Cytonuclear incompatibility contributes to incipient speciation

Karen Beth Barnard-Kubow
Atlanta, GA

B.S., University of Southern California, 2005
M.S., ETH Zürich, 2008

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

Department of Biology

University of Virginia
May, 2015

Abstract

Understanding how new species form is one of the central goals of evolutionary biology. While historically studies have focused on the role of nuclear genetic incompatibilities in speciation, cytonuclear incompatibilities have been proposed to be among the first genetic incompatibilities to arise, influencing the early stages of the speciation process. However, due to the lack of studies characterizing reproductive isolation at the intraspecific level, little is known regarding the role of cytonuclear incompatibilities at these early stages. The effect of cytonuclear incompatibility could be mitigated though biparental plastid inheritance, by increasing the likelihood of individuals inheriting a compatible chloroplast. Alternatively, accelerated plastid evolution could speed up the co-evolution of the plastid and nuclear genomes, increasing the propensity for cytonuclear incompatibility when crossing between populations. This dissertation uses *Campanulastrum americanum* as a system to study how biparental plastid inheritance and accelerated plastid evolution may impact the evolutionary dynamics of cytonuclear incompatibilities and their contribution to incipient speciation. Analysis of chloroplast and nuclear markers distinguished four geographically structured, genetically divergent lineages in *C. americanum*. Crossing studies found substantial reductions in germination and survival when crossing between these lineages. There was also evidence for cytonuclear incompatibility underlying reproductive isolation in survival, supporting the idea that cytonuclear incompatibilities arise early in the speciation process. However, reproductive isolation between divergent lineages appeared to be reduced in the Appalachians where these lineages are likely in secondary contact. In addition, analysis of plastid inheritance patterns found that *C. americanum* has biparental plastid

inheritance, which in genetically-divergent crosses mitigates the impact of cytonuclear incompatibility on reproductive isolation, potentially slowing the speciation process. In contrast, analysis of plastid sequence divergence and polymorphism in *C. americanum* found evidence for increased nucleotide substitution rates acting at the among and within species level, suggesting that accelerated plastid evolution may facilitate the development of cytonuclear incompatibility in this species. Altogether, cytonuclear incompatibility appears to be contributing to incipient speciation in *C. americanum*, but the dynamics of this contribution are being influenced, often in opposing ways, by a variety of mechanisms, including gene flow, biparental plastid inheritance and accelerated plastid evolution.

Acknowledgements

Science is never done in isolation, and there are many people that contributed to the work in this dissertation in a variety of ways. First, I would like to thank my advisor, Laura Galloway. From the first meeting I had with Laura we have always had a great relationship. She has been helpful to me in so many ways throughout my time in graduate school. From helping me with my writing, to acting as a constant source of encouragement and support, she has been essential in shaping my growth as a scientist. I also have a tendency to try to take on more work than I can handle. Laura was instrumental in keeping me focused and on track and helping me to identify which of my ideas were worth following up on, versus those that needed to be saved for some time in the future. Janis Antonovics, my first reader, was always available to talk when I needed feedback on ideas and I greatly appreciate the many hours he spent with me in scientific discussion. He helped me to formulate my ideas and his influence has shaped my dissertation work. Doug Taylor played a similar role early on in my time in graduate school. My discussions with him and his insight helped to formulate my ideas as I was writing up my dissertation proposal. Ben Blackman has also offered much advice, particularly in the area of genomics. Butch Brodie and Dave Carr also gave much appreciated advice on my chapters. I have appreciated having their feedback during my committee meetings, and particularly on the dissertation itself. All of my committee members have given me valuable advice and helped me to see my work in new ways.

I would also like to thank the members of the Galloway lab who overlapped me with me, both past and present. Can Ashley Dai, Catherine Debban, Francis Kilkenny, Holly Prendeville, and Brittany Sutherland provided emotional support and friendship, as

well as giving feedback on my research. When I first started in the program at UVa, I rotated in the Taylor lab working with Dan Sloan, who was a graduate student at the time. In many ways Dan acted as another committee member for me, providing guidance in my research as well as much needed encouragement and support. I could always count on his help, even when he graduated and moved on to a postdoctoral position, and then a faculty position. I also want to thank Deb Triant for helping me with all things related to bioinformatics. Corlett Wood and Hilary Edgington were part of a support group with me for those of us trying to write our dissertations and graduate May 2015. I would not have made it though the past several months without them. Carolyn Beans started in graduate school the same year I did, and her friendship has been invaluable over the years.

The research presented in this dissertation would not have been possible without the many undergraduates who did research with me in the lab: Connor Johnson, Morgan McCoy, Mohan Nagaraja, Shay Nimjareansuk, Tim Park, Nina So, Elissa Trieu, Allison Gaynor, Lauren Wilson, and Remington Wong. In addition to helping me with my research, their independent projects often provided interesting results that were instrumental in shaping my future research directions and ideas. I also want to thank the various technicians and student workers in the Galloway lab that helped me with collecting data: Erin Arnold, Prajakta Bhayade, Anna Greenlee, Karla Platzer, Jie Ren, and Victoria Soler.

I am also indebted to the many other graduate students in my program who gave me feedback on my research and manuscripts, but also helped me out in so many other ways in terms of friendship and support. These include Jessie Abbate, Melissa Aikens, Malcolm Augat, Lelena Avila, Andrea Beradi, Peter Fields, Brian Sanderson, Megan

Sebasky, and Ray Watson. I would also like to thank Wendy Crannage for helping to keep my plants alive in the greenhouse, and AnhThu Nguyen for help with genomics work. Finally, there are many other faculty and various members of the Biology department who have assisted me with support, feedback, and advice: Joanne Chaplin, John Chuckalovcak, Nikki Forrester, Colleen Ingram, AnhThu Nguyen, Deborah Roach, Henry Wilbur, and Martin Wu. Janet Stevens, a faculty at near by Sweet Briar College was also helpful to me in many ways. Finally, the staff in the departmental office also helped me in many ways throughout the last six years.

I would also like to thank the Department of Biology at UVa for their financial support in terms of a dissertation year fellowship. I was also a generous recipient of an ARCS fellowship, which was instrumental in allowing me to have the time and support to obtain training outside of UVa, mentor undergraduates in doing research, and complete my research. This project also benefited from support by the National Science Foundation and Sigma Xi.

Outside of the scientific community, I would like to thank my church family at Meadows Presbyterian Church, particularly the current and past members of the youth group. I greatly appreciate the encouragement and support of my church family. I have also really appreciated the support of my family, particularly my parents, siblings, and grandparents. Most of all, I want to thank my husband, Kris Kubow, for all his help over the years. I would never have made it through graduate school without him. Finally, I want to thank my daughter Amy for helping me keep things in perspective and forcing me to improve my work life balance.

Table of Contents

Introduction

Speciation is often defined as the development of reproductive isolation (Mayr, 1942), and therefore characterizing the barriers to gene flow that result in reproductive isolation is essential for understanding this process by which new species form. Most studies of speciation have focused on characterizing reproductive isolation between closely related species (Scopece *et al.*, 2010). However, this approach can make it difficult to parse out the barriers to gene flow that led to speciation from those barriers that have arisen since (Coyne & Orr, 2004). In order to examine what types of barriers to gene flow arise early in the speciation process, there is a need for studies examining reproductive isolation within a species, or incipient speciation.

Reproductive isolation within a species is likely to be correlated with patterns of genetic divergence, which in turn is strongly influenced by historical processes. For example, glacial cycles have led to cyclic expansions and contractions of many species ranges (Hewitt, 2000). As a result of these range shifts, populations of a given species were often historically isolated in allopatric glacial refugia, which along with recurrent bottlenecks during subsequent recolonization, likely contributed to drift and genetic divergence among populations (Hewitt, 1996; Excoffier *et al.*, 2009). These processes facilitate the development of genetic incompatibility leading to reproductive isolation and speciation (Hewitt, 1996; Avise *et al.*, 1998; Carstens & Knowles, 2007). Therefore, understanding the genetic structure of a species is often a first step to gaining insight into the processes leading to barriers to gene flow and reproductive isolation.

However, the same biogeographical processes that promote genetic divergence can also result in bringing genetically divergent lineages back into contact before reproductive isolation is complete. When this happens, a central question is whether secondary contact and gene flow leads to a breakdown of this initial divergence, reducing reproductive isolation (Seehausen, 2006; Taylor *et al.*, 2006), or whether barriers to gene flow persist or even strengthen, eventually leading to speciation (e.g. reinforcement, Servedio & Noor, 2003).

One of the primary factors that it thought to create permanent barriers to gene flow is genetic incompatibility. Historically studies have focused on incompatibilities between nuclear loci, but recently there has been a growing realization of the potential importance of cytonuclear incompatibilities (negative interactions between organelle and nuclear genes) for the speciation process. Cytonuclear incompatibilities have been found to contribute to reproductive isolation and speciation in plants (e.g. Levy 1991; Fishman & Willis, 2006; Sambatti *et al.*, 2008; Greiner *et al.*, 2011; Leppala & Savolainen, 2011), yeast (reviewed in Chou & Leu, 2010), and animals (reviewed in Burton *et al.*, 2013). The widespread occurrence of asymmetrical fitness in reciprocal hybrids between plant species provides additional evidence for a pervasive role of cytonuclear incompatibility in speciation (Tiffin *et al.*, 2001; Turelli & Moyle, 2007). Asymmetrical reproductive isolation is often considered a hallmark of cytonuclear incompatibility, as reciprocal hybrids contain the same contributions of nuclear DNA from both parents, but differ in their cytoplasmic genomes.

Cytonuclear incompatibilities are likely to contribute to reproductive isolation and the speciation process due to several factors. For example, the nuclear and organelle

genomes must closely cooperate to carry out essential processes such as respiration and photosynthesis (Rand *et al.*, 2004). This cooperation raises the potential for co-adaptation and co-evolution between the two genomes, increasing the likelihood of genetic incompatibilities when the genomes are mismatched between populations (e.g. Ellison & Burton, 2008). On the other hand, there is also the potential for conflict between the nuclear and organelle genomes (e..g. cytoplasmic male sterility, Chase, 2007). In addition, organelle genomes are likely to accumulate deleterious mutations more rapidly due to their smaller effective population size (as they are often effectively haploid and uniparentally inherited) and their lack of recombination with genetically distinct partners (Birky, 2001). Due to these factors, cytonuclear incompatibilities have been proposed to be among the first genetic incompatibilities to arise (Levin, 2003; Greiner *et al.*, 2011; Burton & Barreto, 2012; Burton *et al.*, 2013; Greiner & Bock, 2013). However, few studies have focused on characterizing genetic incompatibilities at the intraspecific level, so there is relatively little evidence of cytonuclear incompatibilities being important at these early stages. Finally, the theory of cytonuclear incompatibilities arising early in the speciation process is based, in part, on the premise of uniparental organelle inheritance, which is not uniform across eukaryotes (Barr *et al.*, 2005).

In angiosperms, biparental inheritance of the chloroplast has arisen multiple times. Approximately 20% of angiosperms are estimated to have the potential for biparental inheritance of the chloroplast, defined as the presence of plastid DNA in the pollen generative cells (Corriveau & Coleman, 1988; Zhang *et al.*, 2003). There has been much discussion regarding the reasons uniparental organelle inheritance appears to be under positive selection (reviewed in Mogensen, 1996; Birky, 2001). However less is

known as to why biparental inheritance may be favored in some taxa. In regards to the plastid genome, the theory most often proposed is that biparental inheritance has evolved as a mechanism to overcome incompatible plastids in species that exhibit cytonuclear incompatibility (Hu *et al.*, 2008; Zhang & Sodmergen, 2010; Jansen & Ruhlman, 2012). The fact that biparental plastid inheritance often occurs in species with cytonuclear incompatibility supports this theory (reviewed in Jansen & Ruhlman, 2012). In addition, plastid-nuclear compatibility has been found to influence patterns of plastid transmission in *Oenothera* (Chiu & Sears, 1993). However, this theory has rarely been experimentally tested.

Biparental inheritance has the potential to influence the evolutionary dynamics of cytonuclear incompatibilities and their contribution to reproductive isolation and speciation. The occurrence of biparental plastid inheritance in crosses that exhibit cytonuclear incompatibility increases the likelihood that hybrids inherit a chloroplast that is compatible with the hybrid nuclear genome. In addition, inheriting multiple chloroplast genomes introduces genetic variation and raises the possibility for selection. For example, if one chloroplast is incompatible with the hybrid nuclear genome, selection against that chloroplast among individuals, along with vegetative sorting within individuals, could lead to loss of that haplotype in the next generation (Birky, 2001) with a concomitant recovery in fitness. Both of these mechanisms would enable biparental chloroplast inheritance to act as a rescue mechanism for cytonuclear incompatibility, reducing its contribution to reproductive isolation. The dynamics of selection occurring within the soma is relatively unexplored, and there is a need to examine the interplay

between selection within and among individuals and how this may influence the role of cytonuclear incompatibility in speciation.

The rate of evolution in the organelle genomes is another mechanism that may influence the evolutionary dynamics of cytonuclear incompatibility and its contribution to reproductive isolation. The chloroplast genome is generally conserved across angiosperms in terms of structure and gene content (Jansen & Ruhlman, 2012), and rates of nucleotide substitution are usually low relative to the nuclear genome (Drouin *et al.*, 2008). However, multiple lineages of angiosperms exhibit accelerated plastid evolution, characterized by extensively rearranged plastid genomes as well as increased rates of nucleotide substitution and elevated dN/dS ratios for some plastid genes (Jansen *et al.*, 2007; Guisinger *et al.*, 2008; Sloan *et al.*, 2012; Sloan *et al*, 2014a). These characteristics have the potential to lead to a more rapid co-evolution of the nuclear and chloroplast genomes (e.g. Sloan *et al.*, 2014b), increasing the likelihood for cytonuclear incompatibilities when crossing between populations.

The goal of this dissertation is to examine the contribution of cytonuclear incompatibility to the early stages of the speciation process, and how factors such as biparental chloroplast inheritance and accelerated plastid evolution may influence the evolutionary dynamics of these incompatibilities and alter their contribution to incipient speciation.

To investigate the interplay between these factors, I used the plant species, *Campanulastrum americanum*. Previous studies using a limited number of *C. americanum* populations found evidence for substantial, intraspecific reproductive isolation (Galloway & Etterson, 2005; Etterson *et al.*, 2007). Reproductive isolation was

asymmetrical, suggesting a potential contribution of cytonuclear incompatibility. *Campanulastrum americanum* is also a member of the Campanulaceae, which have the potential for biparental chloroplast inheritance (Corriveau & Coleman, 1988; Zhang *et al.*, 2003) and also have highly rearranged plastid genomes (Cosner *et al.*, 2004; Knox, 2014). The potential co-occurrence of cytonuclear incompatibility, biparental plastid inheritance, and accelerated plastid evolution in *C. americanum* make this an informative system for examining the interactions between these processes.

The first chapter examines how glacial cycles have contributed to the current genetic structure of *C. americanum*, and establishes the existence of multiple geographically structured, genetically divergent lineages. Three of the four genetic lineages are restricted primarily to the Appalachians, while a fourth occurs throughout the rest of the species range. These patterns result in the Appalachians being both genetically divergent from and more genetically diverse than the rest of the species range.

The second chapter determines whether reproductive isolation exists between the divergent genetic lineages of *C. americanum*, and if cytonuclear incompatibility contributes to this isolation. Substantial reproductive isolation between lineages was found for both germination (up to 87% reduction) and survival (up to 81% reduction), with the strength of isolation related to the degree of geographic and chloroplast-genetic distance between populations. There was also evidence for cytonuclear incompatibility underlying reproductive isolation in survival, supporting the idea that cytonuclear incompatibilities arise early in the speciation process. However, reproductive isolation in germination appeared to be caused by triploid endosperm interactions or maternal-zygote incompatibilities, suggesting that other incompatibilities may be of equal importance.

The third chapter examined whether patterns of reproductive isolation are altered in the part of the geographic range where divergent lineages now co-occur. When crossing populations within the Appalachians, reproductive isolation between genetic lineages was only expressed for survival and was less consistently predicted by chloroplast genetic distance relative to when crossing between allopatric populations. These results suggest that gene flow among divergent lineages in the Appalachians may be leading to a reduction in reproductive isolation and the merging of once disparate lineages.

The fourth chapter explored the frequency of biparental plastid inheritance in *C. americanum*, and how it interacts with and influences the contribution of cytonuclear incompatibilities to reproductive isolation. Biparental plastid inheritance was found when crossing both within and between genetic lineages, although the degree of biparental inheritance was higher in the presence of genetic divergence. Biparental inheritance acted as a rescue mechanism for cytonuclear incompatibility, increasing the fitness of F1 hybrids in genetically-divergent crosses and leading to recovery in the F2 hybrid generation. However, the importance of this rescue mechanism varied among crosses, depending upon the strength of the cytonuclear incompatibility. Overall, biparental inheritance reduced the contribution of cytonuclear incompatibility to reproductive isolation in *C. americanum*, potentially slowing the speciation process.

The fifth chapter evaluated potential evidence for accelerated plastid evolution in *C. americanum*, and whether accelerated evolution is also found at the intraspecific level. Accelerated rates of nucleotide substitution and elevated dN/dS ratios were found in a number of plastid genes when comparing sequence divergence between *C. americanum*

and other species. A similar subset of genes also showed elevated pN/pS (the within-species equivalent of dN/dS) when looking within *C. americanum*. These genes are interesting candidates for investigating the underlying cause of cytonuclear incompatibility in this species. Overall, the results suggest that mechanisms leading to increased nucleotide substitution rates in the plastid genome of the species are continuing to act at the intraspecific level, and that accelerated plastid evolution could be driving the development of cytonuclear incompatibility in *C. americanum*.

In summary, this research suggests that *C. americanum* is in the early stages of the speciation process. A cytonuclear incompatibility leading to substantial reproductive isolation for survival supports the theory that cytonuclear incompatibilities arise early in the speciation process, although reproductive isolation for germination suggests other genetic incompatibilities may be of equal importance. However, biparental chloroplast inheritance was able to rescue hybrids from the cytonuclear incompatibility, reducing the contribution of this incompatibility to reproductive isolation and potentially slowing the speciation process. Evidence was also found for the mechanisms of accelerated plastid evolution at the intraspecific level, likely facilitating the development of cytonuclear incompatibility in *C. americanum*. Overall, cytonuclear incompatibility appears to be contributing to incipient speciation in *C. americanum*. However, this contribution is being influenced by both accelerated plastid evolution, which may increase the relative contribution of cytonuclear incompatibility to reproductive isolation, and biparental plastid inheritance, which could in turn reduce the impact of cytonuclear incompatibility. Interestingly, these three processes (cytonuclear incompatibility, accelerated plastid evolution, and biparental plastid inheritance) appear to co-occur in several independent

lineages of angiosperms, suggesting the interactions described here are likely to influence the dynamics of reproductive isolation and speciation across multiple taxa.

References

Avise, J.C., Walker, D. & Johns, G.C. 1998. Speciation durations and Pleistocene effects on vertebrate phylogeography. *Proceedings of the Royal Society B-Biological Sciences* **265:** 1707-1712.

Barr, C.M., Neiman, M. & Taylor D.R. 2005. Inheritance and recombination of mitochondrial genomes in plants, fungi, and animals. *New Phytologist* **168:** 39-50.

Birky, C.W. 2001. The inheritance of genes in mitochondria and chloroplast: laws, mechanisms, and models. *Annual Review of Genetics* **35:** 125-148.

Burton, R.S. & Barreto, F.S. 2012. A disproportionate role for mtDNA in Dobzhansky-Muller incompatibilities? *Molecular Ecology* **21:** 4942-4957.

Burton, R.S., Pereira, R.J. & Barreto, F.S. 2013. Cytonuclear genomic interactions and hybrid breakdown. *Annual Review of Ecology, Evolution, and Systematics, Vol 44* **44:** 281-302.

Carstens, B.C. & Knowles, L.L. 2007. Shifting distributions and speciation: species divergence during rapid climate change. *Molecular Ecology* **16:** 619-627.

Chase, C.D. 2007. Cytoplasmic male sterility: a window to the world of plant mitochondrial-nuclear interactions. *Trends in Genetics* **23:** 81-90.

Chiu, W.L. & Sears, B.B. 1993. Plastome-genome interactions affect plastid transmission in *Oenothera*. *Genetics* **133:** 989-997.

Chou, J.Y. & Leu, J.Y. 2010. Speciation through cytonuclear incompatibility: Insights from yeast and implications for higher eukaryotes. *Bioessays* **32:** 401-411.

Corriveau, J.L. & Coleman, A.W. 1988. Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperm species. *American Journal of Botany* **75:** 1443-1458.

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

Coyne, J.A. & Orr, H.A. 2004. *Speciation*. Sinauer Associates, Sunderland, Mass.

Drouin, G., Daoud, H. & Xia, J. 2008. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Molecular Phylogenetics and Evolution* **49:** 827-831.

Ellison, C.K. & Burton, R.S. 2008. Interpopulation hybrid breakdown maps to the mitochondrial genome. *Evolution* **62:** 631-638.

