseli montanum	NC_027451
Setaria italica	NC_022850
Setaria viridis	NC_028075
Silene chalcedonica	NC_023359
Silene conica	NC_016729
Silene conoidea	NC_023358
Silene latifolia	NC_016730
Silene noctiflora	NC_016728
Silene paradoxa	NC_023360
Silene vulgaris	NC_016727
Silybum marianum	NC_028027
Sinopodophyllum hexandrum	NC_027732
Sobralia aff. bouchei HTK-2015	NC_028209
Sobralia callosa	NC_028147
Soja max	NC_007942
Soja soja	NC_022868
Solanum bulbocastanum	NC_007943
Solanum commersonii	NC_028069
Solanum nigrum	NC_028070
Solanum tuberosum	NC_008096
Sorghum bicolor	NC_008602
Sorghum timorense	NC_023800
Spinacia oleracea	NC_002202
Spirodela polyrhiza	NC_015891
Sporobolus heterolepis	NC_029417
Sporobolus maritimus	NC_027650
Sporobolus michauxianus	NC_029416
Stangeria eriopus	NC_026041
Staurastrum punctulatum	NC_008116
Stephania japonica	NC_029432
Stichococcus bacillaris	NC_025527
Stigeoclonium helveticum	NC_008372
Stipa hymenoides	NC_027464
Stipa lipskyi	NC_028444
Stipa purpurea	NC_029390
Stockwellia quadrifida	NC_022414
Strobus bungeana	NC_028421
Strobus gerardiana	NC_011154

Strobus koraiensis	NC_004677
Strobus krempfii	NC_011155
Strobus lambertiana	NC_011156
Strobus longaeva	NC_011157
Strobus monophylla	NC_011158
Strobus nelsonii	NC_011159
Strobus sibirica	NC_028552
Strobus strobus	NC_026302
Syagrus coronata	NC_029241
Syntrichia ruralis	NC_012052
Taiwania cryptomerioides	NC_016065
Taiwania flousiana	NC_021441
Takakia lepidozioides	NC_028738
Tanaecium tetragonolobum	NC_027955
Taxus mairei	NC_020321
Tectona grandis	NC_020098
Tetracentron sinense	NC_021425
Tetraphis pellucida	NC_024291
Tetraplodon fuegianus	NC_029305
Tetrastigma hemsleyanum	NC_029339
Thalictrum coreanum	NC_026103
Thamnocalamus spathiflorus	NC_024724
Theobroma cacao	NC_014676
Thysanolaena latifolia	NC_025238
Tilia amurensis	NC_028588
Tilia mandshurica	NC_028589
Tilia oliveri	NC_028590
Tilia paucicostata	NC_028591
Torreya fargesii	NC_029398
Torreyochloa pallida	NC_027486
Trachelium caeruleum	NC_010442
Treubaria triappendiculata	NC_028578
Trifolium aureum	NC_024035
Trifolium boissieri	NC_025743
Trifolium glanduliferum	NC_025744
Trifolium grandiflorum	NC_024034
Trifolium meduseum	NC_024166
Trifolium repens	NC_024036
Trifolium strictum	NC_025745

Trifolium subterraneum	NC_011828
Trigonobalanus doichangensis	NC_023959
Trillium cuneatum	NC_027185
Trillium decumbens	NC_027282
Trillium maculatum	NC_027738
Trillium tschonoskii	NC_027739
Trisetum cernuum	NC_027487
Trithuria inconspicua	NC_020372
Triticum aestivum	NC_002762
Triticum macha	NC_025955
Triticum monococcum	NC_021760
Triticum timopheevii	NC_024764
Triticum turgidum	NC_024814
Triticum urartu	NC_021762
Trochodendron aralioides	NC_021426
Tydemania expeditionis	NC_026796
Typha latifolia	NC_013823
Ulva fasciata	NC_029040
unclassified Trebouxiophyceae sp.	NC_018569
Utricularia gibba	NC_021449
Utricularia macrorhiza	NC_025653
Vaccinium macrocarpon	NC_019616
Vanilla planifolia	NC_026778
Veratrum patulum	NC_022715
Vicia sativa	NC_027155
Vigna angularis	NC_021091
Vigna radiata	NC_013843
Vigna unguiculata	NC_018051
Viola seoulensis	NC_026986
Viscum album	NC_028012
Viscum crassulae	NC_027959
Viscum minimum	NC_027829
Vitis aestivalis	NC_029454
Vitis rotundifolia	NC_023790
Vitis vinifera	NC_007957
Viviania marifolia	NC_023259
Watanabea reniformis	NC_025526
Welwitschia mirabilis	NC_010654
Wisteria floribunda	NC_027677

Wisteria sinensis	NC_029406
Wolffia australiana	NC_015899
Wolffiella lingulata	NC_015894
Wollemia nobilis	NC_027235
Woodwardia unigemmata	NC_028543
Xerophyllum tenax	NC_027158
Xylochloris irregularis	NC_025534
Yushania levigata	NC_024725
Zamia furfuracea	NC_026040
Zanthoxylum piperitum	NC_027939
Zea mays	NC_001666
Zingiber spectabile	NC_020363
Zizania aquatica	NC_026967
Zizania latifolia	NC_029401
Zoysia macrantha	NC_029418
Zygnema circumcarinatum	NC_008117

Table S2.3: Small dataset sampling

Species	Nuclear data type	Nuclear data source	Plastome GenBank accession
Acacia aulococarpa/Acacia ligulata	Transcriptome (*A. aulococarpa*)	PlanTransDB (website no longer active)	NC_026134 (*A. ligulata*)
Amborella trichopoda	Genome	Phytozome (10/4/16, v1.0)	NC_005086
Arabidopsis thaliana	Genome	Phytozome (10/4/16, TAIR10)	NC_000932
Cucumis sativus	Genome	Phytozome (10/4/16, v1.0)	DQ119058
Eucalyptus grandis	Genome	Phytozome (10/4/16, v2.0)	NC_014570
Geranium maderense	Transcriptome	PlanTransDB (website no longer active)	NC_029999
Glycine max	Genome	Phytozome (10/4/16, v1)	NC_007942
Gossypium raimondii	Genome	Phytozome (10/4/16, v2.1)	NC_016668
Lathyrus sativus	Transcriptome	Chapman 2015 Applications Plant Sci 3:1400111; Dryad: https://doi.org/10.5061/dryad.k9h76	NC_014063
Liriodendron chinense	Transcriptome	Yang et al. 2014 Gene, 534:155-162	NC_030504
Lobelia siphilitica	Transcriptome	1KP (IZLO)	KY354225
Medicago truncatula	Genome	Phytozome (10/4/16, v1)	NC_003119
Mimulus guttatus/Erythranthe lutea	Genome (*M. guttatus*)	Phytozome (10/4/16, v2.0)	KU705476 (*E. lutea*)
Musa acuminata	Genome	Phytozome (10/4/16, v1)	HF677508
Oenothera biennis	Transcriptome	1KP (MLUJ)	NC_010361
Oryza sativa (Japonica group)	Genome	Phytozome (10/4/16, v7.0)	NC_001320
Plantago maritima	Transcriptome	1KP (YKZB)	NC_028519
Populus trichocarpa	Genome	Phytozome (10/4/16, v3.0)	NC_009143
Prunus persica	Genome	Phytozome (10/4/16, v2.1)	NC_014697
Ricinus communis	Genome	Phytozome (10/4/16, v0.1)	NC_016736
Silene latifolia	Transcriptome	Sloan et al. 2014 Mol Biol Evol 31:673-682; NCBI SRA SRX353047	NC_016730
Silene noctiflora	Transcriptome	Sloan et al. 2014 Mol Biol Evol 31:673-682; NCBI SRA SRX353048	NC_016728
Solanum lycopersicum	Genome	Phytozome (10/4/16, v2.3)	NC_007898
Sorghum bicolor	Genome	Phytozome (10/4/16, v3.1)	NC_008602
Vitis vinifera	Genome	Phytozome (10/4/16)	NC_007957

Table S2.4: Rate variation species sampling

First subset	Second subset	
Slow background lineages	**Accelerated lineages**	**Slow lineages**
Acorus calamus	*Acacia ligulata*	*Agrostemma githago*
Akebia trifoliata	*Aquilaria sinensis*	*Allium cepa*
Alsophila spinulosa	*Berberis bealei*	*Amborella trichopoda*
Amborella trichopoda	*Brighamia insignis*	*Dorcoceras hygrometricum*
Ananas comosus	*Carex siderosticta* 1	*Elodea canadensis*
Anthoceros angustus	*Epimedium sagittatum*	*Eugenia uniflora*
Arabidopsis thaliana	*Erodium carvifolium*	*Francoa sonchifolia*
Buxus microphylla	*Eustrephus latifolius*	*Habenaria pantlingiana*
Chaetosphaeridium globosum	*Geranium palmatum*	*Lactuca sativa*
Coccomyxa subellipsoidea	*Hypseocharis bilobata*	*Lathraea squamaria*
Cycas taitungensis	*Jasminum nudiflorum*	*Leucaena trichandra*
Equisetum arvense	*Lathyrus pubescens*	*Musa textilis*
Geminella minor	*Lonicera japonica*	*Nuphar advena*
Ginkgo biloba	*Medicago truncatula*	*Olea europaea*
Habenaria pantlingiana	*Monsonia speciosa*	*Orobanche californica*
Huperzia lucidula	*Najas flexilis*	*Primula poissonii*
Illicium oligandrum	*Oenothera biennis*	*Schefflera delavayi*
Liriodendron tulipifera	*Orobanche crenata*	*Silene latifolia*
Lygodium japonicum	*Orobanche gracilis*	*Sinopodophyllum hexandrum*
Marchantia polymorpha	*Oryza sativa*	*Soja max*
Nelumbo lutea	*Pelargonium x hortorum 1*	*Theobroma cacao*
Nicotiana tabacum	*Pisum sativum*	*Typha latifolia*
Nuphar advena	*Plantago maritima*	
Physcomitrella patens	*Plantago media*	
Spinacia oleracea	*Ravenala madagascariensis*	
Tetracentron sinense	*Schwalbea americana*	
Vitis vinifera	*Silene chalcedonica1*	
	Silene conica	
	Silene noctiflora	
	Silene paradoxa	
	Tanaecium tetragonolobum	
	Trachelium caeruleum	
	Trifolium subterraneum	
	Trithuria inconspicua	
	Vaccinium macrocarpon	
	Vanilla planifolia	
	Viviania marifolia	
	Wisteria floribunda	

Table S2.5: Catalytic site gain/loss in ClpP and ClpR subunits across 25 angiosperm species