Etterson, J.R., Keller, S.R. & Galloway, L.F. 2007. Epistatic and cytonuclear interactions govern outbreeding depression in the autotetraploid *Campanulastrum americanum*. *Evolution* **61:** 2671-2683.

Excoffier, L., Foll, M. & Petit, R.J. 2009. Genetic consequences of range expansions. *Annual Review of Ecology Evolution and Systematics* **40:** 481-501.

Fishman, L. & Willis, J.H. 2006. A cytonuclear incompatibility causes anther sterility in *Mimulus* hybrids. *Evolution* **60:** 1372-1381.

Galloway, L.F. & Etterson, J.R. 2005. Population differentiation and hybrid success in *Campanula americana*: geography and genome size. *Journal of Evolutionary Biology* **18:** 81-89.

Greiner, S., Rauwolf, U., Meurer, J. & Herrmann, R.G. 2011. The role of plastids in plant speciation. *Molecular Ecology* **20:** 671-691.

Greiner, S. & Bock, R. 2013. Tuning a menage a trois: co-evolution and co-adaptation of nuclear and organellar genomes in plants. *Bioessays* **35:** 354-365.

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. *Proceedings of the National Academy of Sciences* **105:** 18424-18429.

Hewitt, G. 2000. The genetic legacy of the Quaternary ice ages. *Nature* **405:** 907-913.

Hewitt, G.M. 1996. Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society* **58:** 247-276.

Hu, Y.C., Zhang, Q., Rao, G.Y. & Sodmergen 2008. Occurrence of plastids in the sperm cells of Caprifoliaceae: biparental plastid inheritance in angiosperms is unilaterally derived from maternal inheritance. *Plant and Cell Physiology* **49:** 958-968.

Jansen, R. & Ruhlman, T. 2012. Plastid genomes of seed plants. In: *Genomics of Chloroplasts and Mitochondria*, Vol. 35 (R. Bock & V. Knoop, eds), pp. 103-126. Springer Netherlands.

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., Depamphilis, C.W., Leebens-Mack, J., Muller, 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. *Proceedings of the National Academy of Sciences* **104:** 19369-19374.

Knox, E.B. 2014. The dynamic history of plastid genomes in the Campanulaceae sensu lato is unique among angiosperms. *Proceedings of the National Academy of Sciences* **111:** 11097-11102.

Leppala, J. & Savolainen, O. 2011. Nuclear-cytoplasmic interactions reduce male fertility in hybrids of *Arabidopsis lyrata* subspecies. *Evolution* **65:** 2959-2972.

Levin, D.A. 2003. The cytoplasmic factor in plant speciation. *Systematic Botany* **28:** 5-11.

Levy, F. 1991. A genetic analysis of reproductive barriers in *Phacelia dubia*. *Heredity* **67:** 331-345.

Mayr, E. 1942. *Systematics and the origin of species from the viewpoint of a zoologist*. Columbia University Press, New York,.

Mogensen, H.L. 1996. The hows and whys of cytoplasmic inheritance in seed plants. *American Journal of Botany* **83:** 383-404.

Rand, D.M., Haney, R.A. & Fry, A.J. 2004. Cytonuclear coevolution: the genomics of cooperation. *Trends in Ecology & Evolution* **19:** 645-653.

Sambatti, J.B.M., Ortiz-Barrientos, D., Baack, E.J. & Rieseberg, L.H. 2008. Ecological selection maintains cytonuclear incompatibilities in hybridizing sunflowers. *Ecology Letters* **11:** 1082-1091.

Scopece, G., Lexer, C., Widmer, A. & Cozzolino, S. 2010. Polymorphism of postmating reproductive isolation within plant species. *Taxon* **59:** 1367-1374.

Seehausen, O. 2006. Conservation: Losing biodiversity by reverse speciation. *Current Biology* **16:** R334-R337.

Servedio, M.R. & Noor, M.A.F. 2003. The role of reinforcement in speciation: theory and data. *Annual Review of Ecology Evolution and Systematics* **34:** 339-364.

Sloan, D.B., Alverson, A.J., Wu, M., Palmer, J.D. & Taylor, D.T. 2012. Recent acceleration of plastid sequence and structural evolution coincides with extreme mitochondrial divergence in the angiosperm genus *Silene*. *Genome Biology and Evolution* **4:** 294-306.

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.

Sloan, D.B., Triant, D.A., Wu, M. & Taylor, D.R. 2014b. Cytonuclear interactions and relaxed selection accelerate sequence evolution in organelle ribosomes. *Molecular Biology and Evolution* **31:** 673-682.

Taylor, E.B., Boughman, J.W., Groenenboom, M., Sniatynski, M., Schluter, D. & Gow, J.L. 2006. Speciation in reverse: morphological and genetic evidence of the collapse of a three-spined stickleback (*Gasterosteus aculeatus*) species pair. *Molecular Ecology* **15:** 343-355.

Tiffin, P., Olson, M.S. & Moyle, L.C. 2001. Asymmetrical crossing barriers in angiosperms. *Proceedings of the Royal Society of London Series B-Biological Sciences* **268:** 861-867.

Turelli, M. & Moyle, L.C. 2007. Asymmetric postmating isolation: Darwin's corollary to Haldane's rule. *Genetics* **176:** 1059-1088.

Zhang, Q., Liu, Y. & Sodmergen 2003. Examination of the cytoplasmic DNA in male reproductive cells to determine the potential for cytoplasmic inheritance in 295 angiosperm species. *Plant and Cell Physiology* **44:** 941-951.

Zhang, Q. & Sodmergen 2010. Why does biparental plastid inheritance revive in angiosperms? *Journal of Plant Research* **123:** 201-206.

Chapter 1:

Multiple glacial refugia lead to genetic structuring and the potential for reproductive isolation in an herbaceous plant

This article is formatted as a co-authored manuscript (Barnard-Kubow KB, Debban CL, and Galloway, LF)

Abstract

Glacial cycles have impacted the current genetic structure of many species. In addition to facilitating genetic divergence, isolation in allopatric glacial refugia may also have contributed to the development of genetic incompatibility and reproductive isolation. In this study we examine the phylogeography of *Campanulastrum americanum*, an herbaceous plant that exhibits strong intraspecific reproductive isolation, to determine whether the current genetic structure reflects a history of multiple allopatric glacial refugia. Chloroplast loci and nuclear RAD sequencing were used to characterize the range-wide phylogeography of *C. americanum* to determine the location of potential glacial refugia as well as recolonization routes. Potential locations of refugia during the last glacial maximum were also identified with ecological niche modeling. Taken together, the chloroplast and nuclear phylogenies found support for the existence of four geographically structured, genetically divergent clades. The current distribution of these clades indicates *C. americanum* likely survived the last glacial maximum in at least three allopatric refugia located in the Appalachians and on the Atlantic and Gulf coasts. The ecological niche model supported both coastal refugia, but provided no evidence for the Appalachian one. The isolation of populations of *C. americanum* in allopatric refugia over multiple glacial cycles has led to geographically structured, genetically divergent clades. Populations of *C. americanum* previously shown to be reproductively isolated belong to divergent clades, suggesting that the allopatric glacial refugia may have contributed to the development of reproductive isolation in this species.

Introduction

The genetic structure of a species is a manifestation of both historic processes and contemporary patterns of gene flow. One of the primary historical processes that have influenced genetic structure is climate change, in particular glacial cycles, which have caused cyclic expansions and contractions of many species' ranges (reviewed in Hewitt, 2000). Phylogeography (Avise *et al.*, 1987) is often used to understand how the current genetic structure of a species has been impacted by glacial cycles, including the number and location of glacial refugia as well as the recolonization routes from these refugia to current locations (Hewitt, 1996; Taberlet *et al.*, 1998; Hewitt, 2000).

Several themes have emerged from phylogeographic studies. These include the frequent location of glacial refugia in the southern reaches of the northern hemisphere, with refugia found in the southern peninsulas of Europe and along the Gulf Coast in the eastern United States (Taberlet *et al.*, 1998; Avise, 2000; Hewitt, 2000; Soltis *et al.*, 2006). There is also now general acceptance of additional "cryptic" refugia where species survived close to the ice sheet both in Europe and North America (Stewart & Lister, 2001; Tribsch & Schonswetter, 2003; McLachlan *et al.*, 2005; Provan & Bennett, 2008). In Europe, distinct species exhibit congruence with respect to the locations of glacial refugia, subsequent colonization routes, and areas of secondary contact (Taberlet *et al.*, 1998; Hewitt, 2000). While some congruence has also been seen in eastern North America (Avise, 2000; Soltis *et al.*, 2006), species' responses to glacial cycles generally appear individualistic and complex (Soltis *et al.*, 2006). This lack of congruence is likely the result of the north-south orientation of the Appalachian Mountain having acted as less of a migration barrier than the east-west orientation of the major mountain ranges in

Europe. This complexity raises the need for phylogeographic studies covering a diverse array of organisms. However, studies in eastern North America have historically been biased towards animals (Soltis *et al.*, 2006), leaving the phylogeography of plant species relatively understudied, particularly for shorter lived taxa (though see Griffin & Barrett, 2004; Gonzales *et al.*, 2008; Li *et al.*, 2013).

Isolation in multiple allopatric glacial refugia may have led to genetic divergence among populations due to drift resulting from small refugial population sizes, along with different selective regimes among refugia (Hewitt, 1996). Glacial periods have been longer than interglacials, meaning that during the 18-20 glacial cycles in the last 2 million years, species spent more time in glacial refugia than in their current widespread distributions (reviewed in Davis 1983). Recolonization would have further contributed to drift and genetic divergence as dispersal events lead to recurrent bottlenecks (Excoffier *et al.*, 2009). These processes may also have facilitated the development of genetic incompatibility and the potential for reproductive isolation and speciation (Hewitt, 1996; Avise *et al.*, 1998; Carstens & Knowles, 2007). If isolation in glacial refugia has contributed to the speciation process, the existence of divergent intraspecific clades identified in phylogeographic studies may correspond to incipient species. However, few studies have directly addressed the interplay between phylogeography and intraspecific reproductive isolation (though see Gomez *et al.*, 2007; April *et al.*, 2013; Pinheiro *et al.*, 2013; Singhal & Moritz, 2013).

To understand how glacial cycles may have contributed to the development of genetic incompatibility and reproductive isolation, we examined the phylogeography of *Campanulastrum americanum*, an herbaceous plant that has strong reproductive isolation

between populations (up to 90% reduction in cumulative fitness; Galloway & Etterson, 2005; Etterson *et al.*, 2007). In this study we address the following questions: 1) Does the phylogeography of *C. americanum* based on chloroplast and nuclear markers support survival in multiple allopatric glacial refugia, and are the patterns concordant between the two types of markers? 2) Does niche modeling also support multiple allopatric glacial refugia, and do the proposed locations match those suggested by phylogeography?

Materials and Methods

Study System

Campanulastrum americanum (L.) Small (=*Campanula americana* L.), in the Campanulaceae is an autotetraploid, monocarpic herb found in the eastern half of the United States, with populations from the Appalachian Mountains to just west of the Mississippi river (Fig 1C). Populations are not evenly distributed across the range, as the species is less common in the south (southern Mississippi, Alabama, and northern Florida). Individuals can be annual or biennial, are insect-pollinated, and primarily outcrossing (Galloway *et al.*, 2003). *C. americanum* typically grows in disturbed habitats in the understory or along the forest edge and seeds are dispersed passively, traits that likely contribute to its patchy population structure.

Chloroplast Sequencing and Phylogeny

For determining the chloroplast phylogeny of *C. americanum,* leaf tissue samples were taken from individuals from 49 populations across the species range (Table 1). Leaf tissue was primarily taken from individuals grown in the greenhouse from field-collected

seed, though a small number of field-collected leaf tissue samples were also used. Leaf tissue from *Triodanus biflora* was used as an outgroup for the phylogeny construction. DNA was extracted for all samples using a modified CTAB procedure.

Screening of pre-existing chloroplast primer sets (Taberlet *et al.*, 1991; Hamilton, 1999) gave limited success. To obtain more loci, we screened for areas of high polymorphism in a 454 dataset that consisted of full chloroplast genome sequence of four pooled *C. americanum* populations (AL29, MN12, OH64, and VA73; Chapter 5). From this screen we obtained four primer sets, which along with a trnL intron and trnL-trnF intergenic spacer (Taberlet *et al.*, 1991), were sequenced in all individuals and used for constructing the chloroplast phylogeny. Due to difficulties with consistent amplification, *C. americanum* specific external primers were designed for the trnL intron and trnL-trnF intergenic space. Primer sequences used in constructing the chloroplast phylogeny and their location in the chloroplast genome are listed in Supp. Table 1.

A single individual was sequenced at all five loci in each of the populations. Additional individuals (5 to 13 per population) were sequenced for 13 of the populations to check for polymorphism (Supp. Table 2). If populations were fixed for chloroplast haplotype or contain only closely related haplotypes, sampling additional individuals per population is unlikely to alter the overall phylogeographic pattern (geographic distribution of clades). To maximize the likelihood of detecting polymorphism, populations for additional sampling were chosen that were located in the proximity of expected southern glacial refugia (from other species) or in the Appalachians where divergent chloroplast clades were found to occur in close proximity. Most of the additional individuals were genotyped at the subset of loci that distinguished them from

nearby haplotypes as we were primarily concerned with detecting shared haplotypes among populations.

Loci were amplified using 5 PRIME HotMasterMix. PCR products were purified using Exonuclease I and Shrimp Alkaline Phosphatase (Affymetrix). Cycle sequencing reactions were carried out using BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Life Technologies) and were cleaned using Sephadex G-50 (GE-Healthcare). The resulting products were analysed on an ABI 3130xl sequencer (Applied Biosystems, Life Technologies), and sequences were checked and manually aligned using Biodit (Hall, 1999) or Codon Code Aligner v3.5 (CodonCode Corporation).

Sequences were concatenated across loci for each individual. Maximum likelihood trees were constructed in Mega v5.2.2 (Tamura *et al.*, 2011) for each locus as well as for the concatenated sequences. A GTR model of evolution was used for C11 and rps4, while HKY was used for ycf1 and CLF, and GTR + I was used for rps2 and the concatenated sequences. Models of evolution were chosen based on results of jModeltest v2.1.5 (Guindon & Gascuel, 2003; Darriba *et al.*, 2012). Bayesian analysis was also run on the concatenated sequences using Mr. Bayes v3.2.2 (Ronquist *et al.*, 2012), again using a GTR + I model of evolution. Two independent analyses were run starting with random trees and using one cold and three heated chains. Each analysis ran for 1,000,000 generations, sampling one tree every 1000 generations. At the end of the run the standard deviation of split frequencies was below 0.01, indicating convergence. The first 25% of samples were discarded from the cold chain, and posterior probabilities for supported clades were determined by majority-rule consensus of trees retained after burn-in. A

statistical parsimony network was constructed for the concatenated sequences using TCS v1.21 (Clement *et al.*, 2000) with a 95% parsimony connection limit.

Nuclear Sequencing and Phylogeny

To determine the nuclear phylogeny of *C. americanum*, 25 populations were chosen (Table 1) for RAD-seq (Davey *et al.*, 2011). This technique enables the random sampling of SNPs from across the genome. Populations were chosen to represent all chloroplast haplotypes as well as to achieve a broad geographic sampling with a focus on the Appalachian region where the chloroplast clades co-occur. For each population, leaf tissue samples were taken from 6 individuals grown from field-collected seed, except for VA93 where first generation greenhouse seed was used. DNA was extracted using a modified CTAB procedure. The six DNA extractions from each population were evenly pooled and then barcoded.

The barcoded populations were digested with SacI and then pooled for library construction following standard protocols for sequencing on an Illumina Hiseq. Due to low sequence complexity resulting from the restriction enzyme cut site, a 50% PhiX control was spiked into the RAD library before sequencing on two lanes of an Illumina HiSeq with 100bp reads. Library construction was performed at the Genomics Core Facility in the University of Virginia's Department of Biology, while sequencing was performed at Beckman Coulter Genomics.

The fastq files were run through *process_radtags* in Stacks (Catchen *et al.*, 2011; 2013) to demultiplex the data and filter out low quality reads. The processed fastq files were individually mapped using bowtie2 (Langmead & Salzberg, 2012) to a partial

reference genome for *C. americanum* obtained from earlier sequencing of a whole genome shotgun library (unpublished data). The resulting alignments were used as input for Stacks (Catchen *et al.*, 2011; 2013) to identify and genotype loci using the following programs: *pstacks*, *cstacks*, *sstacks* and *populations*. A minimum read depth of 12 was used when constructing stacks to increase the chances that each individual within a pool was represented by at least a single read. Due to *C. americanum* being an autotetraploid, and to ensure SNPs were between homologous sequences instead of paralogous ones, we only used SNPs that were fixed within populations but variable between. Therefore, we used the fixed SNP model in *pstacks* (Catchen *et al.*, 2013). The barcode error frequency was estimated using the percent of reads that were dropped in *process_radtags* due to ambiguous barcodes. To ensure that SNPs were from homologous sequence, only one mismatch was allowed between tags when generating the catalog in *cstacks*. *Populations* was run using loci present in at least eight populations. Analyses were also run using loci present in a minimum of four, six, or twelve populations with qualitatively similar results.

A perl script was used to concatenate the sequences from the Stacks fasta output for each population (contained full sequence of each allele from each sample locus). The concatenated sequences were then used to construct a maximum likelihood tree in RAxML v8.0.6 (Stamatakis, 2014) using 100 bootstrap replicates and a GTR plus gamma model of evolution. The tree was visualized in Mega (Tamura *et al.*, 2011) and rooted according to the chloroplast phylogeny. For comparison, the RAD data was also run through Stacks denovo, using *ustacks* instead of *pstacks* with similar parameter options.

Ecological Niche Modeling and Climate Analysis

To identify possible glacial refugia for *C. americanum*, an ecological niche model was created in MaxEnt version 3.3.3k (Phillips *et al.*, 2006). This form of modeling uses known occurrence data and current environmental data to generate a model that predicts the location of suitable habitat for a species. This model can then be projected onto past climates to predict how the location of suitable habitat has changed over time. For the occurrence data, 963 locations were obtained from the Galloway lab, the Global Biodiversity Information Facility (www.gbif.org), the Wisconsin Vascular Plants WISCOMP database, the Illinois Natural History Survey, Miami University Herbarium, Ohio State University herbarium, and the University of North Carolina herbarium. To reduce spatial autocorrelation in these points, data were rarefied using SDMtoolbox (Brown, 2014) so that no occurrence was within 40km of another, leaving 301 occurrence points for analysis.

Five climatic variables were chosen from 19 bioclimatic variables (www.worldclim.org) of present (Hijmans *et al.*, 2005) and Pleistocene climates (Braconnot *et al.*, 2007) with a resolution of 2.5 arc-minutes: maximum temperature of the warmest month, minimum temperature of the coldest month, precipitation of the wettest month, precipitation of the driest month, and precipitation of the warmest quarter. These variables were selected by choosing those likely to be most biologically important to *C. americanum* based on its growing season and susceptibility to drought, followed by removal of variables with a correlation coefficient greater than 0.8 (Sheppard, 2013). Climate layers were projected using an Albers equal-area projection to train the model to prevent latitudinal bias in pseudoabsence point selection (Brown 2014). To prevent overfitting due to selection of pseudoabsence points outside the range of equilibrium

(Anderson & Raza, 2010), the area where the model trained was restricted to a minimum convex polygon with a buffer of 150km outside the outermost points. A regularization multiplier of 1.3 was selected because it created a model with the lowest AIC_C (Warren & Seifert, 2011). Maxent uses regularization multipliers to determine how closely to fit the model to the data; values too low may lead to an over-fit model, while higher values give smoother models. Ten-fold cross-validation was used to create the model.

To examine whether genetic lineages of *C. americanum* occur in different ecological niches, data for the five bioclimatic variables were extracted for the 49 populations used for chloroplast sequencing. These bioclimatic variables were run through principal components analysis (PROC PRINCOMP, SAS 9.4, SAS Institute, INC. 2011) and the populations were graphed according to the first and second principal component. As the second principal component (PC2) appeared to separate Appalachian and non-Appalachian populations, ANOVA was used to determine whether chloroplast and/or nuclear clade explained significant variation in PC2 (PROC GLM, SAS 9.4, SAS Institute, INC. 2011).

Results

Chloroplast Sequencing and Phylogeny

Individual chloroplast loci distinguished between three to nine haplotypes (Supp. Table 1). Concatenation of the five loci resulted in 2437 base pairs of sequence with 33 SNPs distinguishing 14 haplotypes. The maximum likelihood trees generated for the individual loci and the concatenated sequence did not contradict each other. Therefore, only the

concatenated tree is presented. Sequences for each locus are deposited in Genbank (KP053958-KP054027).

Both the maximum likelihood and Bayesian phylogenetic analyses, as well as the statistical parsimony network, found support for three chloroplast clades (Fig 1A, B). Percent sequence divergence ranged from 0.04 to 0.25 within clades, from 0.74 to 0.86 between the Appalachian and Eastern clades, from 0.21-0.41 between the Eastern and Western clades, and from 0.95-1.15 between the Appalachian and Western clades. An estimate of chloroplast synonymous substitution rate (Wolfe *et al.*, 1987) places the divergence of the Appalachian clade to 2.3-7 million years ago (Ma), the Eastern and Western clade to 0.7-2.3 Ma, and the divergence of the Western clade branches to 0.5-1.6 Ma. These dates place the divergence of the Appalachian clade as likely occurring pre-Pleistocene, while the other divergences appear to have occurred during the Pleistocene.

The clades are geographically structured with the highly divergent Appalachian clade (haplotypes A-D) restricted to the Appalachian Mountains. Within the Appalachian clade the populations in the southern Appalachians (haplotypes A, B) are differentiated from the more northern populations (haplotypes C, D). The Eastern clade (haplotypes E, F) contains populations in the eastern Appalachians along with a coastal disjunct population in southeastern Virginia, while the Western clade (haplotypes G-N), shows a much wider distribution with populations throughout the range, including the southern Appalachians. The Western clade also shows some sub-structuring with two branches occurring either east or proximate to and west of the Mississippi. The basal populations of the Western clade (haplotype I) occur in the most southern part of the range (Mississippi, Alabama, Florida), as well as in the southern Appalachians.

Of the 13 populations where additional individuals were sequenced, 10 were monomorphic (Supp. Table 2). Two of the polymorphic populations had private haplotypes. Only one population had haplotypes shared by multiple populations (MS55, haplotypes I and J), and both of these haplotypes fall in the Western chloroplast clade. Therefore, sampling additional individuals from populations does not appear likely to change the overall phlylogeographic pattern.

Nuclear Sequencing and Phylogeny

RAD sequencing produced between five to nine million reads per population. Approximately 40% of reads from each population mapped to the reference genome. Mapping reads to the reference genome resulted in an average depth of coverage of 30x across loci. Running the data through Stacks (Catchen *et al.*, 2011; Raineri *et al.*, 2012; Catchen *et al.*, 2013) resulted in a total of 11,215 loci that were used for constructing the maximum likelihood tree.

The nuclear tree produced similar results to the chloroplast tree as it also found support for three clades, which are similar in pattern (Fig 2). The Appalachian chloroplast clade that was restricted to the Appalachians is recovered in the nuclear tree (haplotypes A-D), though the deep divergence between this clade and the remaining clades is no longer apparent. In addition, similar sub-structuring is seen between populations in the southern (haplotypes A, B) and northern Appalachians (haplotype D). However, bootstrap support for the southern Appalachian populations grouping with the northern Appalachian populations is relatively low compared to the chloroplast tree,

suggesting that some of the nuclear data supports an alternative grouping for these southern populations.

The Eastern chloroplast clade found in the eastern Appalachians and the Atlantic coast was also recovered in the nuclear tree, but these populations are now grouped with the Western chloroplast clade populations that occur in the southern Appalachians (haplotypes I, J), henceforth referred to as the Smokies populations, after the Great Smokey Mountains in which several of these populations occur. However, there is also strong support for the Eastern chloroplast populations (haplotypes E, F) forming a monophyletic group. Within this clade there is support for the coastal disjunct population, VA93, being basal to the remaining populations.

The third clade in the nuclear tree contains all of the populations west of the Appalachians and corresponds well to the Western chloroplast clade with the exception of the basal Appalachian populations discussed above. The nuclear tree again shows support for the populations west of the Mississippi grouping together, and there is some support for the north-central populations (IN68 and MI44) being basal to these. The remaining populations, which are primarily southern, are difficult to resolve in terms of their exact placement on the tree as reflected by low bootstrap values, though there was support for AL29, TN34, and PA27 grouping together. Bootstrap values can be inflated in large-scale data sets. However, the alternative de novo analysis of the RAD-seq data produced a similar topology (Supp. Fig 1), providing further confirmation of these patterns.

Taken together, the chloroplast and nuclear phylogenies indicate the existence of four genetically divergent lineages within *C. americanum*. Three of these, including the

highly divergent Appalachian clade, are restricted primarily to the Appalachians, while the western clade occurs throughout the range outside of the Appalachians. This distribution results in populations located within the Appalachian Mountains being both genetically divergent and more genetically diverse than the rest of the species range.

Ecological Niche Modeling and Climate Analysis

The current suitable range indicated by our species distribution model corresponds well to the known current range of *C. americanum* (Figure 3A vs 1C), except for the northeast, where *C. americanum* does not occur, despite the prediction of suitable habitat. This discrepancy may be due to ongoing recolonization in this area. The predicted highly suitable areas in the western part of the range overlap with areas of apparent recent recolonization by haplotypes G, H, and L, while the unsuitable area in the deep south corresponds to sparser distribution of populations. The average AUC of the cross-validated models was 0.700. While the AUC of the model is in the low range of acceptable values (Swets, 1988; Araujo *et al.*, 2005), low AUC is not uncommon for widespread species (Lobo et al. 2008). In addition, AUC may not be the best way of identifying a good model (Lobo et al. 2008), so we also used binomial omission tests to analyze our model. By this test, our model's predictions were significantly better than random using multiple thresholds, including thresholds that minimize the difference between sensitivity and specificity, thresholds that maximize the sum of sensitivity and specificity, and a ten-percentile training presence threshold. These thresholds have been shown to be useful measures of model performance (Jiménez-Valverde and Lobo 2007). When the model was projected onto Pleistocene climate data (Figure 3B), potential

glacial refugia were apparent along the Gulf and Atlantic coasts, near present-day Texas, Florida, and North Carolina. Another apparent refugium appears inland in Alabama.

Principal components analysis of bioclimatic variables for all 49 populations resulted in first and second principal components that explain 55.06% and 29.65% of the total variance respectively. Most of the variation in PC1 is found among populations of the western nuclear clade, while PC2 separates out the Appalachian populations (Appalachian and Eastern nuclear clades), including the Smokies populations (haplotypes I and J) that fall out in the Western chloroplast, but Appalachian nuclear clade (Supp. Fig 2). Both nuclear and chloroplast clade explained variation in the values of PC2 ($p < 0.01$ and $p = 0.04$ respectively). However, more of the variation was explained by nuclear clade ($R^2=0.52$ versus $R^2=0.13$), which better delineates Appalachian and non-Appalachian populations. These results suggest that Appalachian populations occupy a different climatic niche than non-Appalachian populations. Altogether, the Smokies I and J populations appear to group closer with the other Appalachian populations in terms of climate and nuclear genetics, though they fall out with the non-Appalachian populations in terms of chloroplast haplotype.

Discussion

Both the phylogeography and ecological niche modeling support the presence of multiple glacial refugia in *C. americanum*, leading to several geographically structured, genetically divergent lineages. The distributions of these lineages, including a Mississippi discontinuity, an east-west Appalachian discontinuity, as well as an Appalachian refugium, all match recurrent patterns found in previous studies of both plants and

animals in eastern North America (Soltis *et al.*, 2006). However, rarely are all of these patterns found within a single species. In comparison to studies of other plants in eastern North America, including trees and herbaceous species, (Griffin & Barrett, 2004; McLachlan *et al.*, 2005; Gonzales *et al.*, 2008; Morris *et al.*, 2008; Li *et al.*, 2013), we generally find a greater degree of genetic divergence and geographic structuring in *C. americanum*. Potential reasons for the greater divergence and structuring in *C. americanum* could be its widespread distribution, relatively short generation time, or patchy population structure. More studies on species with similar life histories are needed to determine whether patchy population structure and/or short generation times generally lead to greater phylogeographic structure.

Much of the current range of *C. americanum* is occupied by the Western clade, with three haplotypes (G, H, L) occurring in areas of widespread homogeneity suggestive of recent recolonization from southern glacial refugia (Hewitt, 1996). These areas of widespread homogeneity also overlap with much of the highly suitable habitat designated by the current-climate ecological niche model, suggesting rapid recolonization occurred as suitable habitat shifted northward after the last glacial maximum. Due to the low mutation rate of the chloroplast genome (Wolfe *et al.*, 1987), the majority of modern chloroplast haplotypes likely predate the most recent postglacial recolonization, allowing inference of recolonization routes from the current geographic distribution of chloroplast haplotypes (McLachlan *et al.*, 2005). Accordingly, the location of the two Western chloroplast clade branches east (haplotype L), and west (haplotypes G and H) of the Mississippi suggests separate recolonization routes, possibly from separate refugia. The ecological niche model provides some evidence for where these refugia may have been

located. Recolonization west of the Mississippi may have occurred from a refugium on the Texas coast, while populations east of the Mississippi may have come from a refugium located inland in Alabama. A similar east-west Mississippi discontinuity has been observed in other species and is a recurrent pattern in the phylogeography of eastern North America (Soltis *et al.*, 2006).

The ecological niche model also finds strong support for a refugium on the Gulf Coast of Florida. Although populations of *C. americanum* likely existed in this refugium during the last glacial maximum, the current deep-south populations do not contain any widespread haplotypes, suggesting that these refugial populations did not contribute greatly to post-glacial recolonization. However, these populations do contain the ancestral haplotype for the Western clade (haplotype I), suggesting the Florida refugium played an important role in past glacial cycles. While populations containing Western-clade haplotypes G, H, and L do not appear to have moved as far south as the Florida refugium during the last glacial maximum, Florida may have provided the sole refugium for Western-clade haplotypes in previous glacial cycles. Populations in this refugial area appear to have stayed put for some time and may now be relatively isolated. These deep-south populations may also be differently adapted, as they occur in a unique environment (Supp. Fig 2) designated as unsuitable habitat according to the ecological niche model, and are phenotypically different from remaining populations (Prendeville *et al.*, 2013, Barnard-Kubow personal observation). There are other plant species that appear to have genetic lineages mostly restricted to Florida or the deep south, including tulip poplar (Parks *et al.*, 1994) and red maple (McLachlan *et al.*, 2005).

In addition to southern refugia, there is evidence for *C. americanum* having survived the last glacial maximum in a refugium located on the Atlantic coast as suggested by the location of the Eastern chloroplast clade (haplotypes E and F) in the eastern Appalachians and on the Virginia coast. The ecological niche model and the basal position of the Virginia coastal population (VA93) in the nuclear tree provide support for a coastal refugium for this clade. Coastal refugia along the Atlantic coast have been proposed for other species, including the tiger salamander (Church *et al.*, 2003) and American sweetgum (Morris *et al.*, 2008). The coastal plain in the Carolinas has also been considered a likely location for glacial refugia due to the high endemism observed there for plant species (Sorrie & Weakley, 2001). Overall this clade appears to have contributed much less to recolonization than the Western chloroplast and nuclear clade, likely due to the barrier imposed by the Appalachian Mountains.

Restriction of the Appalachian clade to the Appalachian Mountains suggests these populations survived the last glacial maximum in an Appalachian glacial refugium. Furthermore, the deep genetic divergence in the chloroplast tree indicates this clade has likely remained in the Appalachians through many glacial cycles. The divergence of this clade was estimated to have occurred 2.3-7 Ma; indicating the divergence of the Appalachian lineage likely predates the Pleistocene (2.5 – 0.01 Ma). Appalachian refugia have been proposed for other taxa, including both plants (McLachlan *et al.*, 2005; Gonzales *et al.*, 2008) and animals (Brant & Orti, 2003; Church *et al.*, 2003; Austin *et al.*, 2004; Lee-Yaw *et al.*, 2008). In contrast to the other refugia suggested by phylogeography, the ecological niche model provides little support for an Appalachian refugium. However, a major limitation of ecological niche models is that the resolution of

the climatic datasets is generally not fine enough to detect variation due to microclimates, particularly when modeling past climates (Gavin *et al.*, 2014). Mountain ranges are likely to have fine scale climatic variation due to topography, raising the possibility of microrefugia, such as sheltered valleys or south-facing slopes that would be difficult to detect via niche modeling. Therefore, weak support from the ecological niche model does not preclude the existence of an Appalachian refugium for *C. americanum*. A similar situation was found in the millipede genus *Narceus*, where an Appalachian refugium was supported by molecular data, but not the niche-based distribution model (Walker *et al.*, 2009).

Overall there is concordance between the nuclear and chloroplast phylogenies, indicating little gene flow or hybridization between clades. However, there does appear to be less depth of divergence among clades when looking at the nuclear versus chloroplast markers. This difference may be due to the lower effective population size of the chloroplast relative to the nuclear genes, which can lead to increased drift and stronger population differentiation (Birky *et al.*, 1989; Levy & Neal, 1999). Although *C. americanum* has been shown to have increased rates of nucleotide substitutions in a subset of chloroplast genes (Chapter 5), which could contribute to a greater depth of divergence in the chloroplast tree, the deep divergence between the Mountain and Western/Eastern chloroplast clades is found in all five chloroplast loci, including the non-coding regions, indicating the depth of this divergence is likely not driven by increased substitution rates in chloroplast genes. A similar depth of divergence between genetic lineages has not been found in other phylogeographic studies of plant species in eastern

North American (Griffin & Barrett, 2004; McLachlan *et al.*, 2005; Gonzales *et al.*, 2008; Morris *et al.*, 2008; Li *et al.*, 2013).

While overall the chloroplast and nuclear phylogenies give similar results, there are some differences; particularly as regards the basal Western-clade chloroplast haplotypes I and J. Populations with these haplotypes appear to have differing evolutionary histories. The Smokies I and J populations group with the Eastern-chloroplast clade populations in the nuclear tree and occur in a similar environment to the other Appalachian populations, distinct from the remaining Western-clade populations (Supp. Fig 2). Altogether these Smokies populations appear to constitute a fourth genetic lineage, with an evolutionary history distinct from the other western- and eastern-clade populations. Perhaps they originated from one of the same glacial refugia as the other western-clade populations but have now experienced secondary contact and gene flow with other Appalachian populations, or they may have survived and diverged in a separate glacial refugium altogether.

The extensive genetic structuring of a species, as observed here, has the potential to contribute to speciation by facilitating the evolution of genetic incompatibility and reproductive isolation (Hewitt, 1996; Carstens & Knowles, 2007). Is there any evidence for this in *C. americanum*? Strong reproductive isolation is found between populations of *C. americanum*, in particular when crossing between populations from within and outside of the Appalachians, (Galloway & Etterson, 2005; Etterson *et al.*, 2007; Chapters 2 and 3). Several populations with substantial reproductive isolation included in the current study (e.g. IN68 and VA73) belong to disparate genetic clades, while populations that show little reproductive isolation belong to the same clade. This pattern suggests that

reproductive isolation in *C. americanum* likely occurs between genetically divergent lineages that recolonized from different glacial refugia. Accordingly, the historical processes of climate fluctuations and glacial cycles appear to have facilitated the early stages of the speciation process in *C. americanum*, as phylogeographic lineages appear to be incipient species. Our results are consistent with other studies that have found varying degrees of reproductive incompatibility among phylogeographic lineages (Levy 1991; Gomez *et al.*, 2007; Pinheiro *et al.*, 2013; Singhal & Moritz, 2013), suggesting a widespread role for climatic shifts in contributing to the speciation process.

Acknowledgements

We thank the many people who assisted us by collecting seed from populations of *C. americanum*. We thank Deborah Triant for writing the perl script used in concatenating the fasta sequence data and AnhThu Nguyen for constructing the RAD-seq library. We also thank Megan Sebasky for discussions and advice on ecological niche modeling, and Janis Antonovics, Ben Blackman, Edmund Brodie, David Carr, and Douglas Taylor for guidance throughout this project. This work was supported by DEB-0922214 and the ARCS Foundation.

References

Anderson, R.P. & Raza, A. 2010. The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela. *Journal of Biogeography* **37:** 1378-1393.

April, J., Hanner, R.H., Dion-Cote, A.M. & Bernatchez, L. 2013. Glacial cycles as an allopatric speciation pump in north-eastern American freshwater fishes. *Molecular Ecology* **22:** 409-422.

Araujo, M.B., Pearson, R.G., Thuiller, W. & Erhard, M. 2005. Validation of species-climate impact models under climate change. *Global Change Biology* **11:** 1504-1513.

Austin, J.D., Lougheed, S.C. & Boag, P.T. 2004. Discordant temporal and geographic patterns in maternal lineages of eastern north American frogs, *Rana catesbeiana* (Ranidae) and *Pseudacris crucifer* (Hylidae). *Molecular Phylogenetics and Evolution* **32:** 799-816.

Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel, J.E., Reeb, C.A. & Saunders, N.C. 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics* **18:** 489-522.

Avise, J.C., Walker, D. & Johns, G.C. 1998. Speciation durations and Pleistocene effects on vertebrate phylogeography. *Proceedings of the Royal Society B-Biological Sciences* **265:** 1707-1712.

Avise, J.C. 2000. *Phylogeography: the history and formation of species*. Harvard University Press.

Birky, C.W., Fuerst, P. & Maruyama, T. 1989. Organelle gene diversity under migration, mutation, and drift - equilibrium expectations, approach to equilibrium, effects of heteroplasmic cells, and comparison to nuclear genes. *Genetics* **121:** 613-627.

Braconnot, P., Otto-Bliesner, B., Harrison, S., Joussaume, S., Peterchmitt, J.Y., Abe-Ouchi, A., Crucifix, M., Driesschaert, E., Fichefet, T., Hewitt, C.D., Kageyama, M., Kitoh, A., Laine, A., Loutre, M.F., Marti, O., Merkel, U., Ramstein, G., Valdes, P., Weber, S.L., Yu, Y. & Zhao, Y. 2007. Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum - Part 1: experiments and large-scale features. *Climate of the Past* **3:** 261-277.

Brant, S.V. & Orti, G. 2003. Phylogeography of the Northern short-tailed shrew, *Blarina brevicauda* (Insectivora : Soricidae): past fragmentation and postglacial recolonization. *Molecular Ecology* **12:** 1435-1449.

Brown, J.L. 2014. SDMtoolbox: a python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses. *Methods in Ecology and Evolution* **5:** 694-700.

Carstens, B.C. & Knowles, L.L. 2007. Shifting distributions and speciation: species divergence during rapid climate change. *Molecular Ecology* **16:** 619-627.

Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A. & Cresko, W.A. 2013. Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22:** 3124-3140.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W. & Postlethwait, J.H. 2011. Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics* **1:** 171-182.

Church, S.A., Kraus, J.M., Mitchell, J.C., Church, D.R. & Taylor, D.R. 2003. Evidence for multiple pleistocene refugia in the postglacial expansion of the eastern tiger salamander, *Ambystoma tigrinum tigrinum*. *Evolution* **57:** 372-383.

Clement, M., Posada, D. & Crandall, K.A. 2000. TCS: a computer program to estimate gene genealogies. *Molecular Ecology* **9:** 1657-1659.

Darriba, D., Taboada, G.L., Doallo, R. & Posada, D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9:** 772-772.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M. & Blaxter, M.L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12:** 499-510.

Etterson, J.R., Keller, S.R. & Galloway, L.F. 2007. Epistatic and cytonuclear interactions govern outbreeding depression in the autotetraploid *Campanulastrum americanum*. *Evolution* **61:** 2671-2683.

Excoffier, L., Foll, M. & Petit, R.J. 2009. Genetic consequences of range expansions. *Annual Review of Ecology Evolution and Systematics* **40:** 481-501.

Galloway, L.F., Etterson, J.R. & Hamrick, J.L. 2003. Outcrossing rate and inbreeding depression in the herbaceous autotetraploid, *Campanula americana*. *Heredity* **90:** 308-315.

Galloway, L.F. & Etterson, J.R. 2005. Population differentiation and hybrid success in *Campanula americana*: geography and genome size. *Journal of Evolutionary Biology* **18:** 81-89.

Gavin, D.G., Fitzpatrick, M.C., Gugger, P.F., Heath, K.D., Rodriguez-Sanchez, F., Dobrowski, S.Z., Hampe, A., Hu, F.S., Ashcroft, M.B., Bartlein, P.J., Blois, J.L., Carstens, B.C., Davis, E.B., de Lafontaine, G., Edwards, M.E., Fernandez, M., Henne, P.D., Herring, E.M., Holden, Z.A., Kong, W.S., Liu, J., Magri, D., Matzke, N.J., McGlone, M.S., Saltre, F., Stigall, A.L., Tsai, Y.H. & Williams, J.W. 2014. Climate refugia: joint inference from fossil records, species distribution models and phylogeography. *The New Phytologist* **204:** 37-54.

Gomez, A., Hughes, R.N., Wright, P.J., Carvalho, G.R. & Lunt, D.H. 2007. Mitochondrial DNA phylogeography and mating compatibility reveal marked genetic structuring and speciation in the NE Atlantic bryozoan *Celleporella hyalina*. *Molecular Ecology* **16:** 2173-2188.

Gonzales, E., Hamrick, J.L. & Chang, S.M. 2008. Identification of glacial refugia in south-eastern North America by phylogeographical analyses of a forest understorey plant, *Trillium cuneatum*. *Journal of Biogeography* **35:** 844-852.

Griffin, S.R. & Barrett, S.C.H. 2004. Post-glacial history of *Trillium grandiflorum* (Melanthiaceae) in eastern North America: inferences from phylogeography. *American Journal of Botany* **91:** 465-473.

Guindon, S. & Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52:** 696-704.

Hall, T.A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* **41:** 95-98.

Hamilton, M.B. 1999. Four primer pairs for the amplification of chloroplast intergenic regions with intraspecific variation. *Molecular Ecology* **8:** 521-523.