Subunit	Serine	Histidine	Aspartate
ClpP1	Replaced in *P. maritima*	Replaced in *P. maritima* and *S. noctiflora*	Replaced in *A. aulococarpa*, *O. biennis*, *P. maritima*, and *S. noctiflora*
ClpP3	Replaced in *P. maritima*	Replaced in *G. maderense*, *O. biennis*, and *P. maritima*	Replaced in *P. maritima*
ClpP4	Conserved across all species	Replaced in *L. sativus* and *M. truncatula*	Conserved across all species
ClpP5	Conserved across all species	Conserved across all species	Missing in *S. noctiflora*
ClpP6	Replaced in *L. siphilitica* and *P. maritima*	Replaced in *L. sativus*, *M. truncatula*, *P. maritima*, and *S. noctiflora*	Replaced in *P. maritima*
ClpR1		*S. noctiflora* and *S. bicolor* have regained a His in this position	
ClpR2	*G. maderense* and *G. max* have regained a Ser in this position		
ClpR3	*L. siphilitica* has regained a Ser in this position		
ClpR4			

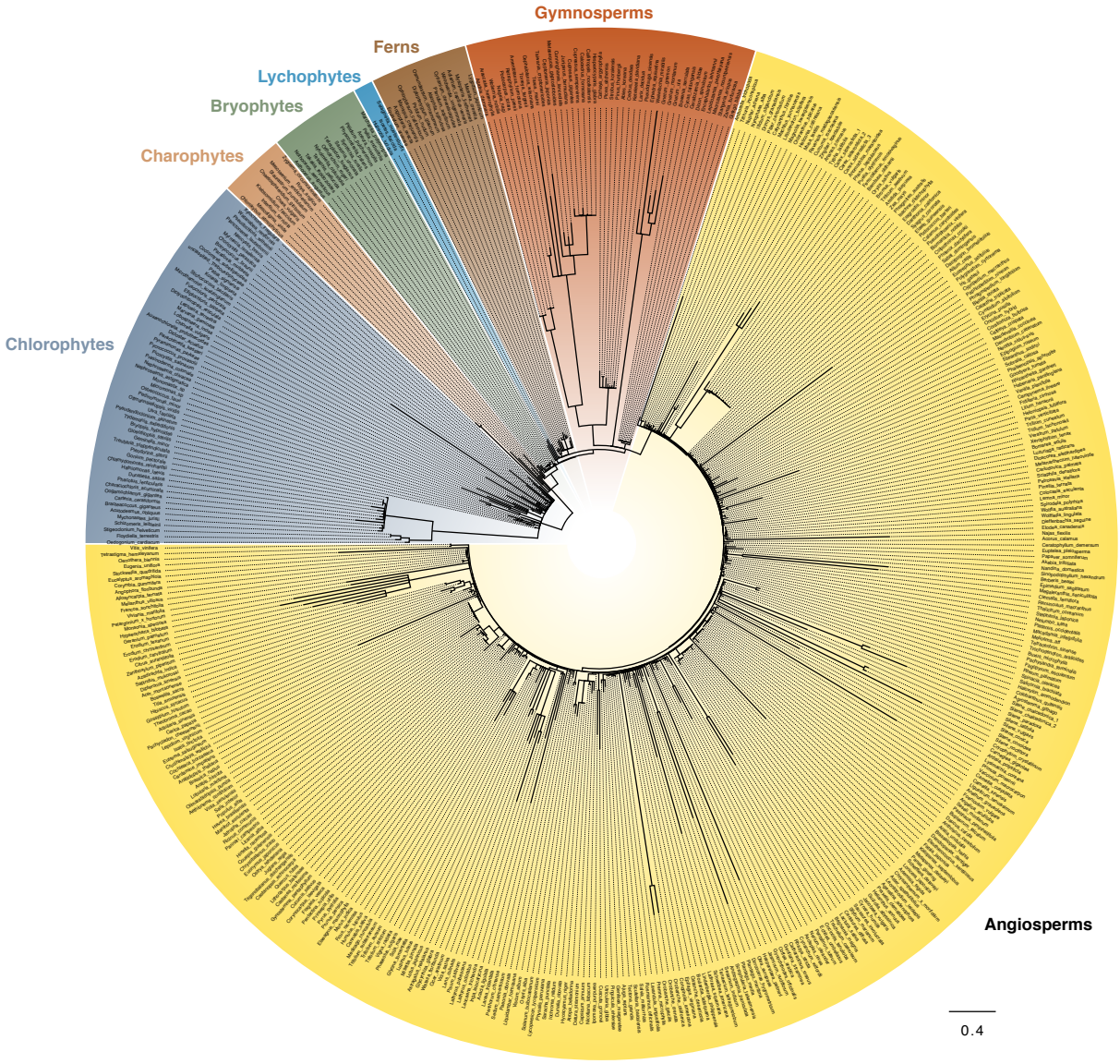


Figure S2.1: ClpP1 evolutionary rates across green plants. Tree is the same as shown in Figure 2 but includes species names for each branch. Branch length represents amino acid substitutions per site.

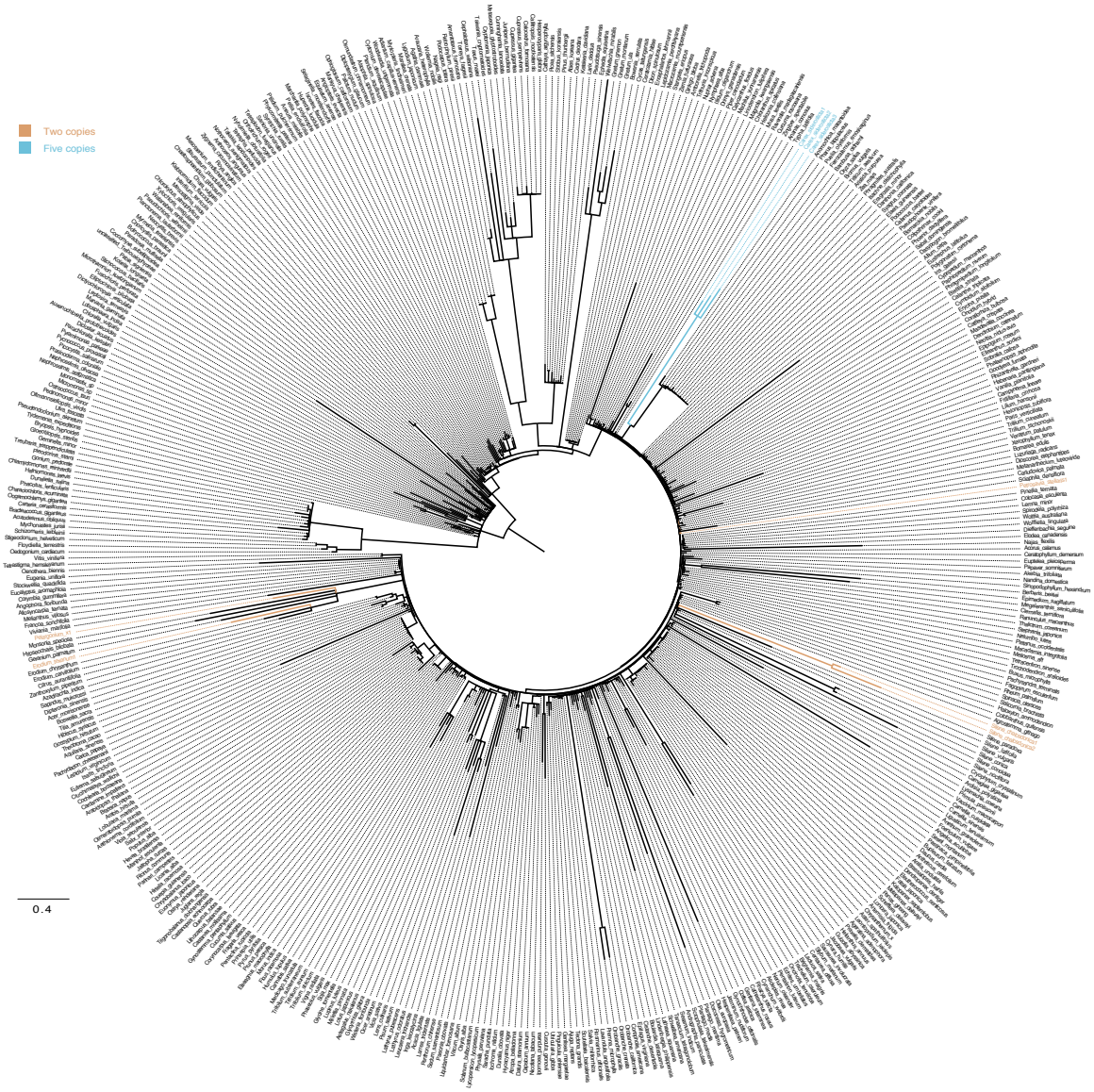


Figure S2.2: *clpP1* duplications across green plants superimposed on ClpP1 evolutionary rate tree. Gene duplicates with identical sequences were collapsed into a single branch. Internal branches are colored based on simple parsimony. Positive association between sequence divergence and gene duplication in angiosperms: $P = 0.006$, binaryPGLMM.

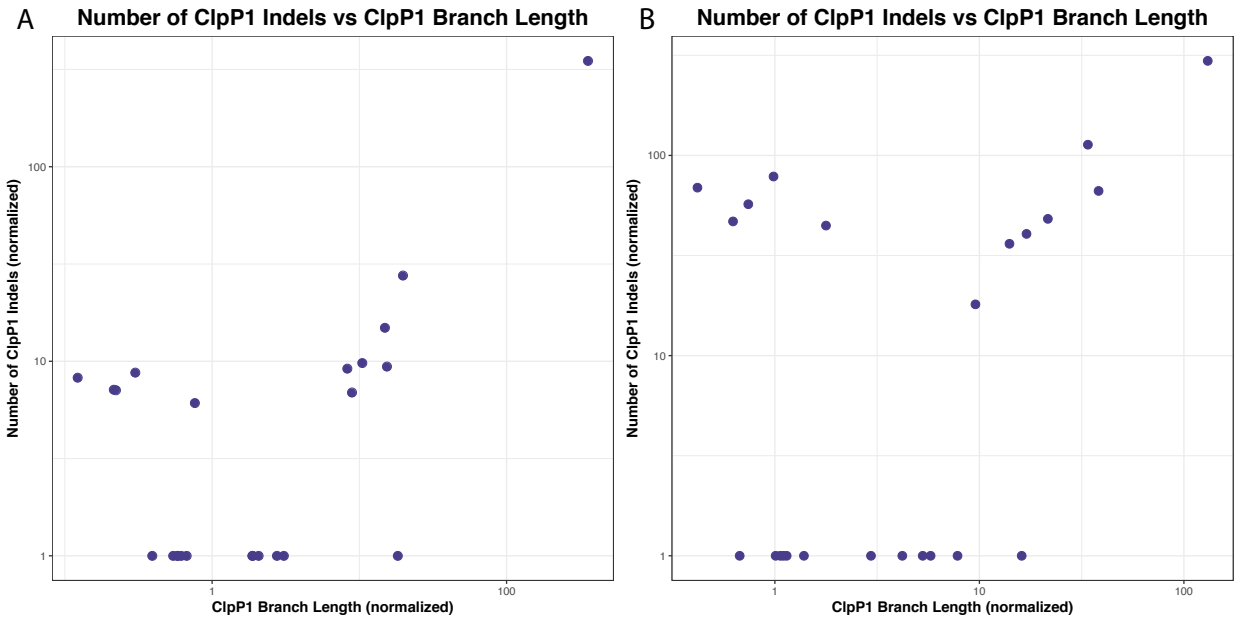


Figure S2.3: Correlation between ClpP1 indels and branch length (sequence divergence). Each point represents the number of *clpP1* indels divided by ClpP1 branch length (both normalized) for a particular species. Values of 0 were set to 1 for purposes of plotting on a logarithmic scale. A) Normalization of both axes was achieved using independent sets of non-Clp nuclear-encoded proteins (n=10, concatenated). Spearman's rho = 0.36, $P = 0.08$. B) Normalization of both axes was achieved using independent sets of plastid-encoded photosynthetic proteins (n=22 each, concatenated). Spearman's rho = 0.20, $P = 0.34$.