Hewitt, G. 2000. The genetic legacy of the Quaternary ice ages. *Nature* **405:** 907-913.

Hewitt, G.M. 1996. Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society* **58:** 247-276.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25:** 1965-1978.

Langmead, B. & Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9:** 357-359.

Lee-Yaw, J.A., Irwin, J.T. & Green, D.M. 2008. Postglacial range expansion from northern refugia by the wood frog, *Rana sylvatica*. *Molecular Ecology* **17:** 867-884.

Levy, F. 1991. A genetic analysis of reproductive barriers in *Phacelia dubia*. *Heredity* **67:** 331-345.

Levy, F. & Neal, C.L. 1999. Spatial and temporal genetic structure in chloroplast and allozyme markers in *Phacelia dubia* implicate genetic drift. *Heredity* **82:** 422-431.

Li, P., Li, M.M., Shi, Y., Zhao, Y.P., Wan, Y., Fu, C.X. & Cameron, K.M. 2013. Phylogeography of North American herbaceous *Smilax* (Smilacaceae): combined

AFLP and cpDNA data support a northern refugium in the Driftless Area. *American Journal of Botany* **100:** 801-814.

McLachlan, J.S., Clark, J.S. & Manos, P.S. 2005. Molecular indicators of tree migration capacity under rapid climate change. *Ecology* **86:** 2088-2098.

Morris, A.B., Ickert-Bond, S.M., Brunson, D.B., Soltis, D.E. & Soltis, P.S. 2008. Phylogeographical structure and temporal complexity in American sweetgum (*Liquidambar styraciflua*; Altingiaceae). *Molecular Ecology* **17:** 3889-3900.

Parks, C.R., Wendel, J.F., Sewell, M.M. & Qiu, Y.L. 1994. The significance of allozyme variation and introgression in the *Liriodendron tulipifera* complex (Magnoliaceae). *American Journal of Botany* **81:** 878-889.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* **190:** 231-259.

Pinheiro, F., Cozzolino, S., de Barros, F., Gouveia, T.M.Z.M., Suzuki, R.M., Fay, M.F. & Palma-Silva, C. 2013. Phylogeographic structure and outbreeding depression reveal early stages of reproductive isolation in the neotropical orchid *Epidendrum denticulatum*. *Evolution* **67:** 2024-2039.

Prendeville, H.R., Barnard-Kubow, K., Dai, C., Barringer, B.C. & Galloway, L.F. 2013. Clinal variation for only some phenological traits across a species range. *Oecologia* **173:** 421-430.

Provan, J. & Bennett, K.D. 2008. Phylogeographic insights into cryptic glacial refugia. *Trends in Ecology & Evolution* **23:** 564-571.

Raineri, E., Ferretti, L., Esteve-Codina, A., Nevado, B., Heath, S. & Perez-Enciso, M. 2012. SNP calling by sequencing pooled samples. *BMC Bioinformatics* **13:** 239.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A. & Huelsenbeck, J.P. 2012. MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61:** 539-542.

Sheppard, C.S. 2013. How does selection of climate variables affect predictions of species distributions? A case study of three new weeds in New Zealand. *Weed Research* **53:** 259-268.

Singhal, S. & Moritz, C. 2013. Reproductive isolation between phylogeographic lineages scales with divergence. *Proceedings of the Royal Society B-Biological Sciences* **280**.

Soltis, D.E., Morris, A.B., McLachlan, J.S., Manos, P.S. & Soltis, P.S. 2006. Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology* **15:** 4261-4293.

Sorrie, B.A. & Weakley, A.S. 2001. Coastal plain vascular plant endemics: phytogeographic patterns. *Castanea* **66:** 50-82.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*.

Stewart, J.R. & Lister, A.M. 2001. Cryptic northern refugia and the origins of the modern biota. *Trends in Ecology & Evolution* **16:** 608-613.

Swets, J.A. 1988. Measuring the Accuracy of Diagnostic Systems. *Science* **240:** 1285-1293.

Taberlet, P., Gielly, L., Pautou, G. & Bouvet, J. 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology* **17:** 1105-1109.

Taberlet, P., Fumagalli, L., Wust-Saucy, A.-G. & Cosson, J.-F. 1998. Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology* **7:** 453-464.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28:** 2731-2739.

Tribsch, A. & Schonswetter, P. 2003. Patterns of endemism and comparative phylogeography confirm palaeoenvironmental evidence for Pleistocene refugia in the Eastern Alps. *Taxon* **52:** 477-497.

Walker, M.J., Stockman, A.K., Marek, P.E. & Bond, J.E. 2009. Pleistocene glacial refugia across the Appalachian Mountains and coastal plain in the millipede genus *Narceus*: Evidence from population genetic, phylogeographic, and paleoclimatic data. *BMC Evolutionary Biology* **9**.

Warren, D.L. & Seifert, S.N. 2011. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecological Applications* **21:** 335-342.

Wolfe, K.H., Li, W.H. & Sharp, P.M. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences* **84:** 9054-9058.

Table 1: Information for *C. americanum* populations sampled including location, chloroplast haplotype, and whether populations were included in nuclear RAD sequencing.

Pop ID	State	Latitude	Longitude	Haplotype	RAD
29	Alabama	34.6514	-86.5017	M/L	yes
69	Alabama	31.8485	-86.6402	I	
58	Arkansas	35.9106	-92.6316	H	
16	Florida	30.7690	-85.2419	I	
83	Florida	30.5647	-84.9598	I	yes
22	Georgia	34.4681	-84.4298	K	yes
20	Illinois	38.6074	-89.8992	G	yes
68	Indiana	39.1643	-86.5269	L	yes
10	Iowa	42.0728	-93.6725	G	yes
25	Iowa	41.6884	-93.7121	G	
45	Iowa	42.7554	-96.6069	H	
60	Kansas	39.0474	-95.6815	H	
51	Kentucky	37.9340	-84.2595	L	
5	Maryland	39.6137	-79.1158	D	yes
44	Michigan	42.2909	-85.5942	L	yes
12	Minnesota	44.8178	-93.0075	G	
38	Minnesota	44.8212	-93.3187	G	
55	Mississippi	31.7433	-88.5239	I/J	yes
70	Mississippi	33.9842	-88.4882	K	
7	Missouri	37.1328	-91.2772	H	
50	Missouri	39.1465	-92.6842	G	
59	Nebraska	41.1609	-96.5393	H	yes
36	North Carolina	36.1850	-81.6683	I	yes
90	North Carolina	35.7669	-82.1636	I	yes
91	North Carolina	35.5862	-83.0663	A	yes
11	Ohio	39.2718	-84.2841	L	
64	Ohio	41.1147	-81.5181	L	
61	Oklahoma	33.9464	-94.5669	H	yes
13	Pennsylvania	40.2721	-79.8850	D	
27	Pennsylvania	41.0079	-80.0833	K	yes
94	Pennsylvania	41.4670	-80.0113	L	
95	Pennsylvania	40.4752	-78.2808	C	
102	Pennsylvania	40.3225	-80.1109	D	
87	Tennessee	36.1211	-82.3178	J	
88	Tennessee	35.9833	-82.4989	J	yes
92	Tennessee	35.6758	-83.5259	B	yes
19	Tennessee	35.7583	-88.0687	L	
32	Tennessee	35.0748	-85.6250	L	
34	Tennessee	36.0822	-86.2961	N	yes
71	Virginia	38.3305	-78.4901	E/F	yes
73	Virginia	37.3534	-80.5522	D	yes
74	Virginia	37.9500	-78.8900	E	yes
85	Virginia	37.7576	-79.1876	E	yes
86	Virginia	36.6344	-81.5881	I	yes
93	Virginia	37.2067	-76.9476	E	yes
100	Virginia	37.2774	-80.6126	D	
72	West Virginia	37.9931	-80.3618	D	yes
84	West Virginia	38.2266	-80.2231	D	
14	Wisconsin	43.3362	-89.9467	G	

Figure 1: A) Maximum likelihood tree and B) statistical parsimony network for concatenated chloroplast sequence data from *Campanulastrum americanum*. Numbers show bootstrap support and Bayesian posterior probabilities ("-" Bayesian posterior probabilities below 50). In the statistical parsimony network the colored circles represent individual haplotypes, while lines represent single mutations, and the black tick marks represent unobserved haplotypes. The sizes of the circles in the statistical parsimony network are scaled according to how frequently that haplotype was observed (1-9x). C) Map shows location of populations with grey shaded area representing an approximation of *C. americanum's* species range. Populations colored according to chloroplast haplotype, with pie charts representing polymorphic populations.

Figure 1:

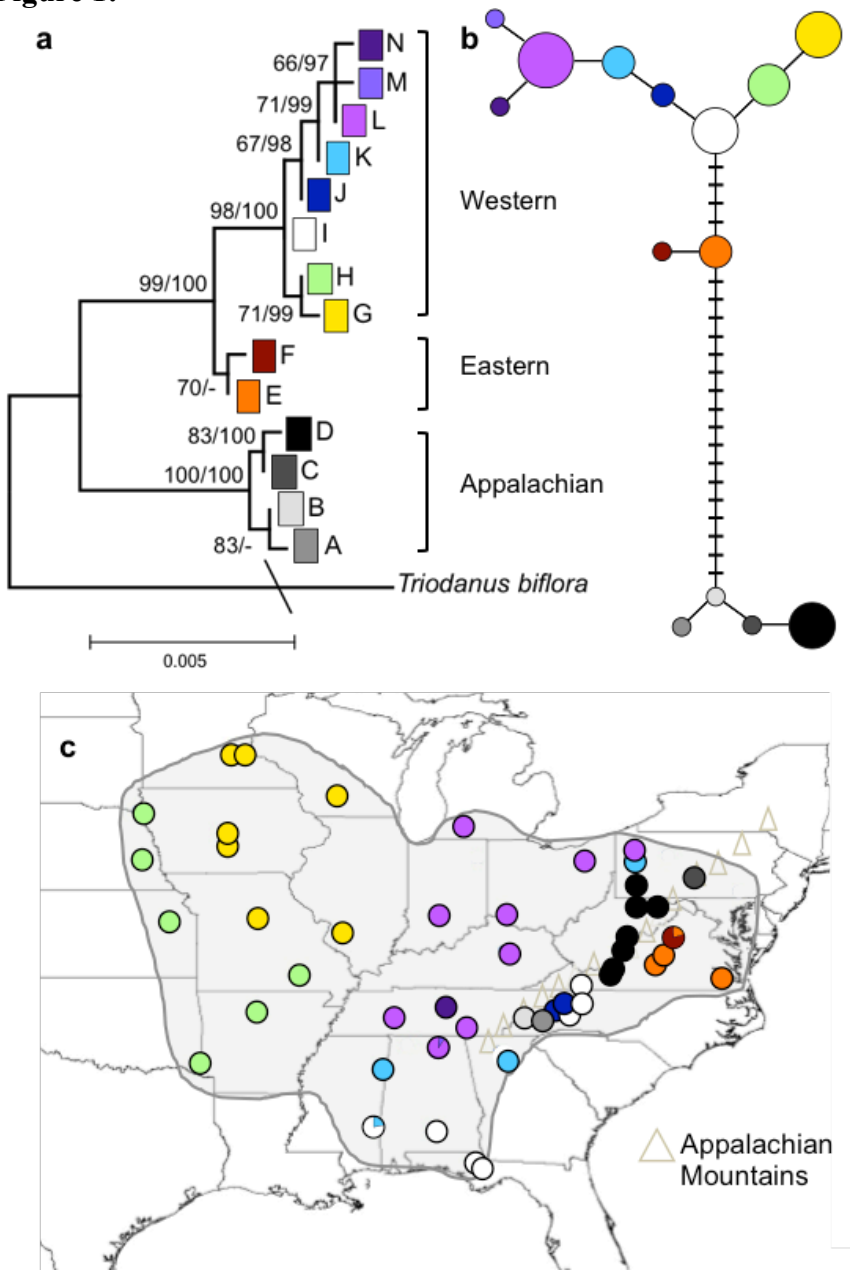

a
66/97
71/99
67/98
98/100
99/100
71/99
70/-
83/100
100/100
83/-
N
M
L
K
J
I
H
G
F
E
D
C
B
A
Western
Eastern
Appalachian
Triodanus biflora
0.005
b
c
Appalachian
Mountains

Figure 2: Maximum likelihood tree for *Campanulastrum americanum* nuclear data from reference based Stacks analysis. Rooted according to position of root in chloroplast tree. Numbers show bootstrap support values above 60. Population names match Table 1 and refer to a population's state and ID. Colors refer to chloroplast haplotype as in Figure 1.

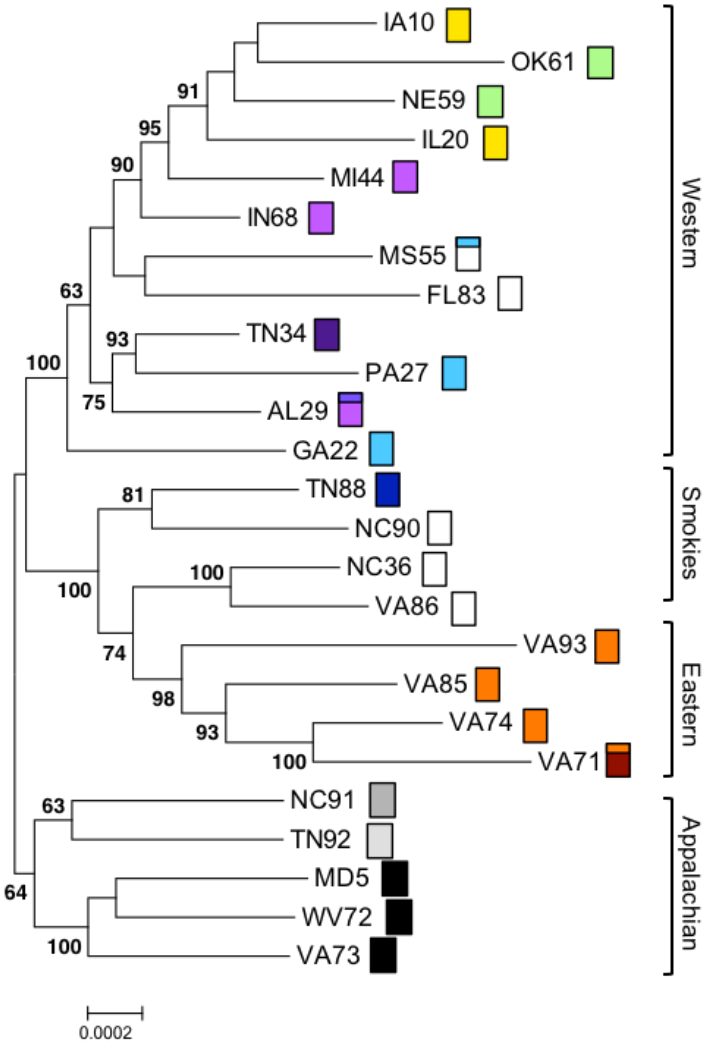
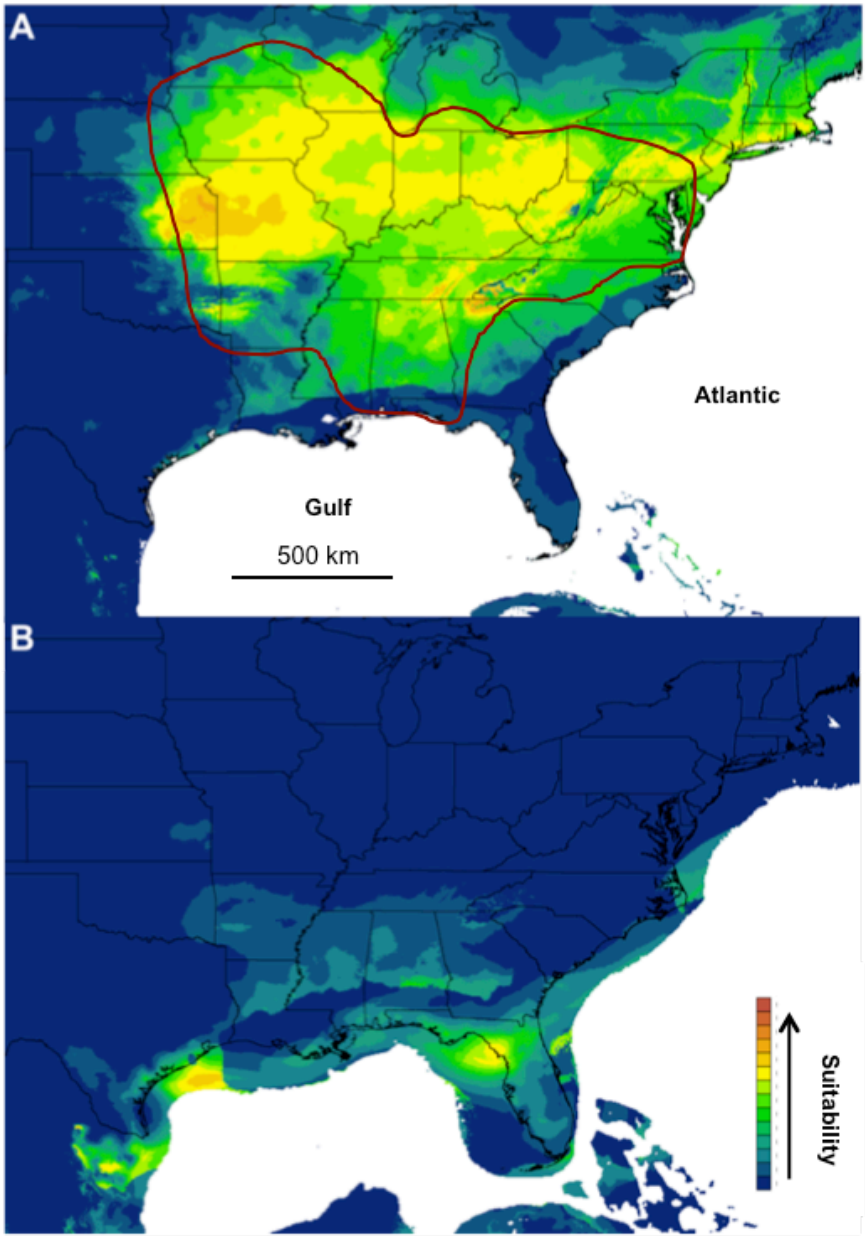


Figure 3: Species distribution models for *Campanulastrum americanum* showing predicted habitat suitability during (a) present day and (b) the last glacial maximum, ~20,000 years ago. Log suitability is expressed between 0 and 1, and higher values (here in warmer colors) indicate an increased probability of finding *C. americanum*. Dark red line represents and approximation of *C. americanum's* species range.



Supplementary Table 1: Information for the five loci used in construction of the chloroplast phylogeny, including primer sequences, and number of SNPs and haplotypes observed.

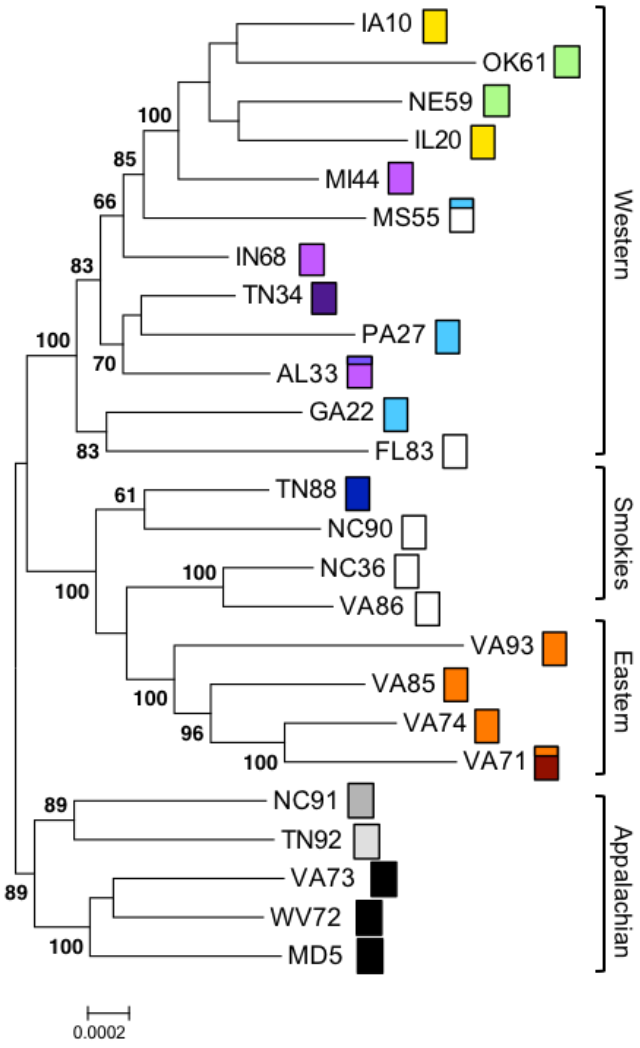
Locus	Location	Primer Names	Primer Sequences	Product Length (bp)	#SNPs	#Haplotypes
LF	trnL intron	CamL1	TCAAATTCAGAGAAACCCTG	587	5	4
		L1("C")*	CGAAATCGGTAGACGCTACG			
		L2(D")*	GGGGATAGAGGGACTTGAAC			
	trnL-trnF intergenic spacer	trnL("E")*	GGTTCAAGTCCCTCTATCCC			
		trnF("F")*	ATTTGAACTGGTGACACGAG			
		CamF	TCCTGACTATTCCCGATGTT			
C11	ndhB-trnI intergenic spacer	C11F	TACCCTGGTGATAAGGGGGCCTA	288	4	3
		C11R	TTCAACATCGAAGAGGGTCCACGTT			
ycf1	ycf1 partial cds	C308F	CCTTGCTCGCGCCACTTTCG	429	6	3
		C308R	ACACGGGTTTGGGCCGATCC			
rps2	rps2 partial cds	C343F	TGTCTTGGCCATGGATCTAGGGCT	600	10	9
		C343R	CGGCCCTCACGAATTGCAAATACT			
rps4	rps4 partial cds	cprps4F	TGTCACGTTACCGAGGACCTTKT	533	8	5
		cprps4R	AGTGGTCAAAGTTGTCCGCACCT			

* Indicates primers from Taberlet et al (1991).

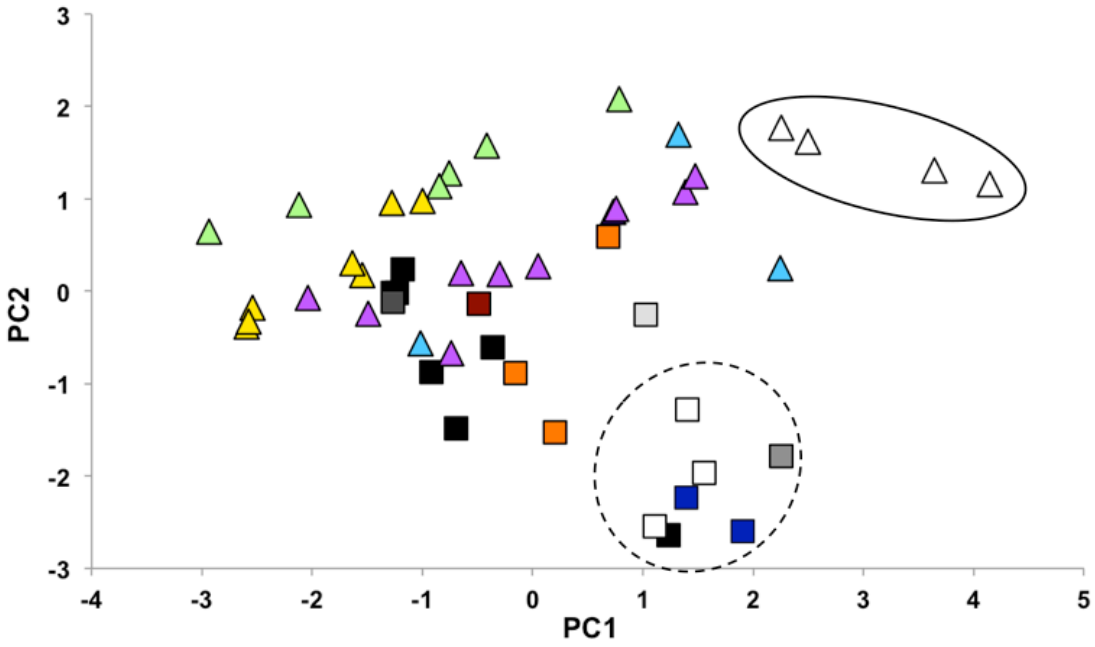
Supplementary Table 2: Sequencing information for populations in which multiple individuals were sequenced for chloroplast haplotype including the number of individuals sequenced at all five loci, the number of individuals sequenced at a subset of loci, which subset of loci were used, and whether or not populations were found to be polymorphic.

POP ID	State	Choroplast haplotype	# Sequenced at all loci	# Sequenced at a subset of loci	Subset of loci sequenced	Polymorphic
29	Alabama	M(1)/L(9)	1	8	C343 and rps4	yes
69	Alabama	I	8	0		no
16	Florida	I	1	8	C343 and C11	no
83	Florida	I	5	0		no
55	Mississippi	I(7)/J(2)	6	2	C11 and C343	yes
90	North Carolina	I	0	7	C343 and C11	no
91	North Carolina	A	0	8	C11	no
88	Tennessee	J	0	7	C343 and C11	no
71	Virginia	E(2)/F(8)	0	10	C343 and C11	yes
73	Virginia	D	5	0		no
85	Virginia	E	0	6	C343 and C11	no
86	Virginia	I	0	7	C343 and C11	no
72	West Virginia	D	0	14	C308 and rps4	no

Supplementary Figure 1: Maximum likelihood tree for nuclear data from denovo based Stacks analysis. Rooted according to position of root in chloroplast tree. Numbers show bootstrap support values above 60. Population names match Table 1 and refer to a population's state and ID. Colors refer to chloroplast haplotype as in Figure 1.



Supplementary Figure 2: Principal components analysis of 19 bioclimatic variables for the study populations. PC1 and PC2 explain 60.42 and 21.77 percent of variance respectively. Color refers to chloroplast haplotype as in Figure 1 and shape refers to whether or not populations are Appalachian, based on nuclear clade (triangle=western and square=eastern/mountain). The solid circle marks the deep-south populations while the dashed circle marks the Smokies populations that fall out in the eastern-nuclear, but western-chloroplast clade.



Chapter 2:

Cytonuclear incompatibility contributes to the early stages of speciation

This chapter is formatted for a journal where the detailed methods are listed at the end.

Abstract

In recent years, there has been a growing interest in the contribution of cytonuclear incompatibilities to speciation and the idea that they may be among some of the first genetic incompatibilities to arise. However, clear examples of cytonuclear incompatibilities contributing to reproductive isolation within species are relatively rare. We show evidence for a strong cytonuclear incompatibility leading to substantial postzygotic reproductive isolation (up to 81% reduction in survival) in first generation hybrids between differentiated populations of an herbaceous plant, with the strength of cytonuclear incompatibility linearly related to genetic distance between populations (r^2=0.87). Accelerated chloroplast genome evolution in this species, a phenomenon found in multiple independent lineages of angiosperms, may have facilitated development of this cytonuclear incompatibility.

Introduction

Understanding genetic incompatibility that arises within a species, leading to reproductive isolation (RI), is one of the primary goals of speciation research. While historically studies have focused on incompatibilities between nuclear loci, there has been a growing realization of the potential importance of cytonuclear incompatibilities for the speciation process as they have been found to contribute to RI and speciation in plants (e.g. Levy 1991; Fishman & Willis, 2006; Sambatti *et al.*, 2008; Greiner *et al.*, 2011; Leppala & Savolainen, 2011), yeast (reviewed in Chou & Leu, 2010), and animals (e.g. Ellison & Burton, 2008).

Increasing interest in the potential importance of cytonuclear incompatibilities is also apparent in the recent publication of several review papers detailing examples of these incompatibilities and the factors that predispose them to be strong contributors to RI and speciation (Levin, 2003; Greiner *et al.*, 2011; Burton & Barreto, 2012; Burton *et al.*, 2013; Greiner & Bock, 2013). These factors include the likelihood of cytonuclear coevolution, as the two genomes must interact closely to carry out essential processes such as respiration and photosynthesis (Rand *et al.*, 2004), an increased accumulation of deleterious mutations in organelle genomes due to their reduced population size and lack of normal recombination (Birky, 2001), and the potential for intergenomic conflict (e.g. cytoplasmic male sterility, Chase, 2007). Due to these factors, cytonuclear incompatibilities have been proposed be among the first genetic incompatibilities to arise, thereby influencing the earliest stages of the speciation process (Levin 2003; Fishman and Willis, 2006; Burton *et al.*, 2013). However, there are relatively few examples of cytonuclear incompatibilities at early stages, in particular within a species, as few studies

characterizing genetic incompatibilities have focused at the intraspecific level (but see Burton *et al.*, 2013).

We characterized genetic incompatibilities arising early in the speciation process by measuring range-wide patterns of RI among populations of the herb, *Campanulastrum americanum*. Previous studies found substantial postzygotic RI in a limited sample of *C. americanum* populations (Galloway & Etterson, 2005; Etterson *et al.*, 2007). Asymmetrical patterns of RI, along with observations of chlorosis (insufficient production of chlorophyll), suggested a potential role for cytonuclear incompatibility. Building on these results, we estimated postzygotic RI across the life cycle including germination, survival to reproduction, and reproductive traits using first-generation (F1) hybrids from pairwise crosses between 29 populations spread across the species range. We also examined whether geographic or genetic distance between parental populations could predict patterns of RI, giving insight into the underlying genetic incompatibilities. We further ask: 1) Which traits show RI? 2) Is RI caused by cytonuclear incompatibility? 3) If RI is found for multiple traits, are the underlying genetic incompatibilities independent of one another?

Methods

C. americanum is an autotretraploid, monocarpic herb found in the eastern half of the United States (Fig 1). Individuals are annual or biennial, insect-pollinated, and highly outcrossing (Galloway *et al.*, 2003). *C. americanum* typically grows in disturbed habitats and there seeds are dispersed passively, traits that likely contribute to its patchy population structure. Chloroplast and nuclear markers resolve four genetic clades:

Western, Smoky Mountains (Smokies), Appalachian Mountains (Appalachian), and Eastern (Chapter 1). The Appalachian clade is the oldest clade, with the Eastern, Smokies, and Western clades progressively more derived. These clades differ in their spread with the Western clade found through much of the species range, and the other three primarily restricted to the Appalachian Mountains (Chapter 1).

We carried out 28 pairwise crosses among the 29 populations sampled from across the species range (Fig 1), 10 of which were between-clades, and grew the hybrid offspring under common greenhouse conditions. Multiple genotypes from each population were used for crossing between populations to produce hybrid seed, as well as within populations to produce parental population seed (Supp. Methods). To determine levels of postzygotic RI, seed was grown under controlled conditions and germination, survival to reproduction, an index of flower number, seed number, and pollen viability were measured as fitness components. Cumulative fitness was calculated as a product of all traits (Supp. Methods). Relative hybrid performance was calculated for each cross/trait combination as the percent deviation in performance of the hybrids from the mean of the parental populations (midparent) [((hybrid-midparent)/ midparent)*100]. Negative values for this index indicate RI, also termed hybrid breakdown. We then examined whether variation in strength of RI could be explained by geographic, chloroplast-genetic (# SNPs between parental chloroplast haplotypes), or nuclear-genetic (maximum likelihood distance based on RAD-seq data) distance between parental populations (Supp. Methods).

Results and Discussion

Substantial postzyogtic RI was found in crosses between-clades, with up to a 92% reduction in cumulative fitness in F1 hybrids. No RI was found in within-clade crosses, results substantiated in subsequent work (Chapter 3). The two viability traits, germination and survival, showed strong RI. Geographic distance between populations was a good predictor of RI for germination (r^2=0.70, Table 1, Fig 2A), with larger distances leading to a reduction in germination. In contrast, chloroplast-genetic distance predicted RI for survival (r^2=0.87, Table 1, Fig 2D), with decreased survival at larger distances. Prediction of RI for germination and survival by different non-correlated measures of distance (geographic versus chloroplast-genetic distance, r=-0.01, p=0.98) (Supp. Methods), suggests different genetic incompatibilities underlie isolation in these traits. When evaluating individual crosses, RI for these traits also occurred independently of one another (Supp. Table 2), with some crosses showing reduced germination, but not survival, and vice versa, providing further support that separate mechanisms underlie RI in these traits. The presence of multiple, independent mechanisms leading to RI in *C. americanum* suggests that even at the earliest stages speciation can be a complex process, involving multiple genes and incompatibilities.

Reproductive traits exhibited little to no RI. RI was only found for pollen viability, where about half of within-and between-clade crosses exhibited modest fitness reductions (up to 25%) (Supp. Figure 1, Supp. Table 2). In contrast to RI for survival and germination, RI for pollen viability was not associated with geographic or genetic distance (Table 1). In total, between-population RI in *C. americanum* primarily impacts viability, with individuals that survive to flower having only a moderate reduction in reproduction.

RI for viability traits in between-clade crosses was frequently asymmetrical, with the strength of RI dependent upon crossing direction (Fig 3). For germination, six of ten between-clade crosses exhibited asymmetrical RI, while the total was seven of ten for survival (Supp. Table 2). The difference between crossing directions increased with chloroplast-genetic distance for survival (r^2=0.51, p=0.02), but not with geographic distance for germination (r^2=0.00, p=0.99) and was stronger for survival than germination (mean difference of 38% versus 22%; Fig 3), although this difference was not significant (t=-1.58, p=0.15). The asymmetry in survival was the result of stronger RI when populations with more derived chloroplasts were maternal (up to 81% reduction in survival; t=-3.63, p=0.01), although low to moderate RI also occurred in the other crossing direction (up to 44% reduction in survival) (Supp. Table 2).

Asymmetrical patterns of RI for post-zygotic traits are often considered a hallmark of cytonuclear incompatibilities (Tiffin *et al.*, 2001), as hybrids contain the same nuclear genetic makeup, but different cytoplasmic backgrounds. RI was strongly asymmetrical for survival in *C. americanum*, suggesting underlying cytonuclear incompatibility. The influence of chloroplast-genetic distance on RI for survival provides additional support for a cytoplasmic contribution. Chlorosis and variegation were also frequently observed in between-clade F1 hybrids (Fig 4) indicating that one of the parental chloroplasts was unable to function effectively on the hybrid nuclear background. A cytonuclear basis for this incompatibility was evidenced by chlorosis never being observed in parental populations. Chlorotic tissue was correlated with parental chloroplast haplotypes (Chapter 4), further confirming cytonuclear incompatibility underlies chlorosis and variegation, rather than other mechanisms, such

as an autoimmune response (e.g. Bomblies & Weigel, 2007). Finally, *C. americanum* has a rapidly evolving plastid genome (Chapter 5), which could lead to selection for rapid coevolution of the nuclear and plastid genome and an increased likelihood of cytonuclear incompatibility.

Although RI was also asymmetrical for germination, this pattern does not appear to reflect an underlying cytonuclear incompatibility. It was not predicted by chloroplast-genetic distance. In addition, backcross data for two sample crosses found recovery for germination in both maternal and paternal backcrosses (Supp. Fig 2) (Supp. Methods). Backcrossing F1 hybrids to parental populations leads to a partial restoration of parental nuclear gene combinations regardless of direction. However, depending upon backcross direction there is either a partial restoration (maternal) or an increased disruption (paternal) of parental cytonuclear gene combinations (e.g. Ellison & Burton, 2008). Therefore, if cytonuclear incompatibilities contribute to RI, the maternal, but not the paternal, backcross would be expected to show recovery, but this was not found. Instead there was complete recovery, regardless of direction, suggesting cytonuclear incompatibilities are not contributing to RI for germination. RI in germination is therefore likely to be caused by mechanisms such as triploid endosperm interactions or maternal-zygote incompatibilities, both of which are expected to cause asymmetric RI, similar to cytonuclear incompatibility (Turelli & Moyle, 2007). However, further work is needed to confirm which of these alternative mechanisms underlies RI in germination.

Cytonuclear incompatibility has been hypothesized to be one the earlier genetic incompatibilities to arise in the speciation process and to be widespread (Levin, 2003; Greiner *et al.*, 2011), but so far there are relatively few examples of cytonuclear

incompatibility contributing to intraspecific reproductive isolation (but see Burton *et al.*, 2013). Our results support a role for cytonuclear incompatibility at the early stages of the speciation process, as seen by substantial reductions in survival in among-population F1 hybrids of *C. americanum* that are strongly asymmetric and influenced by chloroplast-genetic distance. These results fit with findings in *Oenothera*, where plastid-nuclear incompatibility is the only genetic barrier between several recently diverged species (reviewed in Greiner *et al.*, 2011). However, strong RI for germination in *C. americanum* appears to be caused by triploid endosperm interactions or maternal-zygote incompatibilities, suggesting other incompatibilities may be of equal importance role early in speciation.

The presence of such a strong intraspecific cytonuclear incompatibility in *C. americanum* raises the question as to how it evolved. *C. americanum* has a structurally unstable plastid genome (similar to other members of the Campanulaceae, Cosner *et al.*, 2004; Haberle *et al.*, 2008; Knox, 2014), as well as increased nucleotide substitution rates (Chapter 5), two traits frequently shown to co-occur (Jansen *et al.*, 2007). We hypothesize that this accelerated plastid genome evolution underlies the cytonuclear incompatibility found in *C. americanum*. Similar accelerated plastid evolution has been found in taxa from other independent lineages of Angiosperms (Jansen *et al.*, 2007), raising the possibility of intraspecific cytonuclear incompatibilities in these taxa. Cytonuclear incompatibilities resulting in chlorosis are known from interspecific crosses in *Oenothera* (reviewed in Greiner *et al.*, 2011) and *Pelargonium* (Metzlaf *et al.*, 1982; Weihe *et al.*, 2009). Future work characterizing genetic incompatibilities at the intraspecific level in taxa with accelerated plastid evolution could help reveal how

widespread the patterns observed in *C. americanum* are, and whether accelerated plastid evolution may help drive the early stages of the speciation process.

Acknowledgements

We thank Nina So for helping to collect the backcross data, and Shay Nimjareansuk, Can Dai, Francis Kilkenny, Prajakta Bhayade, Tim Park, Karla Platzer, Jie Ren, Lauren Wilson, and Remington Wong for assistance with greenhouse work. Janis Antonovics, Ben Blackman, Edmund Brodie, David Carr, and Douglas Taylor provided helpful discussion and comments. This work was supported by the ARCS Foundation.

References

Birky, C.W. 2001. The inheritance of genes in mitochondria and chloroplast: laws, mechanisms, and models. *Annual Review of Genetics* **35:** 125-148.

Bomblies, K. & Weigel, D. 2007. Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. *Nature Reviews Genetics* **8:** 382-393.

Burton, R.S. & Barreto, F.S. 2012. A disproportionate role for mtDNA in Dobzhansky-Muller incompatibilities? *Molecular Ecology* **21:** 4942-4957.

Burton, R.S., Pereira, R.J. & Barreto, F.S. 2013. Cytonuclear genomic interactions and hybrid breakdown. *Annual Review of Ecology, Evolution, and Systematics, Vol 44* **44:** 281-302.

Chase, C.D. 2007. Cytoplasmic male sterility: a window to the world of plant mitochondrial-nuclear interactions. *Trends in Genetics* **23:** 81-90.

Chou, J.Y. & Leu, J.Y. 2010. Speciation through cytonuclear incompatibility: Insights from yeast and implications for higher eukaryotes. *Bioessays* **32:** 401-411.

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

Ellison, C.K. & Burton, R.S. 2008. Interpopulation hybrid breakdown maps to the mitochondrial genome. *Evolution* **62:** 631-638.

Etterson, J.R., Keller, S.R. & Galloway, L.F. 2007. Epistatic and cytonuclear interactions govern outbreeding depression in the autotetraploid *Campanulastrum americanum*. *Evolution* **61:** 2671-2683.

Fishman, L. & Willis, J.H. 2006. A cytonuclear incompatibility causes anther sterility in *Mimulus* hybrids. *Evolution* **60:** 1372-1381.

Galloway, L.F., Etterson, J.R. & Hamrick, J.L. 2003. Outcrossing rate and inbreeding depression in the herbaceous autotetraploid, *Campanula americana*. *Heredity* **90:** 308-315.

Galloway, L.F. & Etterson, J.R. 2005. Population differentiation and hybrid success in *Campanula americana*: geography and genome size. *Journal of Evolutionary Biology* **18:** 81-89.

Greiner, S., Rauwolf, U., Meurer, J. & Herrmann, R.G. 2011. The role of plastids in plant speciation. *Molecular Ecology* **20:** 671-691.

Greiner, S. & Bock, R. 2013. Tuning a menage a trois: co-evolution and co-adaptation of nuclear and organellar genomes in plants. *Bioessays* **35:** 354-365.

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. *Journal of Molecular Evolution* **66:** 350-361.

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., Depamphilis, C.W., Leebens-Mack, J., Muller, 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. *Proceedings of the National Academy of Sciences* **104:** 19369-19374.

Knox, E.B. 2014. The dynamic history of plastid genomes in the Campanulaceae sensu lato is unique among angiosperms. *Proceedings of the National Academy of Sciences* **111:** 11097-11102.

Leppala, J. & Savolainen, O. 2011. Nuclear-cytoplasmic interactions reduce male fertility in hybrids of *Arabidopsis lyrata* subspecies. *Evolution* **65:** 2959-2972.

Levin, D.A. 2003. The cytoplasmic factor in plant speciation. *Systematic Botany* **28:** 5-11.

Levy, F. 1991. A genetic analysis of reproductive barriers in *Phacelia dubia*. *Heredity* **67:** 331-345.

Metzlaf, M., Pohlheim, F., Börner, T. & Hagemann, R. 1982. Hybrid variegation in the genus *Pelargonium*. *Current genetics* **5:** 245-249.

Rand, D.M., Haney, R.A. & Fry, A.J. 2004. Cytonuclear coevolution: the genomics of cooperation. *Trends in Ecology & Evolution* **19:** 645-653.