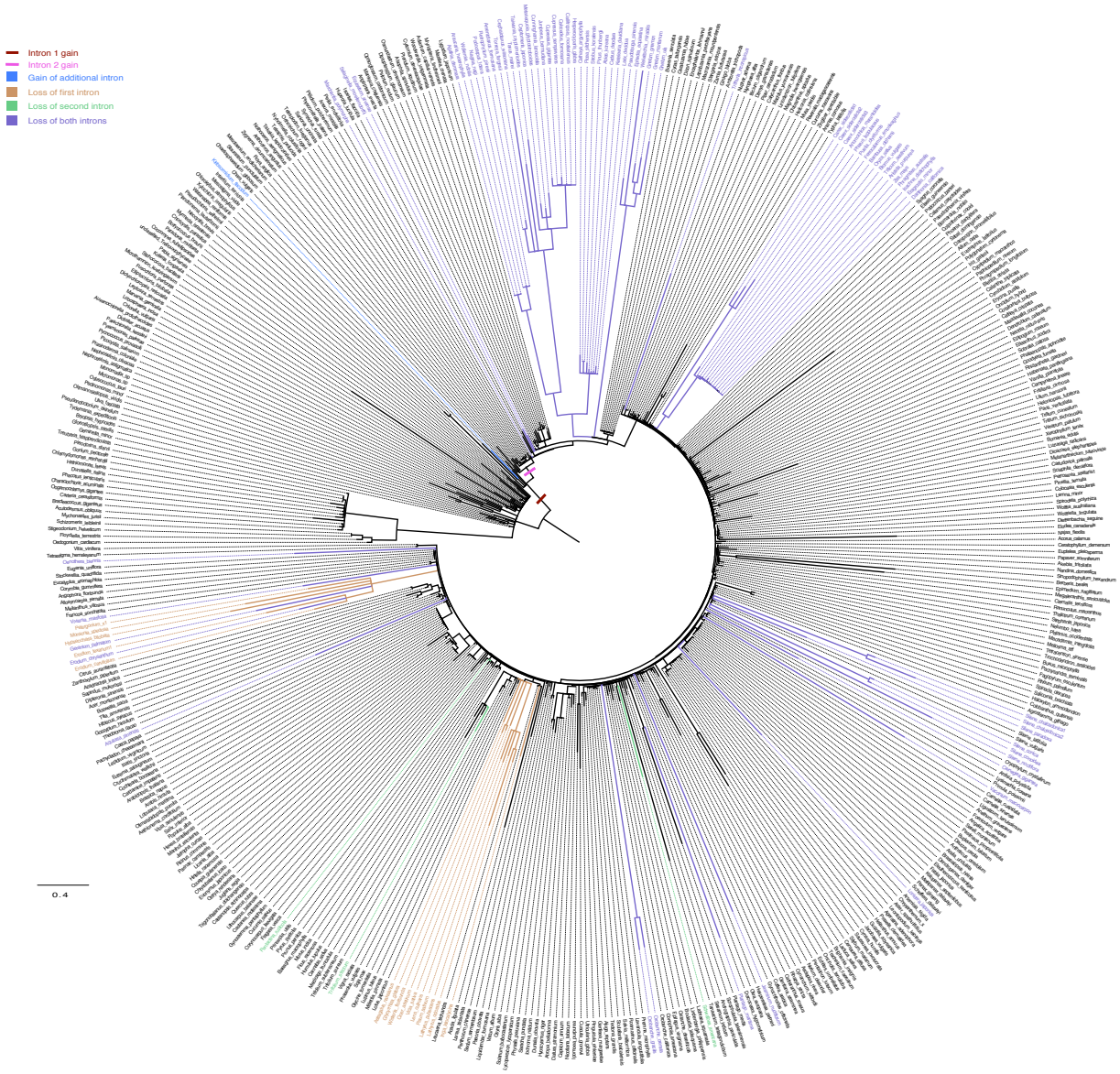


Figure S2.4: Evolutionary history of gains and subsequent losses of *clpP1* introns superimposed on ClpP1 evolutionary rate tree. Internal branches are colored based on simple parsimony. Positive association between sequence divergence and loss of introns in angiosperms: P = 2.7e-9 for intron 1 and P = 6.1e-9 for intron 2, binaryPGLMM.

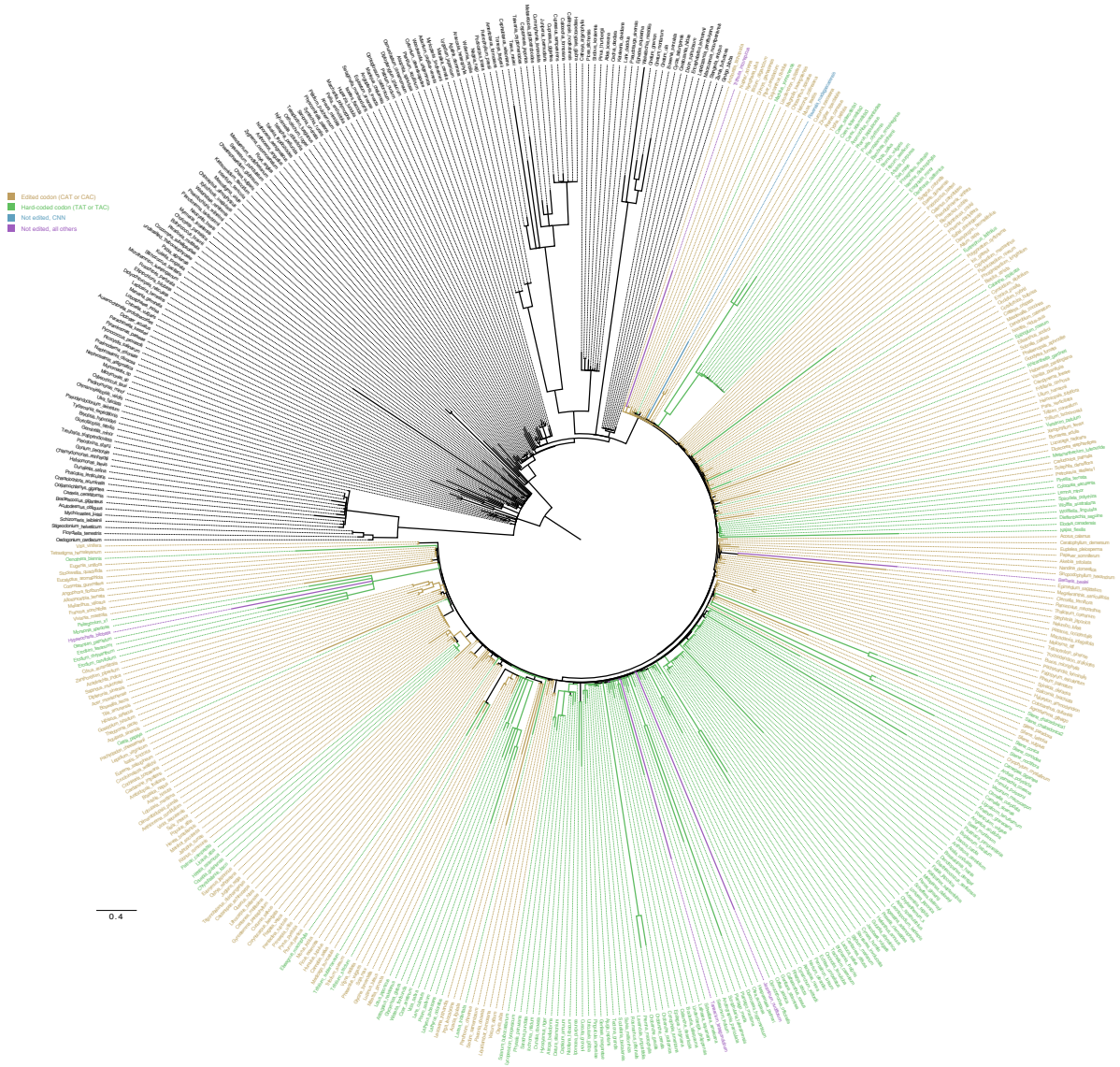


Figure S2.5: RNA editing of *clpP1* codon 187 superimposed on ClpP1 evolutionary rate tree. Internal branches are colored based on simple parsimony. Positive association between sequence divergence and loss of editing in angiosperms: P = 3.2e-6, binaryPGLMM.