Sambatti, J.B.M., Ortiz-Barrientos, D., Baack, E.J. & Rieseberg, L.H. 2008. Ecological selection maintains cytonuclear incompatibilities in hybridizing sunflowers. *Ecology Letters* **11:** 1082-1091.

Tiffin, P., Olson, M.S. & Moyle, L.C. 2001. Asymmetrical crossing barriers in angiosperms. *Proceedings of the Royal Society of London Series B-Biological Sciences* **268:** 861-867.

Turelli, M. & Moyle, L.C. 2007. Asymmetric postmating isolation: Darwin's corollary to Haldane's rule. *Genetics* **176:** 1059-1088.

Weihe, A., Apitz, J., Pohlheim, F., Salinas-Hartwig, A. & Borner, T. 2009. Biparental inheritance of plastidial and mitochondrial DNA and hybrid variegation in *Pelargonium*. *Molecular Genetics and Genomics* **282:** 587-593.

Table 1: ANCOVA testing whether variation in hybrid performance is determined by genetic divergence (crossing within or between clades), and magnitude of geographic, chloroplast-genetic, or nuclear-genetic distance between parental populations of *C. americanum*. Cumulative fitness was calculated as (germination*survival*flower #*((seed # +pollen viability)/2)). *F*-values are given, with values in bold significant at $P < 0.05$.

Trait	Genetic Divergence	Geographic Distance	Gen Div *Geo Dist	Chloroplast Distance	Gen Div *CP Dist	Nuclear Distance	Gen Div *Nuc Dist
Germination	1.58	**41.27*** **	**31.67*** **	0.86	0.07	1.19	1.53
Survival	0.78	1.42	1.20	**9.66** **	**5.46***	0.01	0.09
Flower number	0.27	2.78	0.04	0.08	0.60	0.67	0.41
Seed number	3.18	0.64	**5.31***	3.26[a]	3.14[a]	2.50	2.69
Pollen viability	0.31	0.63	0.58	0.07	0.33	0.08	0.75
Cumulative fitness	0.23	1.46	**6.21***	0.13	0.52	0.29	0.00

a $P < 0.10$; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$
d.f. 1,20

Figure 1: Location of *C. americanum* populations used to evaluate RI. Lines represent pairwise crosses carried out to examine patterns of reproductive isolation with blue and green representing within- and between-clade crosses respectively. Populations are shaded to differentiate the four clades (Supp. Table 1). The shaded area represents an approximation of *C. americanum's* range.

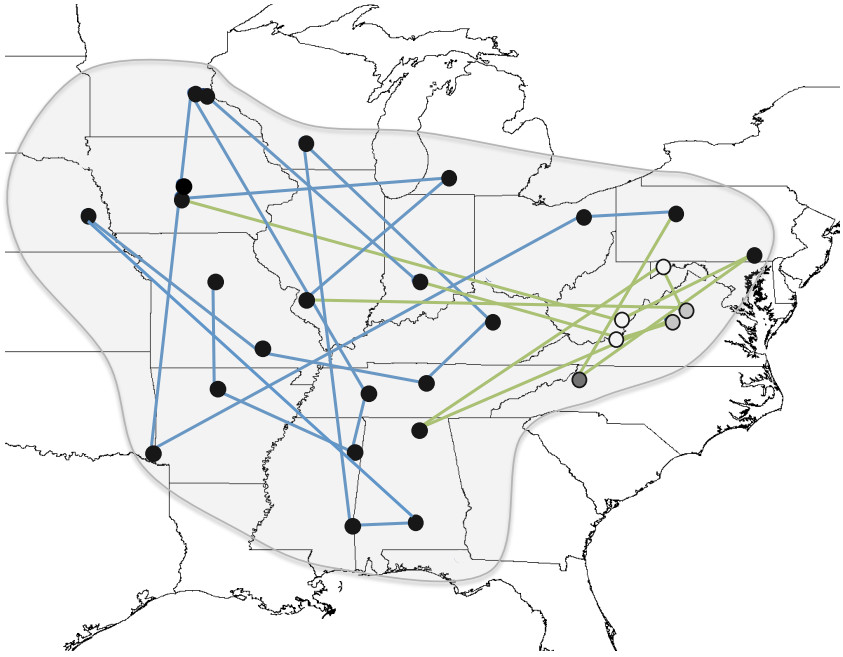


Figure 2: Performance of *C. americanum* hybrids relative to parental populations for germination (A,B) and survival (C,D) graphed against geographic and chloroplast-genetic distance between parental populations. Values were calculated as the percent reduction in hybrid performance relative to the parents, such that negative values represent RI. Each point represents a cross mean. r^2 obtained from regression analyses of between-clade relative hybrid performance against geographic or chloroplast genetic distance (Supp. Methods).

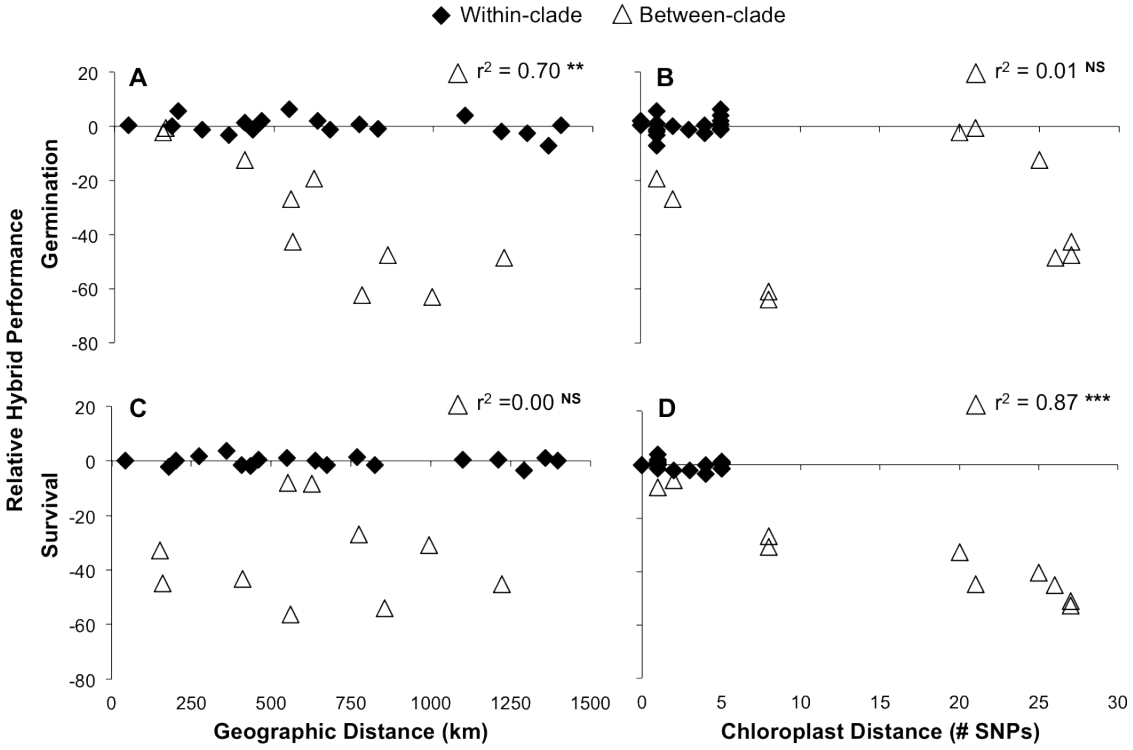


Figure 3: Performance of between-clade *C. americanum* hybrids relative to parental populations for each crossing direction for germination (A) and survival (B) graphed against geographic and chloroplast-genetic distance respectively. Negative values represent RI. Shading of triangles indicates crossing direction. r^2 values obtained from regression analyses of relative hybrid performance against geographic or chloroplast genetic distance for each crossing direction (Supp. Methods).

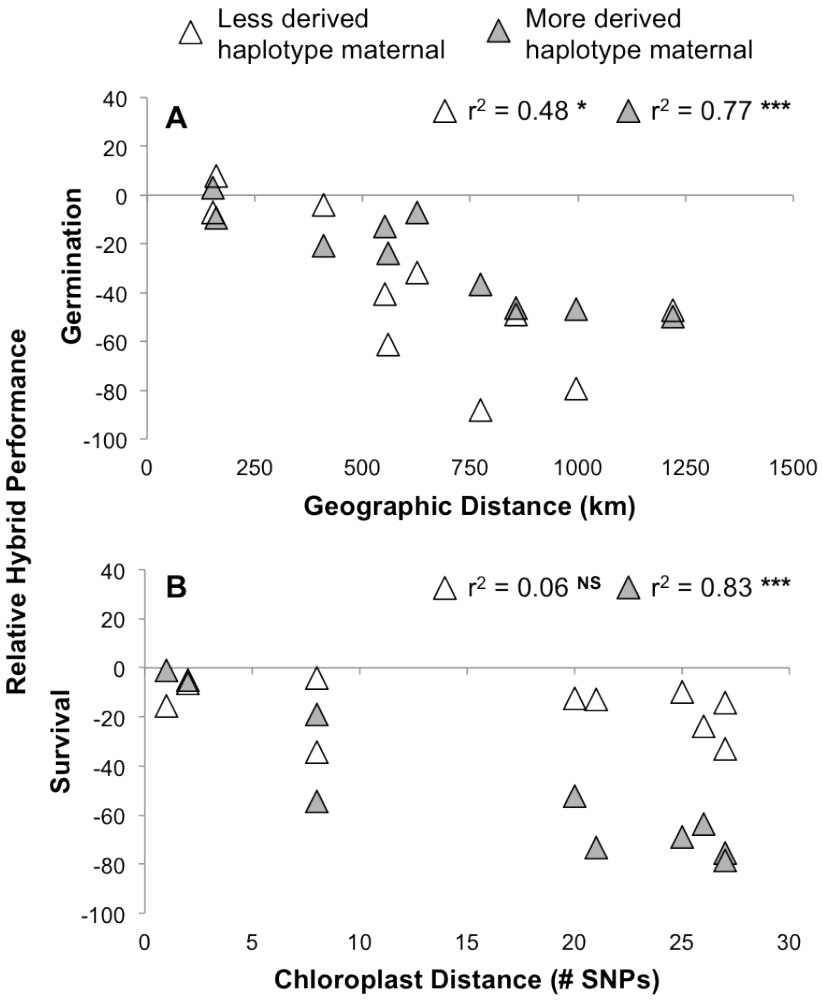
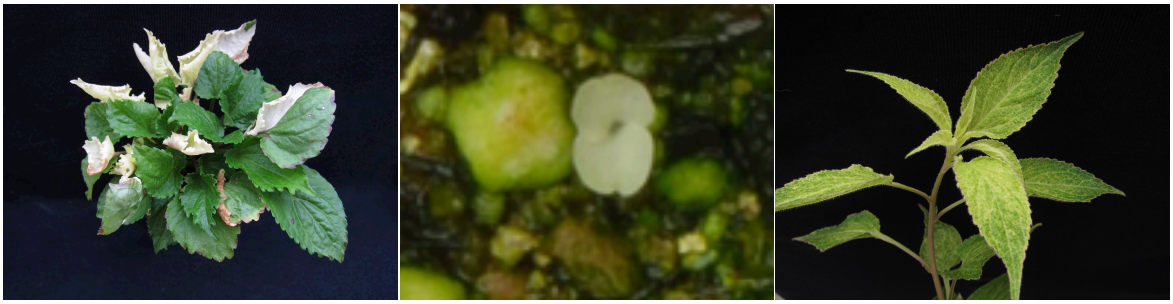


Figure 4: Examples of variegation and chlorosis observed in between-clade *C. americanum* hybrids.



Supplementary Table 1: Populations of *C. americanum* in the eastern United States used to evaluate RI including geographic location and phylogenetic clade.

Pop ID	State	Latitude	Longitude	Clade
5	Maryland	39.6137	-79.1158	Appalachian
73	Virginia	37.3534	-80.5522	Appalachian
72	West Virginia	37.9931	-80.3618	Appalachian
71	Virginia	38.3305	-78.4901	East
74	Virginia	37.9500	-78.8900	East
36	North Carolina	36.1850	-81.6683	Smokies
29	Alabama	34.6514	-86.5017	West
69	Alabama	31.8485	-86.6402	West
58	Arkansas	35.9106	-92.6316	West
20	Illinois	38.6074	-89.8992	West
68	Indiana	39.1643	-86.5269	West
10	Iowa	42.0728	-93.6725	West
25	Iowa	41.6884	-93.7121	West
51	Kentucky	37.9340	-84.2595	West
44	Michigan	42.2909	-85.5942	West
12	Minnesota	44.8178	-93.0075	West
38	Minnesota	44.8212	-93.3187	West
55	Mississippi	31.7433	-88.5239	West
70	Mississippi	33.9842	-88.4882	West
7	Missouri	37.1328	-91.2772	West
50	Missouri	39.1465	-92.6842	West
59	Nebraska	41.1609	-96.5393	West
64	Ohio	41.1147	-81.5181	West
61	Oklahoma	33.9464	-94.5669	West
27	Pennsylvania	41.0079	-80.0833	West
52	Pennsylvania	39.9406	-76.3464	West
19	Tennessee	35.7583	-88.0687	West
34	Tennessee	36.0822	-86.2961	West
14	Wisconsin	43.3362	-89.9467	West

Supplementary Table 2: Individual cross-level analyses of hybrid performance for each trait. Down arrows indicate crosses where hybrids performed significantly worse than their midparent (RI), while up arrows indicate significantly better performance (heterosis). In crosses where the reciprocal F1 hybrids were significantly asymmetrical, results for the reciprocals are shown separately, with the crossing direction where the more derived population was maternal represented by arrows on the left, and the other crossing direction by arrows to the right. Within clade crosses are organized by geographic distance between parental populations, with smaller distances at the top.

Crosstype	Clade	Germination	Survival	Flower #	Seed #	Pollen	Fitness	Cross
within	West							25x10
within	West							55x69
within	West							19x70
within	West							34x51
within	West							58x50
within	West							27x64
within	West			↑		↓		70x58
within	West				↓	↓*		7x34
within	West					↓*		20x44
within	West					↓*		59x7
within	West					↓*		44x25
within	West			↑		↓*		51x14
within	West					↓*		12x68
within	West			↑*			↑	38x19
within	West					↓*		61x38
within	West			↑*			↑	44x55
within	West	↓*			↑*			69x59
within	West							64x61
between	West x Smokies	↓	↓*	↑*		↓*		27x36
between	West x Smokies		↓*	↑*		↓*		52x36
between	East x App		↓*				↓	71x5
between	East x App		↓* ↓	↓*			↓*	74x73
between	West x East	↓	↓* ↓*			↓*	↓*	29x74
between	West x East	↓*	↓* ↓*				↓*	20x71
between	West x App	↓*	↓*			↓* ↓*	↓*	52x72
between	West x App	↓*	↓* ↓* ↓*			↑	↓*	68x73
between	West x App	↓*	↓* ↓			↓*	↓*	29x5
between	West x App	↓*	↓* ↓*			↓ ↓*	↓*	10x72

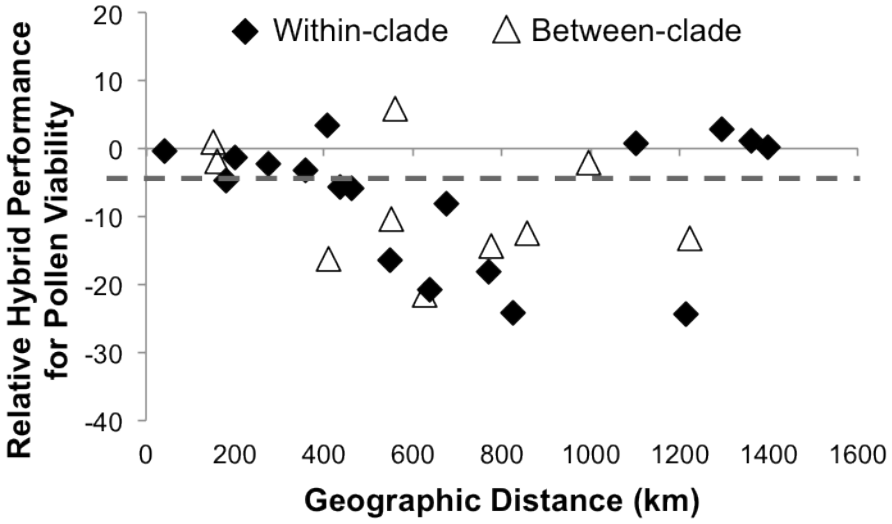
Grey: $P < 0.075$, black $P < 0.05$, * $P < 0.0125$ (significant after correction for multiple non-independent contrasts)

Supplementary Table 3: ANCOVA testing whether variation in hybrid performance in between-clade crosses is determined by crossing direction, and magnitude of geographic, chloroplast-genetic, or nuclear-genetic distance between parental populations of *C. americanum*. *F*-values (ANCOVA) are given, with values in bold significant at $P < 0.05$.

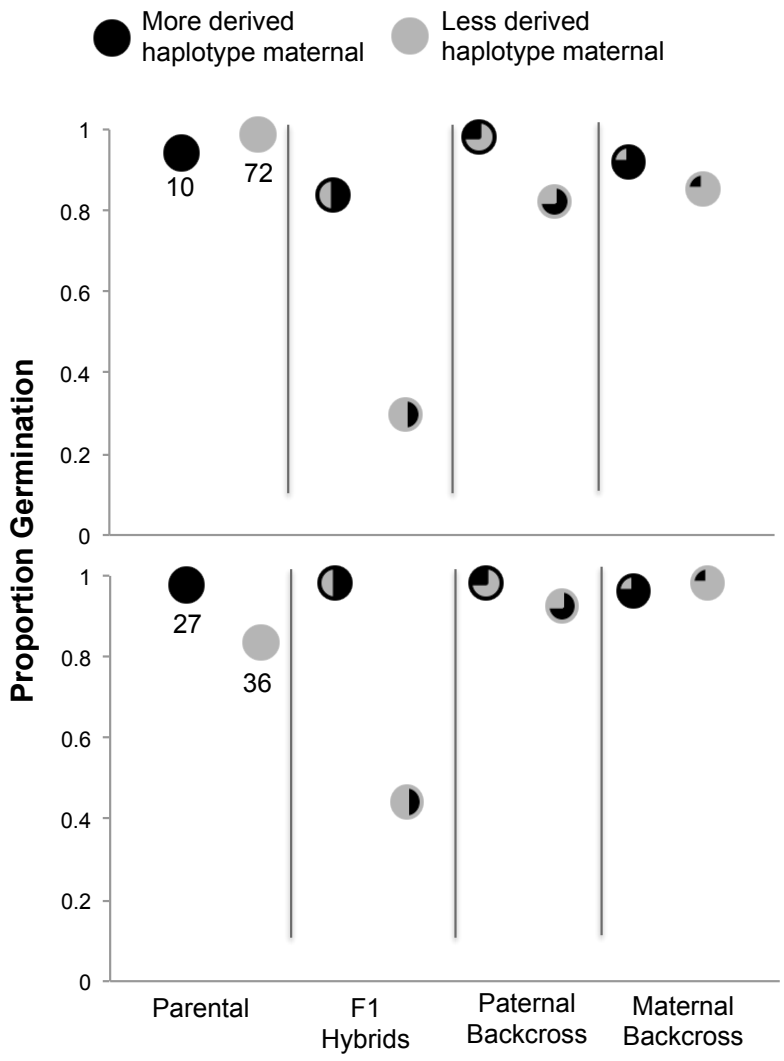
Trait	Direction	Geographic Distance	Direction* Geo Dist	Chloroplast Distance	Direction* CP Dist	Nuclear Distance	Direction* Nuc Dist
Germination	2.21	**22.28*** **	0.26	0.70	4.02[a]	1.03	1.29
Survival	4.22[a]	0.21	0.01	**22.60*** **	**25.77*** **	0.03	4.56
Flower number	0.45	0.77	0.21	0.53	0.95	0.58	0.04
Seed number	1.96	3.17	0.35	0.00	0.34	4.25[a]	3.60[a]
Pollen viability	0.09	1.67	0.20	0.63	0.29	0.29	0.02
Cumulative fitness	1.33	**7.26***	2.47	3.28[a]	**7.92***	0.17	2.06

a $P < 0.05$; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$
d.f. 1,12

Supplementary Figure 1: Relative hybrid performance for pollen viability graphed against geographic distance. Each point represents a cross mean, with diamonds and triangles representing within- and between-clade crosses respectively. Values were calculated as the percent reduction in hybrid performance relative to the parents, such that negative values represent RI. Points below the dotted line represent crosses where hybrids showed a significant reduction in pollen viability relative to the parental populations (Supp. Table 2).



Supplementary Figure 2: Proportion of seeds germinating for parental, F1 hybrid, and backcross generations for two between-clade crosses. Outer circle color depicts the origin of the cytoplasm, while the inner quadrants are the relative proportion of the nuclear genome from each parental population. Numbers below the parental circles indicate population IDs.