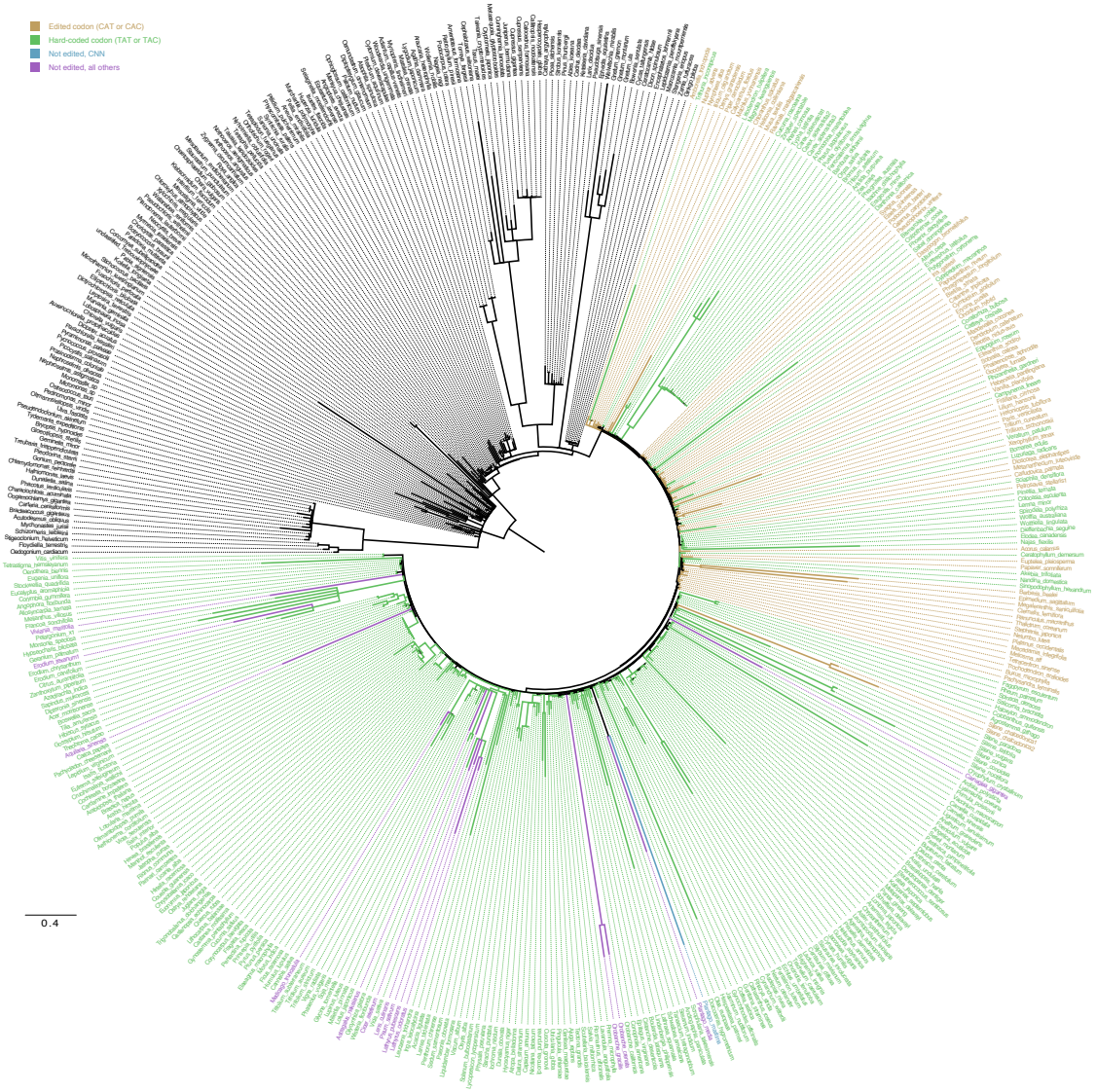


Figure S2.6: RNA editing of *clpP1* codon 28 superimposed on ClpP1 evolutionary rate tree. Internal branches are colored based on simple parsimony. Positive but non-significant association between sequence divergence and loss of editing in angiosperms: $P = 0.43$, binaryPGLMM.

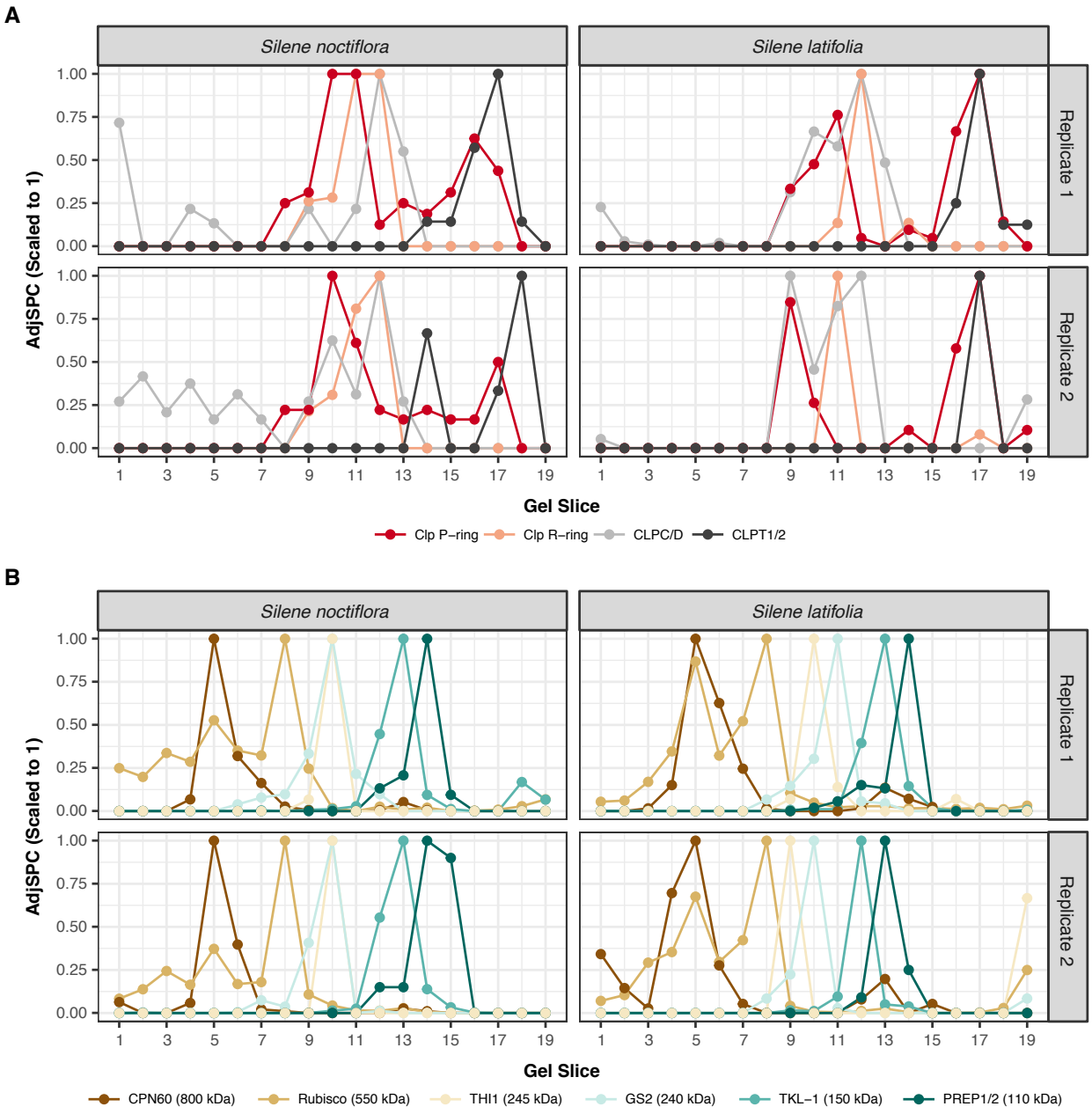


Figure S2.7: Size-fractionated mass spectrometry analysis of *Silene* plastid proteins. Gel slices correspond to native gel shown in Figure 4. A) Scaled AdjSPC summed across all subunits in different components of the Clp complex in *S. noctiflora* and *S. latifolia*. Values are scaled to 1 by dividing by the maximum observed value in any of the 19 gel slices. B) Control complexes used for internal calibration. Scaled AdjSPC values are calculated as in panel A.

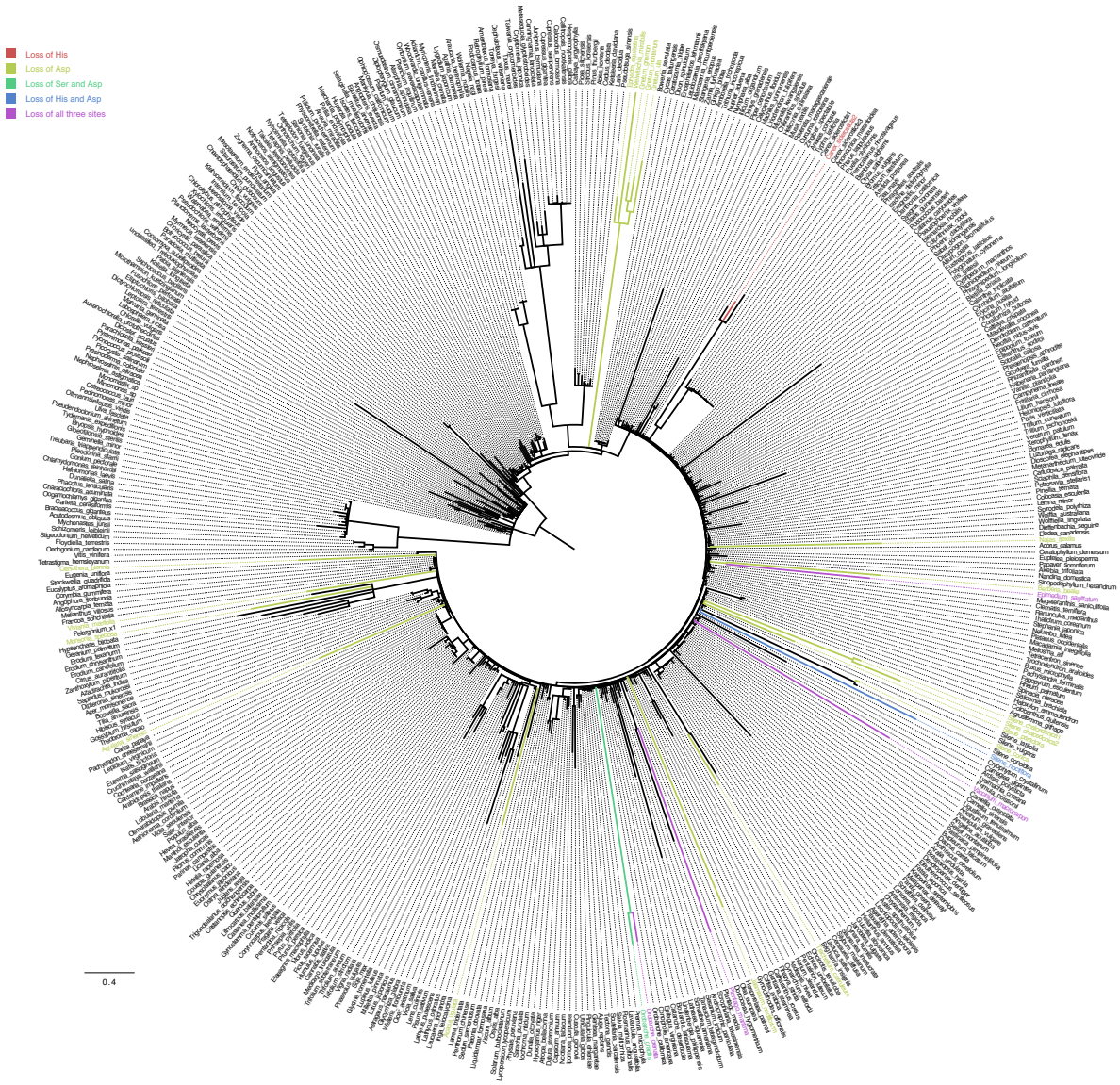


Figure S2.8: Loss of ClpP1 catalytic sites superimposed on ClpP1 evolutionary rate tree. Internal branches are colored based on simple parsimony. Positive association between sequence divergence and loss of catalytic residues in angiosperms: $P = 0.12$ for Ser, $P = 0.008$ for His and $P = 4.2e\text{-}7$ for Asp, binaryPGLMM.