Supplementary methods

Evaluation of reproductive isolation (RI)

Seeds were sampled from 29 populations distributed throughout *Campanulastrum americanum's* range (Fig 1, Supp Table 1) with at least one population sampled from each clade. Twenty-eight pairwise crosses were conducted with all but two populations serving in two crosses (Fig 1). Eighteen of these crosses were within-clade, while ten were between-clades (parental populations belonged to different clades in either the chloroplast phylogeny, nuclear phylogeny, or both, Chapter 1)). 10 to 15 genotypes from each population were used in crosses, with reciprocal F1 hybrids created by using the same set of individuals as pollen donors and pollen recipients for alternate crossing directions. In all but three populations, each genotype came from a separate maternal family. The remaining three populations contained genotypes from between three and nine families due to sampling constraints. The same to 10 to 15 genotypes per population were also crossed pairwise within populations to produce parental population seed in the same greenhouse conditions as the hybrid seed. Altogether this crossing design resulted in 85 cross-types (29 within-population crosses + 28 between-population crosses x 2 crossing directions).

To determine performance of the first-generation (F1) hybrids relative to their parental populations, seed was planted and grown under controlled conditions. For each cross-type, 25 replicates of 5 seeds each were surface sown onto potting medium (3:1 Promix:Turface) in 2.54 cm x 2.54 cm cells in 9x18 germination flats. Replicates were evenly spread among maternal families and only normal looking seed was used for planting. While most hybrid seed was normal in appearance, some seed was less plump

and appeared to have issues with the development of the endosperm or embryo. Seed location was fully randomized and germination flats were placed in growth chambers with a 21°C day/14°C night temperature regime, 12-hour days, and daily watering. Germination success, timing, and seedling mortality, were scored for 33 days. Germination was scored when cotyledons had emerged and separated. At times, germinants were observed where the seed coat was not shed and the cotyledons never emerged. These situations were recorded as germination events, but with 100% mortality. After germination slowed, cells were randomly thinned to one seedling, unless cells contained seedlings with chlorotic phenotypes (i.e. white or variegated), where multiple seedlings were left to maximize the likelihood of at least one seedling surviving to transplant. In crosses where germination was low, seedlings from cells with more than one seedling were retained to increase sample size.

After germination had stopped (33 days post-planting), flats were moved to 5°C for seven weeks to stimulate flowering. Plants were then removed from the cold and any mortality recorded. They were transplanted into conetainers and moved to a greenhouse where supplemental light increased day length to 16 h. All seedlings were transplanted, with most crosses having 25 offspring, but as few as nine individuals for one cross-type (due to low germination and high early mortality). In total 2140 plants were transplanted. Plants were fertilized (20:20:20 N:P:K) every other week until bolting, at which point they were fertilized weekly. Survival until flowering and day of first flower was recorded. Flower counts were made at seven and 30 days after first flower and summed for an index of flower number. Within a few days after first flower, mature anthers were removed from a single bud on each plant prior to dehiscence. Anthers were air-dried and

then placed in lactophenol-aniline blue for at least 24 hours in order to score pollen viability (Kearns & Inouye, 1993). Pollen viability was determined as the fraction of pollen grains stained, with an average of 475 ± 161 grains scored per individual. A single pollination was carried out on each plant to evaluate seed production. All individuals that survived to flower were randomly crossed to two other individuals within that cross-type, with each individual serving once as a pollen recipient and once as a pollen donor. Crosses within maternal families were avoided to minimize any inbreeding.

For each of the 28 between-population crosses, hybrid performance relative to that of the parental populations was calculated for germination, survival to flower, flower number, seed number, pollen viability, and cumulative fitness. Cumulative fitness was calculated as (germination*survival*flower #*((seed # +pollen viability)/2)), with seed number standardized by dividing all seed counts by the largest seed count over all crosses.

Calculating Genetic Distance

Genetic distance was calculated for both chloroplast and nuclear loci using data from a previous study on the phylogeography of *C. americanum* (Chapter 1). Five chloroplast markers were sequenced for all populations and chloroplast genetic distance for each cross was calculated as the number of SNPs between parental haplotypes. Nuclear genetic distance was obtained from RAD-seq data. For all populations involved in between-clade crosses, as well as a subset of the remaining populations, DNA from six individuals was pooled, barcoded, and used for RAD-seq library construction and sequencing. The program Stacks (Catchen *et al.*, 2011; Catchen *et al.*, 2013) was then

used to identify and genotype loci. For more details on RAD sequencing and genotyping see Chapter 1. Using the fasta output from Stacks, maximum likelihood pairwise distances were calculated between all populations using RAxML (Stamatakis, 2014). These distances were used as the nuclear genetic distance for each of the crosses carried out in this study. Calculated distances were available for all the between-clade crosses, enabling a robust analysis of the effect of nuclear-genetic distance on strength of RI (as RI was only found in these crosses). For within-clade crosses where both parental populations were not included in the RAD-seq study, the geographically closest sequenced populations with the same chloroplast haplotype were used to obtain an estimated nuclear genetic distance (one parental population was estimated in 12 crosses, while both parental populations were estimated in five crosses). The nuclear phylogeny for *C. americanum* found geographic structuring of genetic lineages (Chapter 1), indicating that using a geographically close, genetically similar population is a reasonable method for estimating nuclear genetic distance.

Statistical Analysis

Relative hybrid performance was first determined for each F1 hybrid reciprocal separately by subtracting the mean of the two parental populations (midparent) from the mean F1 hybrid value for each reciprocal for each trait, and then dividing that value by the midparent value to get a standardized number depicting the percent hybrid deviation from the midparent value [((hybrid-parent)/parent)*100]. To obtain a mean hybrid performance for each cross, similar calculations were carried out at the cross level using a cross mean F1 hybrid value (mean of the two reciprocal means). Standardizing hybrid

performance allows for comparison among crosses. According to this metric, negative values of relative hybrid performance represent RI, or hybrid breakdown, while positive values represent heterosis or hybrid vigor.

The effect of genetic divergence (crossing within or between clades), geographic, chloroplast-, and nuclear-genetic distance on variation in mean hybrid performance for each trait was determined using ANCOVA. Genetic divergence tested whether or not crossing between genetic clades leads to a greater reduction in hybrid performance (RI) than crossing within clades. The contribution of geographic, chloroplast-genetic, and nuclear-genetic distance between parental populations to variation in hybrid performance was evaluated by including them as direct effects. Finally, we tested for interactions between genetic divergence and each measure of distance. Analyses were conducted on relative hybrid performance for each cross for six traits (germination, survival, flower number, seed number, pollen viability, and cumulative fitness) (PROC GLM, SAS 9.3, SAS Institute, INC. 2011).

A correlation analysis was conducted to check whether the measures of distance (geographic, chloroplast-genetic, or nuclear-genetic) were independent of one another (PROC CORR, SAS 9.3, SAS Institute, INC. 2011). No significant correlation was found: geographic versus chloroplast-genetic ($r=-0.01$, $p=0.98$), geographic versus nuclear-genetic ($r=-0.14$, $p=0.49$), and chloroplast-genetic versus nuclear-genetic ($r=0.30$, $p=0.12$).

To determine whether crossing direction influenced the magnitude of reduced hybrid performance (RI) and/or the relationship of hybrid performance with geographic or genetic distance, an ANCOVA was run using just the between-clade crosses. Separate

measures of relative hybrid performance were used for each F1 hybrid reciprocal. The analysis tested for an effect of crossing direction and included the three measures of distance as direct effects. Interactions between crossing direction and each measure of distance were also tested. Crossing direction was defined by whether or not the parental population with the more derived chloroplast haplotype was maternal.

Patterns of hybrid performance and RI were also examined at the individual cross level. ANOVA was carried out for each trait to test for variation among the four cross-types (2 within-population (parental) and 2 between-population (hybrid reciprocals)) (PROC GLM, SAS 9.3 SAS Institute, INC. 2011). For analysis of pollen viability, log-linear analyses were conducted assuming a gamma distribution and a log link (PROC GENMOD, SAS 9.3 SAS Institute, INC. 2011). If there was significant variation among cross-types, linear contrasts were used to test for RI (whether hybrid performance was significantly reduced relative to the parents) and asymmetry (whether the degree of reduction differed between hybrid reciprocals) within each cross. If there was significant asymmetry, linear contrasts were run to test whether only one or both of the reciprocals had significant RI (i.e. performance of a reciprocal was significantly reduced relative to the parents).

Backcross Germination Data

For one western x Appalachian-clade cross, 10x72, and one western x Smokies-clade cross, 27x36, backcrosses were carried out to test whether cytonuclear incompatibility underlies RI for germination. F1 hybrids were backcrossed to their corresponding maternal and paternal populations. F1 hybrids served as the pollen recipients, while the

parental populations were pollen donors. Maternal backcrosses result in a partial restoration of parental cytonuclear combinations, while paternal backcrosses result in further disruption of these combinations. Altogether, we carried out four backcross types per cross, maternal and paternal backcrosses on each F1 hybrid reciprocal. Multiple genotypes from each hybrid and parental population were used in crosses, and within- and between-parental population crosses were again carried out to make parental and F1 hybrid seed generated under the same greenhouse conditions as the backcross seed. To determine performance of the maternal and paternal backcrosses relative to their F1 hybrid and parental populations, seed was planted and germinated under controlled conditions using the same methods detailed above. Germination success and timing, as well as mortality, were scored for 30 days.

To evaluate whether germination differed among cross-types (2 parental, 2 hybrid reciprocals, 2 maternal backcrosses, and 2 paternal backcrosses) for each cross, log-linear analyses were conducted assuming a binomial distribution (PROC GENMOD, SAS 9.3 SAS Institute, INC. 2011). If there was significant variation among cross-types, linear contrasts were used to test whether each hybrid reciprocal or maternal/paternal backcross differed significantly from the parental populations. For 10x72 and 27x36, the only cross-type that showed significantly reduced germination relative to the parental populations (after Bonferroni correction for multiple non-independent contrasts) was the hybrid reciprocal where the populations with the less-derived chloroplast haplotype was maternal ($p<0.001$).

Literature Cited

Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A. & Cresko, W.A. 2013. Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22:** 3124-3140.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W. & Postlethwait, J.H. 2011. Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics* **1:** 171-182.

Kearns, C.A. & Inouye, D.W. 1993. *Techniques for pollination biologists*. University Press of Colorado.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*.

Chapter 3:

Patterns of intraspecific reproductive isolation are altered by the co-occurrence of genetically divergent lineages

Abstract

Understanding under what circumstances secondary contact and gene flow result in maintenance, enhancement, or loss of divergence and reproductive isolation is a central question to predicting patterns of speciation. In this study we used experimental greenhouse crosses to examine patterns of reproductive isolation (RI) among genetically divergent populations of *Campanulastrum americanum* that occur in close geographic proximity. We then compared these patterns to previously established patterns of RI between geographically isolated, divergent populations to determine how gene flow may have impacted reproductive isolation in this species. Substantial, asymmetric RI due to cytonuclear incompatibility was found for survival (up to 94% reduction). In comparison to patterns of RI among allopatric populations, RI among sympatric divergent lineages was less consistent and expressed in a smaller number of fitness traits. We hypothesize that gene flow among divergent lineages has led to reduced RI in *C. americanum*, with variable levels of gene flow leading to variation in strength of RI.

Introduction

Species ranges are dynamic over time, with processes such as climate change, in particular glacial cycles, having led to cyclic expansions and contractions (Hewitt, 2000). As a result of these range shifts, populations of a given species were often historically isolated in allopatric glacial refugia, which along with recurrent bottlenecks during subsequent recolonization, likely contributed to drift and genetic divergence among populations (Hewitt, 1996). These processes may also have facilitated the development of genetic incompatibility and the potential for reproductive isolation (RI) and speciation (Hewitt, 1996; Avise *et al.*, 1998; Carstens & Knowles, 2007). However, this dynamic nature of species ranges not only facilitates the development of genetic incompatibilities and RI among populations, but also increases the likelihood that divergent lineages will come back into contact before RI is complete. When this occurs, a central question is whether secondary contact and gene flow leads to a breakdown of this divergence, reducing RI (e.g. Seehausen, 2006; Taylor *et al.*, 2006), or whether barriers to gene flow persist and even increase, promoting RI and eventually speciation (e.g. reinforcement, Servedio & Noor, 2003).

Theory predicts that the outcome of secondary contact on genetic divergence and RI depends on the strength of selection and recombination on the divergent loci that contribute to RI (Felsenstein, 1981; Smadja & Butlin, 2011; Abbott *et al.*, 2013). RI is often multi-genic, with multiple barriers to gene flow required to achieve strong RI and complete speciation (Coyne & Orr, 2004). As populations diverge and barriers arise in allopatry, loci underlying barriers to gene flow are necessarily linked with one another. However, secondary contact and gene flow between populations raises the possibility of

recombination breaking up these associations and reducing the strength of RI. One way to reduce this effect of recombination would be strong selection against hybrids, preventing the formation of recombinant individuals. Another would be the building up of associations or linkage disequilibrium between barrier loci, for example through chromosomal inversions, which would help to maintain RI and genetic divergence (Smadja & Butlin, 2011). Accordingly, maintenance of RI would be more likely when a smaller number of barrier loci of large effect underlie RI and/or when there are stronger associations/linkage between these loci (Smadja & Butlin, 2011). In addition, the outcome of secondary contact has also been found to be influenced by both the strength and directionality of gene flow (Servedio & Kirkpatrick, 1997; Servedio & Noor, 2003) Therefore, the outcome of secondary contact will be variable, depending on the number and strength of barrier loci, the associations between these loci, and the patterns of gene flow.

Despite the importance of understanding the consequences of secondary contact and gene flow on RI between incompletely isolated lineages, there remains a general lack of empirical data regarding how frequently gene flow between divergent lineages results in an enhancement, maintenance, or breakdown of RI (Abbott *et al.*, 2013). One way to gain a better understanding of the variable outcomes of secondary contact is to directly examine the impact of gene flow on RI by comparing patterns of RI between divergent lineages of a species both in sympatry and allopatry. The goal of the current study was to examine how secondary contact may have influenced postzygotic RI in an herbaceous species, *Campanulastrum americanum*, by comparing patterns of RI among genetically-divergent populations that occur in close proximity and have likely experienced some

degree of secondary contact and gene flow, versus divergent populations that remain geographically isolated.

A previous study examining the range-wide patterns of RI in *C. americanum* found substantial postyzogtic isolation in first generation (F1) hybrids, in the form of reduced germination and survival, when crossing between divergent genetic lineages (Chapter 2). Reduced germination occurred with increasing geographic distance between parental populations, while increasing genetic distance was associated with decreased survival, which appeared to be due to cytonuclear incompatibility. However, most of the crosses used in this range-wide study were across large geographic distances, involving populations that are unlikely to have experienced any gene flow in nature. In particular, all but two of the ten crosses between genetically divergent lineages involved populations belonging to the Western clade of *C. americanum*, which while widespread through much of the species range, does not co-occur with any of the other genetic lineages (Chapter 1).

In contrast to the Western clade, the remaining three clades of *C. americanum* have limited distributions and are all confined to the Appalachians and the immediate surrounding regions. These divergent lineages co-occur and potentially have experienced secondary contact and gene flow. Here we carry out pairwise crosses between populations from these three Appalachian clades and use F1 hybrids to estimate postzygotic RI across the life cycle including germination, survival to flower, and reproductive traits. We also examine whether geographic or genetic distance between parental populations predicts patterns of RI. Specifically, we ask: 1) What traits exhibit RI? 2) What type of genetic incompatibility appears to underlie RI? 3) How do patterns

of RI among sympatric divergent populations compare to the patterns found among allopatric divergent populations?

Materials and Methods

Study System

C. americanum is an autotretraploid, monocarpic herb found in the eastern half of the United States. Individuals are annual or biennial, insect-pollinated, and highly outcrossing (Galloway *et al.*, 2003). *C. americanum* typically grows in disturbed habitats and seeds are dispersed passively, traits that likely contribute to its patchy population structure. Chloroplast and nuclear markers resolve four genetic lineages: Western, Smoky Mountains (Smokies), Appalachian Mountains (Appalachian), and Eastern (Chapter 1). The Appalachian clade is the oldest, with the Eastern, Smokies, and Western clades progressively more derived. These clades differ in their geographical extent with the Western clade found throughout much of the species range, and the other three primarily restricted to the Appalachians (Chapter 1).

Evaluation of RI

To evaluate patterns of RI in sympatry, nine populations were identified that were distributed throughout the Appalachian region of *C. americanum's* range, with two, three, and four populations sampled from the Eastern, Smokies, and Appalachian clades respectively (Supp. Table 1, Figure 1). Populations were chosen to span the full range of chloroplast divergence present in the Appalachians (0-25 SNPs), which is similar to the study of allopatric divergent populations (0-27 SNPs). Nuclear divergence of populations

in the current study covered a slightly narrower range (0.0010-0.0021) than in the study of allopatric populations (0.0007-0.0024), and had half as much variation (SD of 0.0002 versus 0.0004). For the crossing design, the nine populations were split into two groups with each group containing two Appalachian-clade populations, an Eastern-clade population, and two Smokies-clade populations (one Smokies-clade population, TN88, was used in both crossing groups). This design resulted in the two following groups of five populations each: group 1: VA73, TN92, VA71, VA86, and TN88; and group 2: MD5, NC91, VA85, NC90, and TN88. All pair-wise crosses were then conducted between populations in each group, and additional crosses were carried out between groups to increase the number of within-clade crosses. In total, eight within- and 16 between-clade crosses were conducted, with both reciprocals carried out for each cross. In addition, crosses were also carried out within each of the nine populations to obtain parental seed generated under the same greenhouse conditions. Altogether this crossing design resulted in 57 cross-types (9 within-population crosses + 24 between-population crosses x 2 crossing directions).

Individuals sampled from multiple field-collected families from each parental population were grown under controlled conditions for carrying out within- and between-clade crosses. Thirty to 45 replicates of one seed each were surface sown onto potting medium (3:1 Promix:Turface) in 2.54 cm x 2.54 cm cells in 9x18 germination flats. Replicates were spread evenly among 18-20 families, except for NC90 where only five families were available. Seed location was fully randomized and germination flats were placed in a growth chamber with a 24°C day/14°C night temperature regime and 12-hour days. Seeds were watered daily. After germination had slowed, flats were moved to 5°C

for eight weeks of vernalization to stimulate flowering. Plants were then removed from the cold and a subset of 16 plants from each population was transplanted into conetainers and moved to a greenhouse where supplemental light increased day length to 16 h. Individuals to transplant were chosen in a stratified random manner, as they were evenly distributed across maternal families but randomly selected within families. Sixteen genotypes from each population were then used in carrying out within- and between-clade crosses, with reciprocal crosses created by using the same set of individuals as pollen donors and pollen recipients in alternate crossing directions. No more than two genotypes came from the same maternal family except for population NC90, which included 3-4 genotypes per family as seed was only collected from five plants. The same 16 genotypes were also crossed pair-wise within the populations to produce parental population seed in the same greenhouse conditions as the hybrid seed.

To determine performance of the first generation (F1) hybrids relative to their parental populations, seed was planted and grown under controlled conditions using the same methods detailed above. For each cross-type, 40 replicates of 2 seeds each were planted, with replicates evenly spread among maternal families and using only normal looking seed. Germination success, timing, and seedling mortality were scored for 36 days. Germination was scored when cotyledons had emerged and separated. At times, seedlings were observed where the seed coat was not shed and the cotyledons never emerged. These situations were recorded as mortality following germination events. After germination slowed, all cells were randomly thinned to one seedling, unless cells contained seedlings with chlorotic phenotypes (i.e. white or variegated), where multiple seedlings were left to maximize the likelihood of sufficient seedlings surviving for

transplant. After germination had stopped (36 days post-planting), flats were moved to 5°C for eight weeks of vernalization to stimulate flowering. Plants were then removed from the cold and any mortality recorded. A subset of 25 surviving plants from each cross-type was then transplanted into the greenhouse. Individuals to transplant were randomly chosen, but evenly distributed across maternal families, with one to five individuals per maternal family. Two of 57 cross-types, TN88xMD5 and TN88xVA73, had reduced numbers (22 and 15, respectively), due to high early mortality. In total 1412 plants were transplanted. Plants were fertilized (20:20:20 N:P:K) every other week until bolting, at which point they were fertilized weekly. Survival until flowering and day of first flower was recorded.

Reproductive traits were scored on flowering individuals. Flower counts were carried out 14 days after first flower. Within a few days after first flower, mature anthers were removed from a single bud on each plant prior to dehiscence. Anthers were air-dried and then placed in lactophenol-aniline blue for at least 24 hours in order to score pollen viability (Kearns & Inouye, 1993). Pollen viability was determined as the fraction of pollen grains stained, with an average of 595 ± 210 grains scored per individual. Twenty-five individuals had less than 25 total grains of pollen and were deemed pollen sterile. A single pollination was carried out on each plant to evaluate seed production. All individuals that survived to flower were randomly crossed to two other individuals within that cross-type, with each individual serving once as a pollen recipient and once as a pollen donor. Crosses within maternal families were avoided to minimize inbreeding. Most pollinations were carried out within three weeks of first flowering.