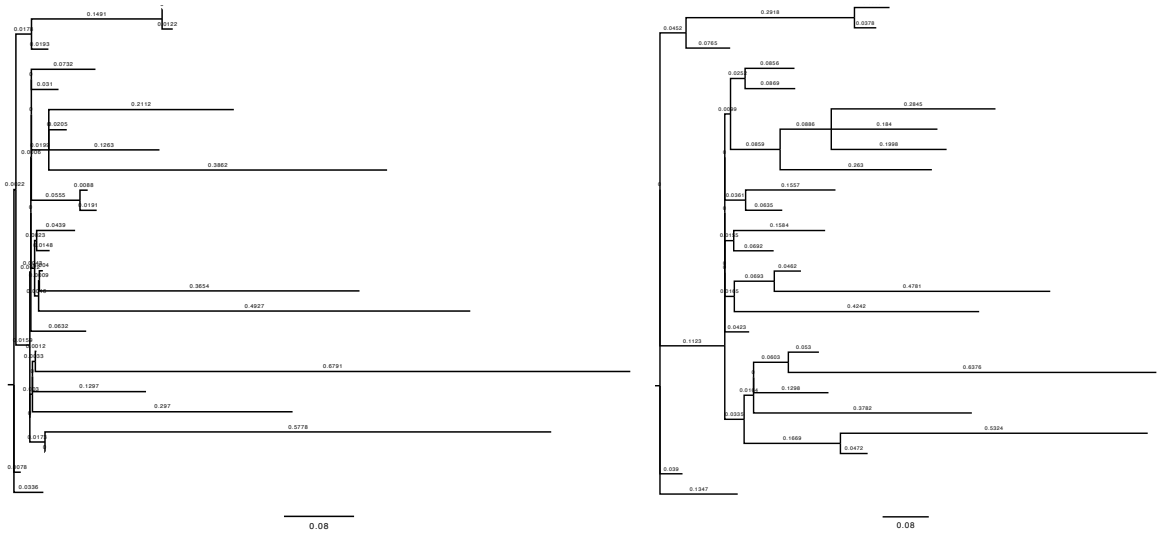


Figure S2.9: *clpP1 dN* (left) and *dS* (right) across 25 angiosperm species. Branch lengths represent either nonsynonymous (d_N) or synonymous (d_S) substitutions per site. Branch lengths are also reported as numerical values over each branch.

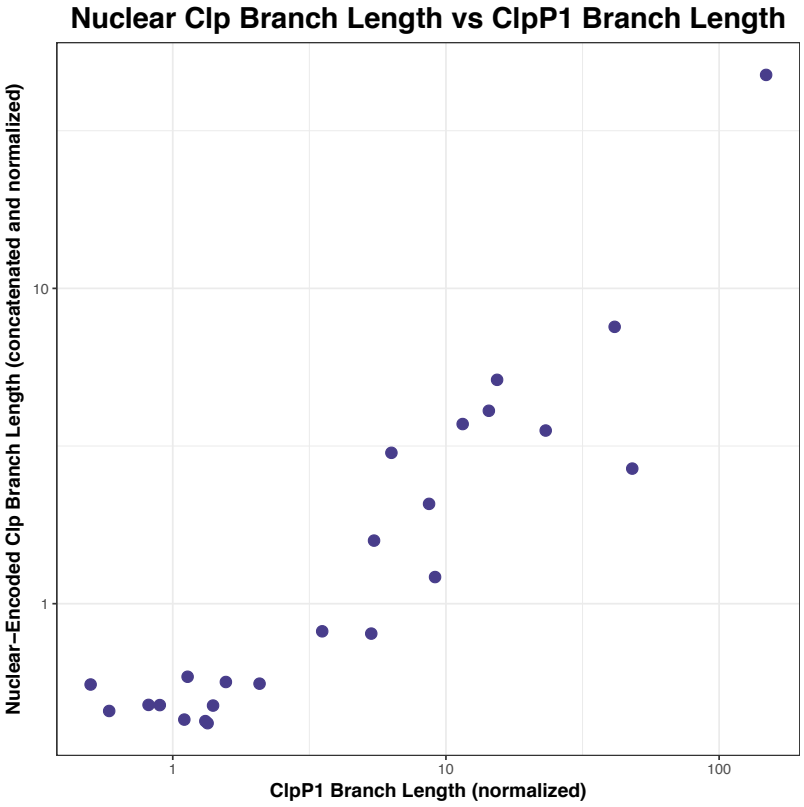


Figure S2.10: Scatterplot comparison of branch lengths (sequence divergence) of nuclear-encoded Clp proteins and ClpP1. Normalization of the x-axis was achieved using a set of non-Clp nuclear-encoded proteins (n=20, concatenated). Normalization of the y-axis was achieved using a set of plastid-encoded photosynthetic proteins (n=44, concatenated). $R^2 = 0.86$, $p << 0.001$.

Table S4.1: Sites inferred to be under positive selection in *ACC2* branches based on a branch-sites test in PAML using the trimmed alignment for *ACC*.

Site	Amino acid	Probability
6	L	1.000
92	G	0.965
125	T	0.998
153	V	0.967
182	L	0.983
200	V	0.982
212	L	0.951
333	E	0.993
387	E	0.952
425	E	0.999
428	S	0.982
429	L	0.999
475	S	1.000
480	R	0.968
533	T	0.968
539	S	0.991
546	V	0.977
562	V	0.989
597	L	0.994
652	L	0.999
660	H	0.982
663	M	0.995

687	R	0.961
693	H	0.996
699	L	0.991
700	G	1.000
713	F	0.999
715	A	0.973
739	L	0.995
740	N	0.999
744	S	0.993
753	Q	0.980
766	D	0.967
769	N	0.990
776	K	0.986
794	L	0.967
797	G	0.996
837	R	0.997
849	S	0.996
859	Q	0.986
865	R	1.000
868	L	0.999
872	K	0.995
922	T	0.979
948	Q	0.985
973	T	0.984
981	T	0.999

982	P	0.999
985	K	0.998
989	N	0.999
991	R	0.997
1015	P	0.999
1027	R	0.999
1042	Q	0.991
1043	W	1.000
1044	H	1.000
1045	R	0.995
1048	L	0.978
1078	E	0.995
1085	W	0.980
1096	L	0.984
1108	T	0.999
1110	H	0.958
1147	M	0.990
1151	Q	0.994
1160	Q	0.998
1161	E	0.957
1168	K	1.000
1177	S	0.999
1197	R	0.998
1200	M	0.986
1212	Y	0.989

1243	A	0.995
1326	A	0.963
1342	I	0.981
1344	R	0.983
1507	S	0.952
1580	K	0.956
1602	R	0.971
1655	S	0.966
1735	L	1.000
1887	V	0.984
1898	A	0.999
1904	Q	0.985
1928	E	0.982
2021	P	0.958
2027	S	0.976
2063	E	1.000
2131	K	0.998
2135	E	0.962
2137	A	0.998
2164	G	0.965
2194	E	1.000
2230	P	1.000
2235	Q	0.992
2241	R	0.985
2245	G	0.976

Table S4.2: Loss of catalytic sites and truncation of nuclear-encoded plastid ClpP core subunits

Species	Protein	Serine	Histidine	Aspartate	Length
Eucalyptus grandis	ClpP3		Replaced with R		
Musa acuminata	ClpP3		Replaced with R		
Plantago maritima 1	ClpP3	Replaced with Y	Replaced with R	Replaced with N	
Plantago maritima 2	ClpP3	Replaced with Y	Replaced with R	Replaced with N	
Lathyrus sativus	ClpP4		Replaced with T		
Medicago truncatula	ClpP4		Replaced with A		
Musa acuminata 2	ClpP4	Gap	Replaced with R	Gap	Truncated
Plantago maritima 2	ClpP4			Gap	
Populus trichocarpa 2	ClpP4	Replaced with M	Gap	Gap	Truncated
Gossypium raimondii 3	ClpP5	Gap			
Gossypium raimondii 4	ClpP5				Truncated
Gossypium raimondii 5	ClpP5	Replaced with N			Truncated

Gossipium Raimondii 6	ClpP5	Replaced with N		Replaced with N	Truncated
Gossypium raimondii 7	ClpP5	Replaced with N		Replaced with F	
Plantago maritima	ClpP5	Gap	Gap	Gap	Truncated
Lathyrus sativus	ClpP6		Replaced with G		
Lobelia siphilitica	ClpP6	Replaced with N			
Medicago truncatula	ClpP6		Replaced with N		
Plantago maritima	ClpP6	Replaced with G	Replaced with E	Replaced with F	
Populus trichocarpa 2	ClpP6				Truncated

*Empty cell indicates presence of catalytic site or full length.

Table S4.3: Truncation of nuclear-encoded plastid ClpR core subunits

Truncated ClpR1 Subunits
Vitis vinifera 2
Populus trichocarpa 2
Musa acuminata 2 (internal stop codon)
Mimulus guttatus 2
Truncated ClpR2 Subunits
Plantago maritima 2
Truncated ClpR3 Subunits
Populus trichocarpa 1
Populus trichocarpa 2
Truncated ClpR4 Subunits
Eucalyptus grandis 2
Medicago truncatula 2
Vitis vinifera 2
Vitis vinifera 3
Vitis vinifera 4
Musa acuminata 3
Musa acuminata 4

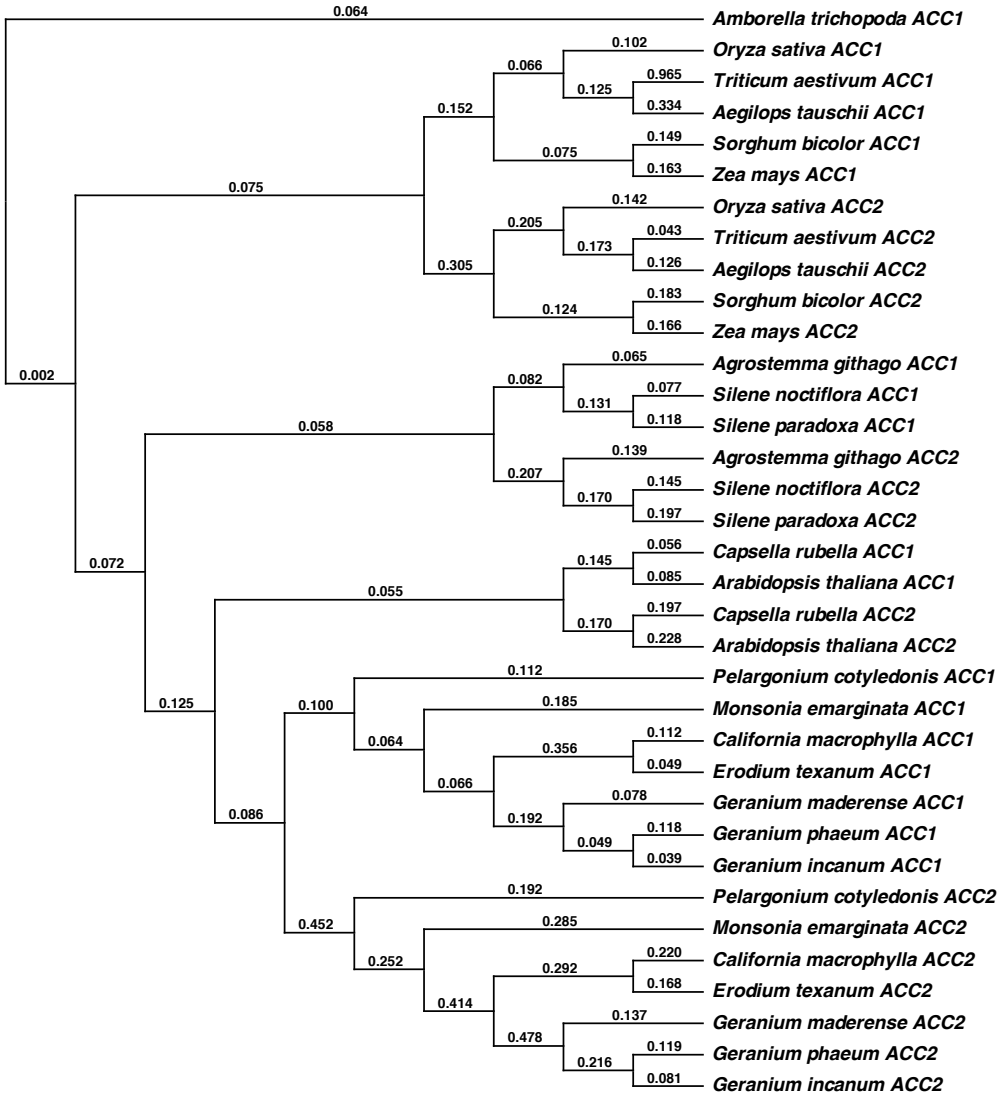


Figure S4.1. *ACC* tree. Branch labels are d_N/d_S values. *ACC1* represents cytosolic-targeted genes while *ACC2* represents plastid-targeted genes.

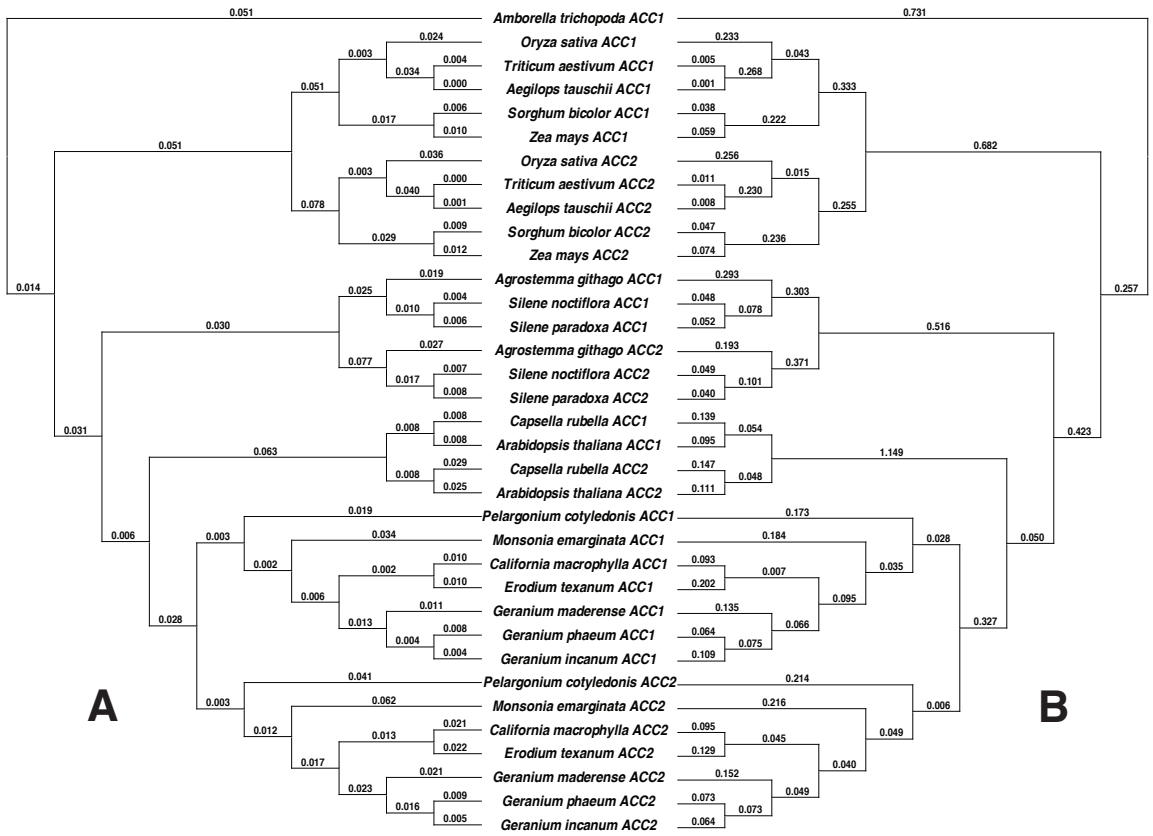


Figure S4.2. *ACC* tree. A) Branch labels are d_N values. B) Branch labels are d_S values.

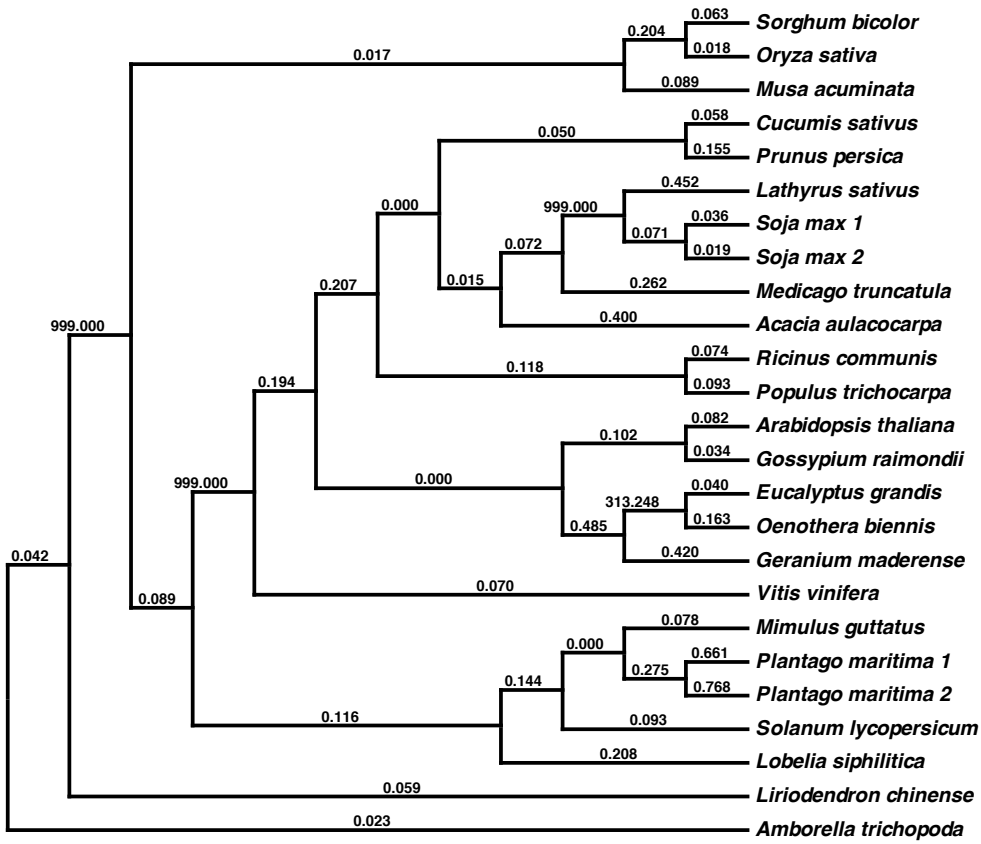


Figure S4.3. *CLPP3* tree. Branch labels are d_N/d_S values.

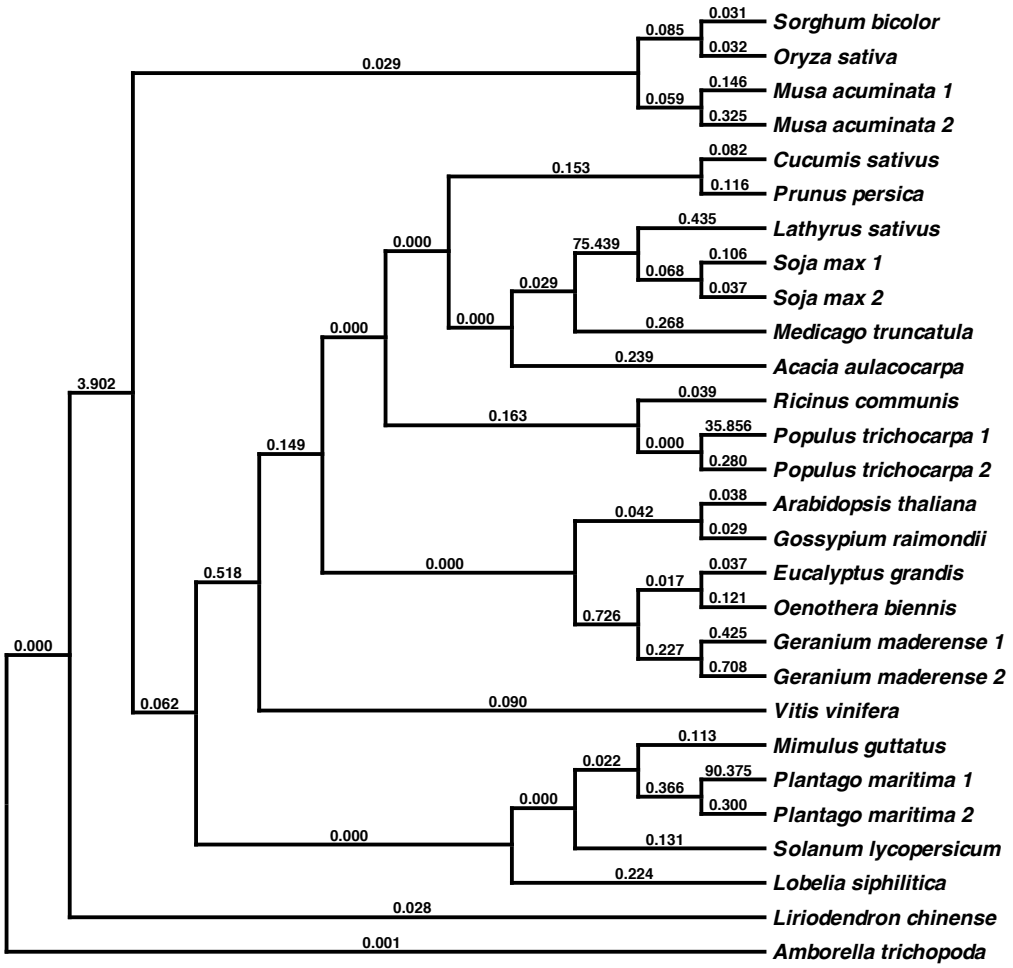


Figure S4.4. *CLPP4* tree. Branch labels are d_N/d_S values.

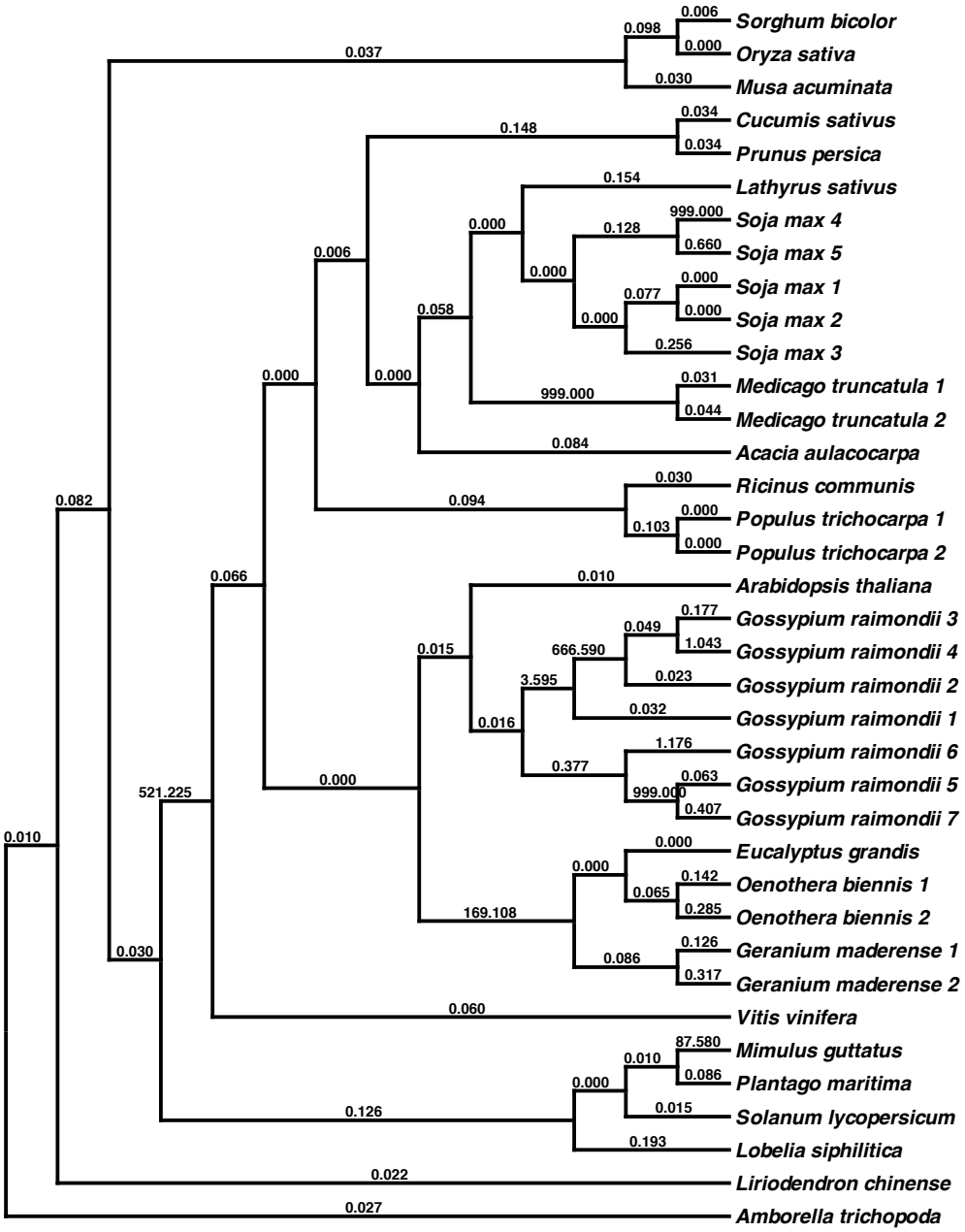


Figure S4.5. *CLPP5* tree. Branch labels are d_N/d_S values.

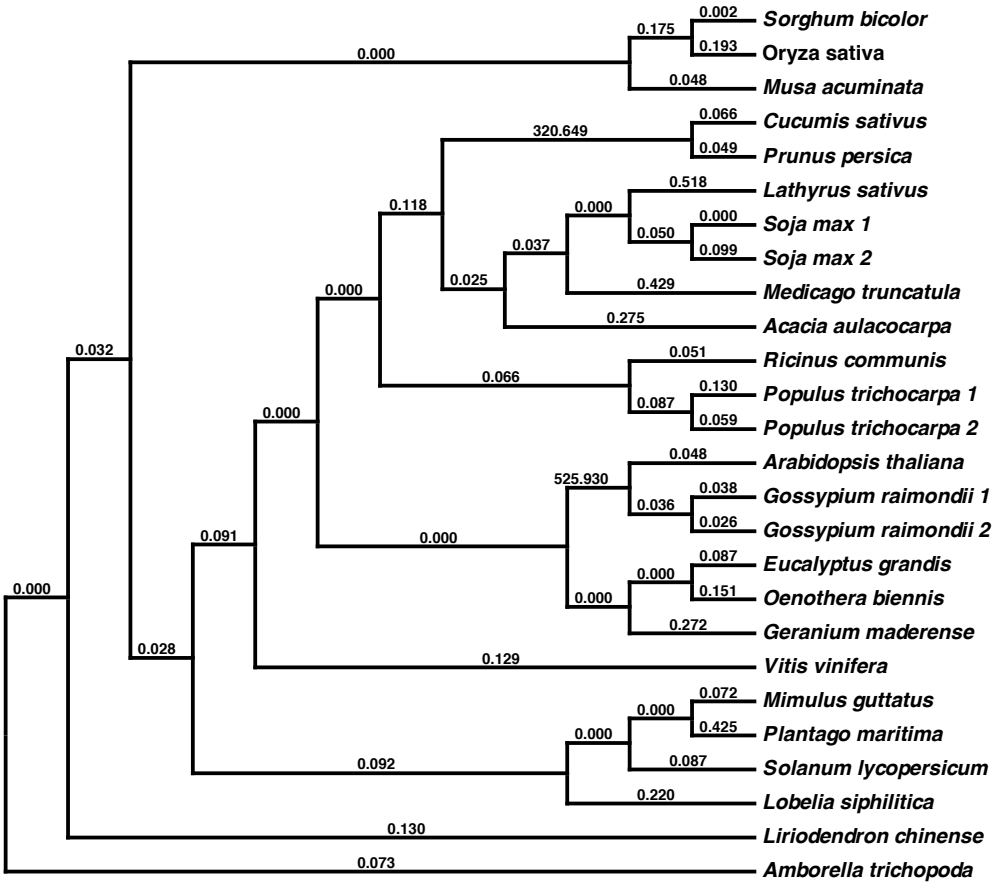


Figure S4.6. *CLPP6* tree. Branch labels are d_N/d_S values.

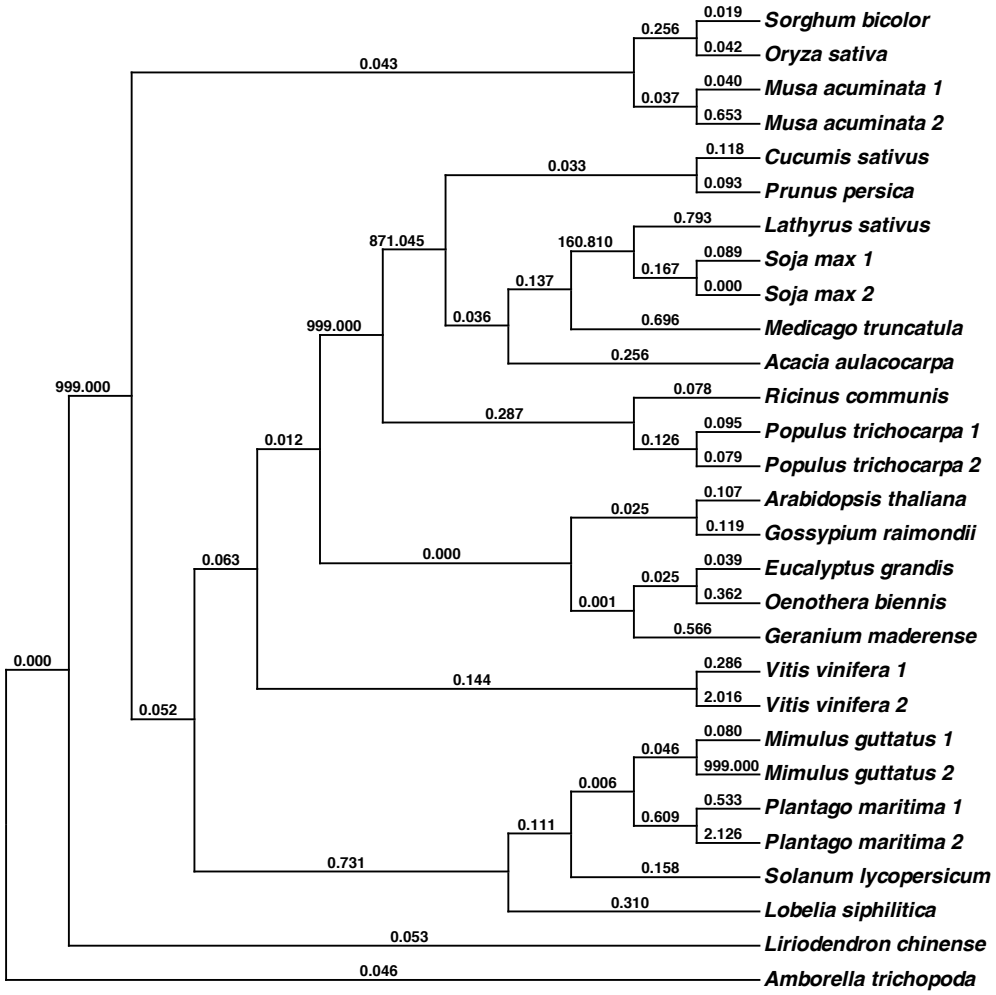


Figure S4.7. *CLPR1* tree. Branch labels are d_N/d_S values.

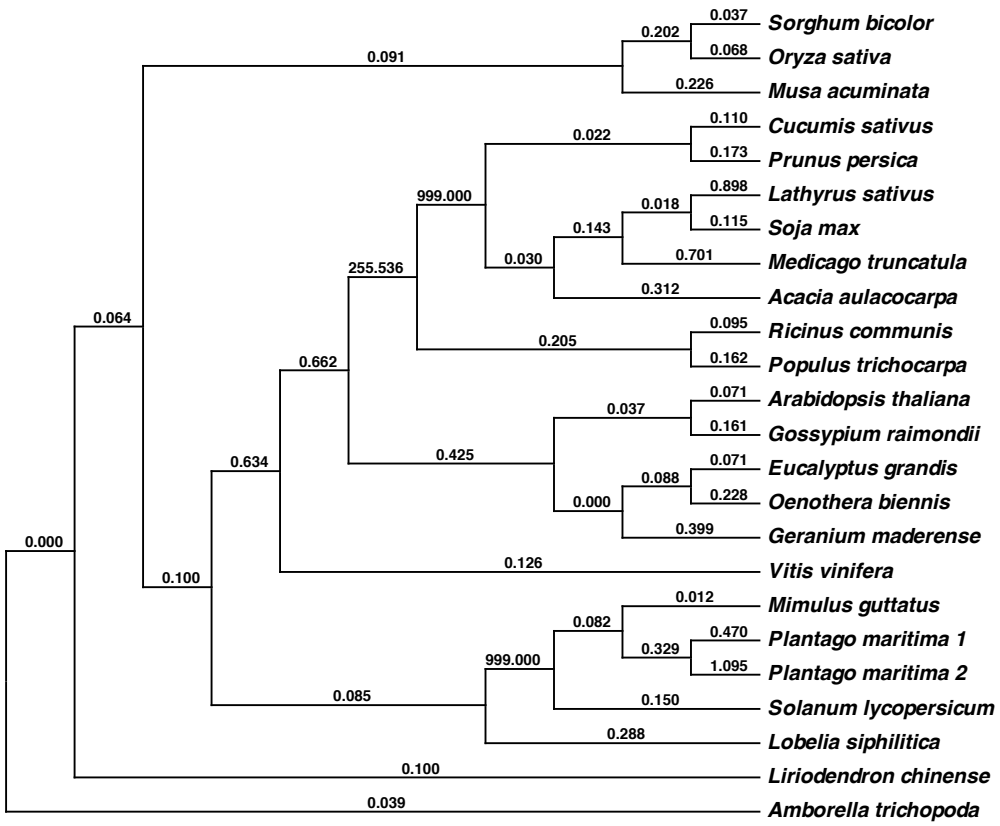


Figure S4.8. *CLPR2* tree. Branch labels are d_N/d_S values.

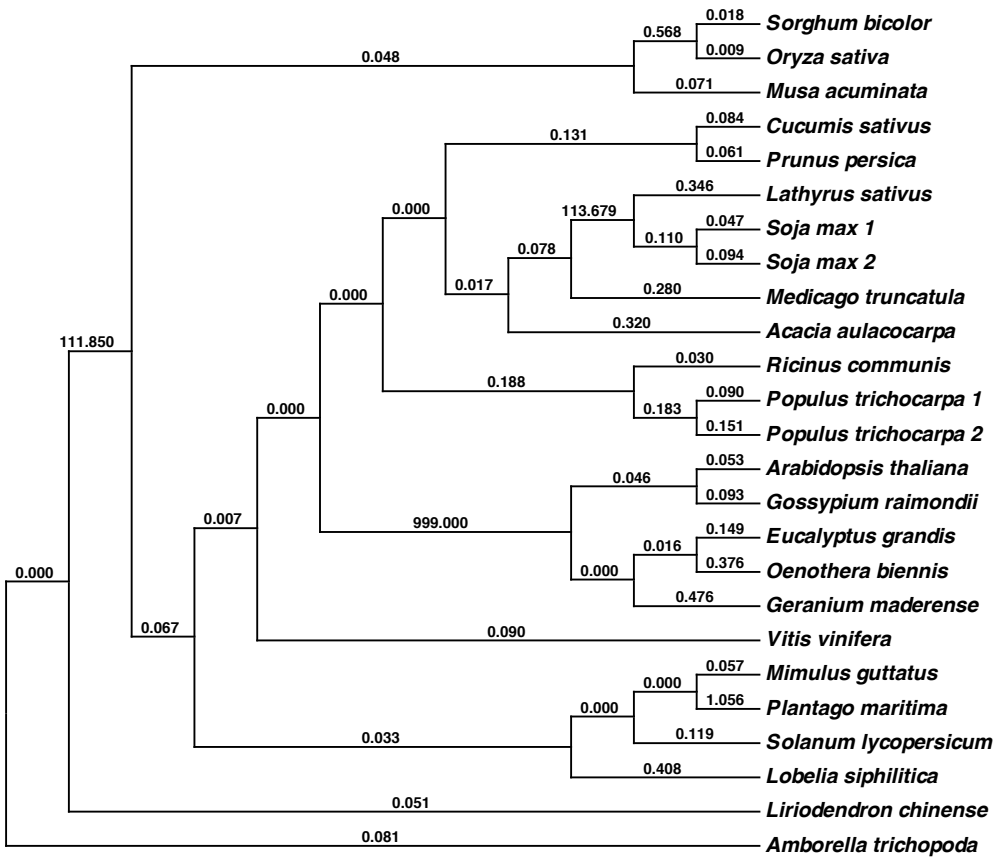


Figure S4.9. *CLPR3* tree. Branch labels are d_N/d_S values.

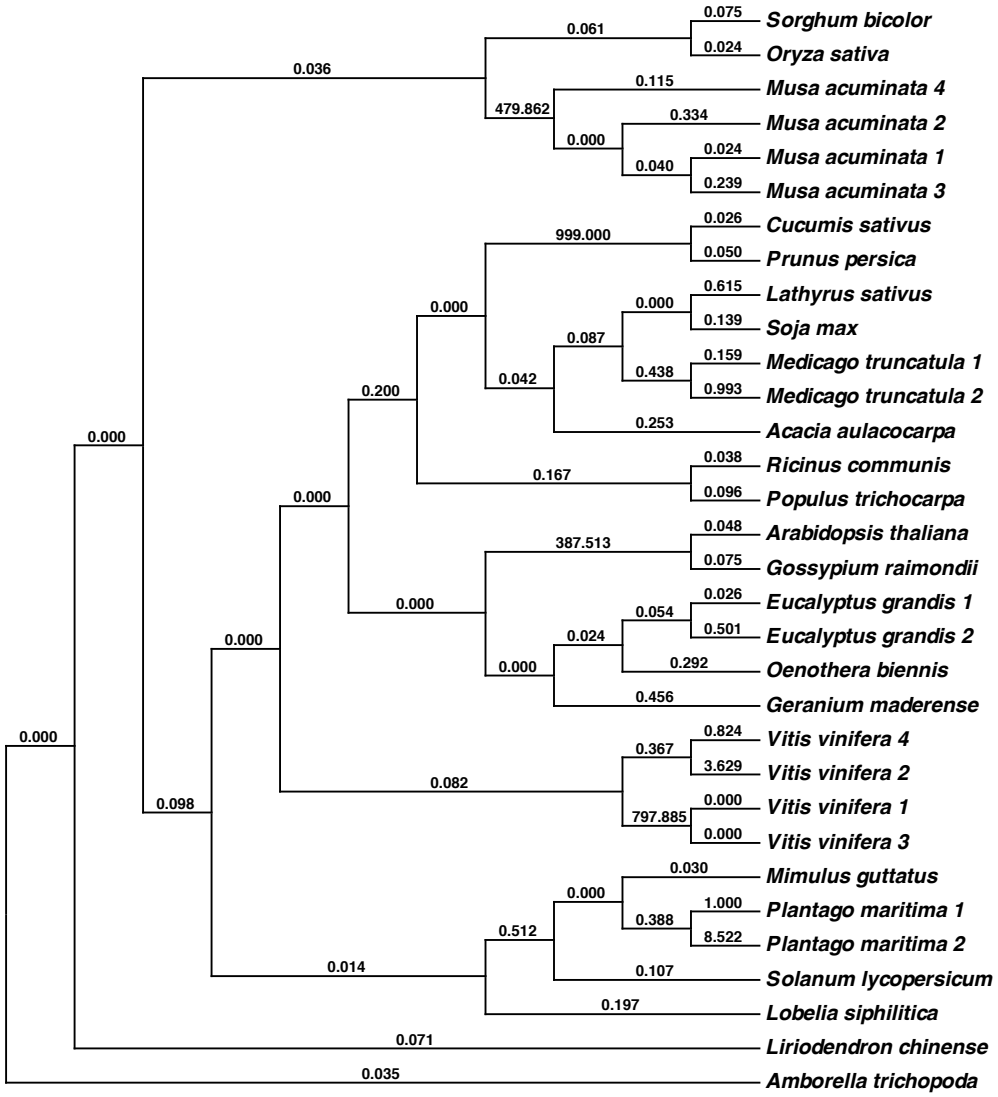


Figure S4.10. *CLPR4* tree. Branch labels are d_N/d_S values.