For each of the 24 between-population crosses (8 within- and 16 between-clade), hybrid performance relative to that of the parental populations was calculated for germination, survival to flower, flower number, seed number, pollen viability, and cumulative fitness. Cumulative fitness was calculated as (germination*survival*total flower #*((seed #+pollen viability)/2)), with seed number standardized by dividing all seed counts by the largest seed count over all cross-types.

Calculating genetic distance

Genetic distance was calculated for both chloroplast and nuclear loci using data from a previous study on the phylogeography of *C. americanum* (Chapter 1). Five chloroplast markers were sequenced for all populations and chloroplast genetic distance for each cross was calculated as the number of SNPs between parental haplotypes. Nuclear genetic distance was obtained from RAD-seq data. For all nine populations, DNA from six individuals was pooled, barcoded, and used for RAD-seq library construction and sequencing. The program Stacks (Catchen *et al.*, 2011; 2013) was then used to identify and genotype loci. For more details on RAD sequencing and genotyping see Chapter 1. Using the fasta output from Stacks, maximum likelihood pairwise distances were calculated between all populations using RAxML (Stamatakis, 2014). These distances were used as the nuclear genetic distance for each of the crosses carried out in this study.

Statistical Analysis

Mean relative hybrid performance was determined for each cross by subtracting the midparent value (mean of the two parental populations) from the mean F1 hybrid value

for each trait, and then dividing that value by the midparent value to get a standardized percent hybrid deviation from the midparent value [((hybrid-parent)/parent)*100]. This allows for comparison among crosses. Negative values represent RI, while positive values represent heterosis or hybrid vigor. Similar calculations were also carried out for each F1 hybrid reciprocal separately (two values per cross) to determine if hybrid performance depended upon crossing direction.

The effect of crossing within versus between clades, and geographic, chloroplast-, and nuclear-genetic distance on variation hybrid performance for each trait was determined using ANCOVA (N=24, 8 within- + 16-between clade crosses). We tested whether crossing between genetic clades leads to greater RI (reduced hybrid performance) than crossing within clades. We also tested whether geographic, chloroplast-genetic, or nuclear-genetic distance between parental populations contributed to the strength of RI by including them in the model as direct effects. Tests for the interactions between genetic divergence and the three measures of distance were only significant for one out of 18 analyses; therefore, the interaction terms were dropped and the multiple regression analysis rerun with only the four main effects. Analyses were conducted on mean relative hybrid performance for each cross for the six traits (germination, survival, flower number, seed number, pollen viability, and cumulative fitness) (PROC GLM, SAS 9.3 SAS Institute, INC. 2011). No significant correlations were detected among the measures of distance used in the ANCOVA analysis (geographic, chloroplast-genetic, or nuclear-genetic; PROC CORR, SAS 9.3, SAS Institute, INC. 2011).

To determine whether crossing direction influenced the magnitude of RI and/or the relationship of RI with geographic, chloroplast-, or nuclear-genetic distance, a similar ANCOVA analysis was run using just the between-clade crosses (N=36, 16 between-clade crosses x 2 crossing directions). Separate measures of relative hybrid performance calculated for each F1 hybrid reciprocal were used and the analyses tested for an effect of crossing direction, the three measures of distance, included as direct effects, and the interactions between crossing direction and the measures of distance. Crossing direction referred to whether or not the parental population with the more derived chloroplast haplotype was maternal.

Patterns of hybrid performance and RI were also examined at the individual cross level, ANOVA was carried out for each trait to test for variation among the four cross-types (2 within-population (parental) and 2 between-population (hybrid reciprocals)) (PROC GLM, SAS 9.3 SAS Institute, INC. 2011). In these analyses, individuals served as the data points. For pollen viability, log-linear analyses were conducted assuming a gamma distribution and a log link (PROC GENMOD, SAS 9.3 SAS Institute, INC. 2011). If there was significant variation among cross-types, linear contrasts were used to test for RI (whether hybrid performance differed from parents) and hybrid asymmetry (whether performance differed between hybrid reciprocals) within each cross. In cases of asymmetry, additional linear contrasts were run to test whether only one or both of the reciprocals had significant RI (i.e. performance of a reciprocal differed from the average of the parents).

Results

Substantial postyzygotic RI was found in between-clade crosses, with up to a 57% reduction in cumulative fitness. RI for cumulative fitness was primarily driven by reduced hybrid survival, the only individual fitness trait to show substantial RI (also up to 57% reduction) (Fig 2A). Chloroplast-genetic distance between parental populations predicted strength of RI for survival (r^2=0.62, Table 1, Fig 2A), with larger distances generally leading to greater reductions in survival. Correspondingly, RI for cumulative fitness was also predicted by chloroplast-genetic distance (Table 1), with larger distances again leading to a greater reduction in fitness. RI for survival in between-clade crosses was strongly asymmetrical, with the occurrence of RI strongly dependent upon crossing direction (Table 2, Fig 2B). For all 10 crosses that showed RI for survival, a reduction in fitness was only observed when the population with the more derived chloroplast was maternal (Table 3). Fitness reductions in this crossing direction were substantial (up to 94% reduction in survival).

RI for pollen viability was also found in a subset of crosses and was influenced by genetic divergence and chloroplast-genetic distance (Table 1). However, in contrast to other traits that exhibited RI, reduced pollen viability occurred in both within- and between-clade crosses, and only in crosses involving a single Smokies-clade population, VA86 (Table 3). When graphing RI for pollen viability against chloroplast genetic distance, it appeared as though one cross, VA86xVA71 with a 52% reduction, was driving the effect of genetic divergence and chloroplast-genetic distance on strength of RI (Fig 3). Further analysis found that this cross was acting as a strong outlier in the multiple regression analysis (Cook's D = 1.05), and removing the cross from the analysis eliminated the significant effect of genetic divergence and chloroplast-genetic distance on

strength of RI for pollen viability. A similar effect explained the significant values for pollen viability when examining the effect of crossing direction, chloroplast-genetic distance, and their interaction on strength of RI (Table 2). Again, removing VA86xVA71 from the analysis eliminated these significant effects. The reason VA86xVA71 was such a strong outlier is that it contained a large number of individuals with only inviable pollen relative to those with viable pollen (Fig 4, see discussion of pollen sterility below), strongly impacting the strength of RI for pollen viability for this cross.

Complete pollen sterility also occurred in four between-clade crosses (Fig 4), independent of the RI for pollen viability mentioned above. Each of these four crosses involved a Smokies-clade population with chloroplast haplotype I, and either an Eastern- or Appalachian-clade population. The occurrence of pollen sterility was also strongly asymmetrical, with pollen sterile individuals only occurring when the Smokies-clade populations were maternal. Pollen sterility in these crosses was manifested in two different ways, hybrid individuals producing little to no pollen (25 grains or less) or only inviable pollen. Pollen sterility was not fixed within a cross, as hybrid individuals varied in their phenotype (no pollen, only inviable pollen, or viable pollen), with the frequency of pollen sterile individuals (no pollen and inviable pollen) ranging from 20-87% (Fig 4). Two of the crosses containing pollen sterile individuals (VA86xVA71 and VA86xVA73, Fig 4) also exhibited significant RI for pollen viability (Table 3).

Discussion

Substantial RI was found when crossing between genetically divergent, sympatric populations of *C. americanum*. RI in between-clade crosses was primarily restricted to a

single fitness trait, survival, with strength of RI predicted by chloroplast-genetic distance. There was also variance in the strength of RI for survival in the between-clade crosses. The less divergent crosses (Smokies x East) showed no reduction in survival, and even among the more divergent crosses there was variance in strength of RI, with some crosses (e.g. those involving NC90) showing little to no reduction in fitness.

RI for survival was strongly asymmetric, with crosses only showing reduced fitness when populations with the more derived chloroplast were maternal. Asymmetrical RI is often indicative of cytonuclear incompatibility (Tiffin *et al.*, 2001; Turelli & Moyle, 2007). Further support for cytonuclear incompatibility was found in that reduced survival of F1 hybrids was primarily due to chlorosis (reduced production of chlorophyll), indicating that the more derived chloroplast was frequently unable to function effectively on the hybrid nuclear background. In addition, in variegated individuals with chlorotic and green leaf tissue, the two tissue types correspond to alternate parental chloroplast haplotypes (Chapter 4). This correlation of chlorosis with chloroplast haplotype further suggests that cytonuclear incompatibility underlies chlorosis in *C. americanum*, rather than other mechanisms, such as autoimmune response (Bomblies & Weigel, 2007).

In addition to low levels of RI for pollen viability, pollen sterility was also found in a small set of crosses. In four crosses, all involving two Smokies populations with the same chloroplast haplotype (I; VA86 and NC90), a subset of F1 hybrids expressed pollen sterility either by producing no pollen, or only inviable pollen. In each of these four crosses, pollen sterility was only observed when the Smokies populations were maternal, suggesting a cytoplasmic contribution and therefore cytoplasmic male sterility (CMS; Chase, 2007). CMS has not previously been documented in *C. americanum*, but its

occurrence in between-clade crosses is in accordance with previous studies finding that crosses between species can reveal CMS in hermaphroditic species (Fishman & Willis, 2006; Leppala & Savolainen, 2011). Pollen sterility was not fixed within crosses, with the frequency of pollen sterile hybrids varying from 20-87%, indicating the existence of genetic variance within populations for either for the mitochondrial CMS-determining alleles or the nuclear restorers. Segregating alleles affecting CMS have also been found in *Mimulus* and *Arabidopsis* (Sweigart *et al.*, 2007; Leppala & Savolainen, 2011). One of the two *C. americanum* populations that caused CMS (VA86) had a general impact on pollen viability, such that RI for pollen viability was found in all crosses in which it was used, including when crossing with the other population containing the same chloroplast haplotype.

In the current study, substantial RI was found for a single fitness trait, survival, which contrasts the pattern found among allopatric populations, where RI was observed for two fitness traits, germination and survival (Chapter 2). There was no RI for germination in the current study, whereas in the allopatric study, RI for germination was observed in between-clade crosses when the parental populations were separated by large geographic distances (>400 km). As populations for the current study were sampled across a more restricted geographic area, many of the between-clade crosses were between populations separated by distances below this threshold. However, five between-clade crosses did occur at distances greater than 400 km, compared to eight crosses in the allopatric study, and even in these five crosses no RI for germination was observed. Altogether, the results suggest that the underlying cause of RI for germination that exists among allopatric divergent populations is not present among the sympatric populations

sampled in the current study. Environmental sources of variation could underlie the variation in the presence of RI for germination in these two studies. However, we think this explanation is unlikely, as RI for germination was replicated for two of the allopatric crosses using F1 hybrid seed produced and grown in a different year in a separate study (Barnard-Kubow, unpublished data).

Although cytonuclear incompatibility led to substantial, asymmetrical RI for survival, with strength of RI predicted by chloroplast-genetic distance, the strength of this relationship differed between studies (r^2=0.62 in sympatry versus r^2=0.87 in allopatry). This difference was due to a less consistent occurrence of RI among genetically-divergent crosses in the current study. Between-clade crosses at smaller chloroplast genetic distances showed no RI for survival, and even some of the more divergent sympatric crosses had little to no reduction in survival. In addition, the asymmetry in RI for survival was much stronger in the current study, with significant RI only observed when populations with the more derived chloroplast were maternal. In the allopatric study moderate RI was also found in the alternate crossing direction. Altogether these findings suggest that the cytonuclear incompatibility found among allopatric populations also operates among sympatric populations, however, the strength of this incompatibility has been reduced in some crosses.

Overall, the patterns of RI observed among allopatric populations are somewhat attenuated in sympatry. RI impacts fewer traits in sympatry, with RI for germination no longer observed, and RI for survival restricted to a single crossing direction and less consistent in its occurrence among genetically divergent populations. Although the maximum reduction in cumulative fitness is similar between studies (94% versus 92%),

suggesting that some pairs of sympatric populations maintain levels of RI seen among allopatric populations, several between-clade crosses in sympatry show reduced or even no RI. We hypothesize that reduction of RI in sympatry is likely due to secondary contact and gene flow among populations. The reason crosses differ in how much RI has been reduced may be due to variance in the strength and directionality of gene flow between populations. Such variation may lead to RI being maintained in some areas of secondary contact, while in others it is broken down. Estimates of gene flow among populations are needed to test this hypothesis.

Determining how frequently secondary contact and gene flow results in enhancement, maintenance, or breakdown of RI is important for informing our understanding of how gene flow impacts the outcome of incompletely isolated lineages (Abbott *et al.*, 2013). Maintenance of RI and genetic divergence in at least some portions of the genome in areas of sympatry has been found in multiple studies (e.g. Carneiro *et al.*, 2013; Larson *et al.*, 2014; Lindtke *et al.*, 2014). Strengthening of RI has also been documented through the increase of prezygotic isolation in sympatric relative to allopatric populations (Servedio & Noor, 2003). In contrast, loss of genetic divergence and RI have also been found, although these have primarily focused on instances of ecological speciation and the loss of prezyogtic or extrinsic postzyogtic RI (Seehausen, 2006; Taylor *et al.*, 2006). Studies examining the impact of secondary contact and gene flow on the strength of intrinsic postzygotic RI remain relatively rare, although this type of situation is likely to be common when lineages that have diverged in allopatry come back into secondary contact. The results from our study show that in the case of *C. americanum*, intrinsic postzygotic RI is moderately attenuated in areas of secondary

contact, suggesting the potential for a loss of RI and genetic divergence overtime. The degree in reduction of RI also varied among crosses, possibly due to differences in the extent or directionality of gene flow. Our results suggest *C. americanum* could be a good system for increasing our understanding of what circumstances are likely to lead to a breakdown versus maintenance of RI and genetic divergence in secondary contact zones between partially isolated genetic lineages.

Acknowledgements

We would like to thank Melissa Aikens for collecting seed from *C. americanum* populations and Lauren Wilson, Erin Arnold, Anna Greenlee, Holly Prendeville, and Victoria Soler for assistance in the greenhouse. We also thank Janis Antonovics, Ben Blackman, Edmund Brodie, David Carr, and Douglas Taylor for helpful discussion and comments. This work was supported by the ARCS Foundation.

References

Abbott, R., Albach, D., Ansell, S., Arntzen, J.W., Baird, S.J.E., Bierne, N., Boughman, J.W., Brelsford, A., Buerkle, C.A., Buggs, R., Butlin, R.K., Dieckmann, U., Eroukhmanoff, F., Grill, A., Cahan, S.H., Hermansen, J.S., Hewitt, G., Hudson, A.G., Jiggins, C., Jones, J., Keller, B., Marczewski, T., Mallet, J., Martinez-Rodriguez, P., Most, M., Mullen, S., Nichols, R., Nolte, A.W., Parisod, C., Pfennig, K., Rice, A.M., Ritchie, M.G., Seifert, B., Smadja, C.M., Stelkens, R., Szymura, J.M., Vainola, R., Wolf, J.B.W. & Zinner, D. 2013. Hybridization and speciation. *Journal of Evolutionary Biology* **26:** 229-246.

Avise, J.C., Walker, D. & Johns, G.C. 1998. Speciation durations and Pleistocene effects on vertebrate phylogeography. *Proceedings of the Royal Society B-Biological Sciences* **265:** 1707-1712.

Bomblies, K. & Weigel, D. 2007. Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. *Nature Reviews Genetics* **8:** 382-393.

Carneiro, M., Baird, S.J.E., Afonso, S., Ramirez, E., Tarroso, P., Teotonio, H., Villafuerte, R., Nachman, M.W. & Ferrand, N. 2013. Steep clines within a highly permeable genome across a hybrid zone between two subspecies of the European rabbit. *Molecular Ecology* **22:** 2511-2525.

Carstens, B.C. & Knowles, L.L. 2007. Shifting distributions and speciation: species divergence during rapid climate change. *Molecular Ecology* **16:** 619-627.

Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A. & Cresko, W.A. 2013. Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22:** 3124-3140.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W. & Postlethwait, J.H. 2011. Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics* **1:** 171-182.

Chase, C.D. 2007. Cytoplasmic male sterility: a window to the world of plant mitochondrial-nuclear interactions. *Trends in Genetics* **23:** 81-90.

Coyne, J.A. & Orr, H.A. 2004. *Speciation*. Sinauer Associates, Sunderland, Mass.

Felsenstein, J. 1981. Skepticism towards Santa Rosalia, or why are there so few kinds of animals. *Evolution* **35:** 124-138.

Fishman, L. & Willis, J.H. 2006. A cytonuclear incompatibility causes anther sterility in *Mimulus* hybrids. *Evolution* **60:** 1372-1381.

Galloway, L.F., Etterson, J.R. & Hamrick, J.L. 2003. Outcrossing rate and inbreeding depression in the herbaceous autotetraploid, *Campanula americana*. *Heredity* **90:** 308-315.

Hewitt, G. 2000. The genetic legacy of the Quaternary ice ages. *Nature* **405:** 907-913.

Hewitt, G.M. 1996. Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society* **58:** 247-276.

Kearns, C.A. & Inouye, D.W. 1993. *Techniques for pollination biologists*. University Press of Colorado.

Larson, E.L., White, T.A., Ross, C.L. & Harrison, R.G. 2014. Gene flow and the maintenance of species boundaries. *Molecular Ecology* **23:** 1668-1678.

Leppala, J. & Savolainen, O. 2011. Nuclear-cytoplasmic interactions reduce male fertility in hybrids of *Arabidopsis lyrata* subspecies. *Evolution* **65:** 2959-2972.

Lindtke, D., Gompert, Z., Lexer, C. & Buerkle, C.A. 2014. Unexpected ancestry of *Populus* seedlings from a hybrid zone implies a large role for postzygotic selection in the maintenance of species. *Molecular Ecology* **23:** 4316-4330.

Seehausen, O. 2006. Conservation: Losing biodiversity by reverse speciation. *Current Biology* **16:** R334-R337.

Servedio, M.R. & Kirkpatrick, M. 1997. The effects of gene flow on reinforcement. *Evolution* **51:** 1764-1772.

Servedio, M.R. & Noor, M.A.F. 2003. The role of reinforcement in speciation: theory and data. *Annual Review of Ecology Evolution and Systematics* **34:** 339-364.

Smadja, C.M. & Butlin, R.K. 2011. A framework for comparing processes of speciation in the presence of gene flow. *Molecular Ecology* **20:** 5123-5140.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*.

Sweigart, A.L., Mason, A.R. & Willis, J.H. 2007. Natural variation for a hybrid incompatibility between two species of *Mimulus*. *Evolution* **61:** 141-151.

Taylor, E.B., Boughman, J.W., Groenenboom, M., Sniatynski, M., Schluter, D. & Gow, J.L. 2006. Speciation in reverse: morphological and genetic evidence of the collapse of a three-spined stickleback (*Gasterosteus aculeatus*) species pair. *Molecular Ecology* **15:** 343-355.

Tiffin, P., Olson, M.S. & Moyle, L.C. 2001. Asymmetrical crossing barriers in angiosperms. *Proceedings of the Royal Society of London Series B-Biological Sciences* **268:** 861-867.

Turelli, M. & Moyle, L.C. 2007. Asymmetric postmating isolation: Darwin's corollary to Haldane's rule. *Genetics* **176:** 1059-1088.

Table 1: Testing for an effect of genetic divergence (crossing within or between clades), and magnitude of geographic, chloroplast-genetic, or nuclear-genetic distance between parental populations of *C. americanum*, on hybrid performance for fitness components. *F*-values are reported with values significant at $P < 0.05$ shown in bold.

Trait	Genetic Divergence	Geographic Distance	Chloroplast Distance	Nuclear Distance
Germination	1.87	0.13	0.21	0.01
Survival	0.23	0.58	**20.94*** **	2.80
Flower number	0.01	0.40	3.27[a]	0.01
Seed number	0.77	0.18	0.94	2.23
Pollen viability	**6.04***	1.03	**7.76***	0.44
Cumulative fitness	2.40	2.54	**6.65***	0.87

a $P < 0.1$; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$
d.f. 1,19

Table 2: Comparison of crossing direction and magnitude of geographic, chloroplast-genetic, or nuclear-genetic distance between parental populations of *C. americanum*, on the strength of RI for fitness components of between-clade crosses. *F*-values are given for testing for an effect of crossing direction (which population is maternal), distance, as well as the interaction between the two. Values in bold are significant at $P < 0.05$.

Trait	Direction	GeoDist	Direction *GeoDist	CpDist	Direction *CpDist	NucDist	Direction *NucDist
Germination	0.02	0.57	0.17	0.00	0.50	0.010	0.03
Survival	0.72	0.53	0.13	**15.04*** **	**14.44*** **	1.68	2.12
Flower number	0.19	0.81	0.60	2.36	1.17	0.12	0.29
Seed number	0.25	0.07	1.18	1.99	0.20	1.24	0.68
Pollen viability	0.86	1.33	0.13	**7.50***	**4.70***	0.63	0.11
Cumulative fitness	0.57	1.42	0.43	**4.65***	**4.39***	0.15	0.23

*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$
d.f. 1,24

Table 3: Individual cross-level analyses of hybrid performance for each trait, including the metric of cumulative fitness (germ*surv*flwr*((seed+pollen)/2). Down arrows indicate crosses where hybrids performed significantly worse than their midparent (RI), while up arrows indicate significantly better performance (heterosis). In crosses where the reciprocal F1 hybrids were significantly asymmetrical, results for the reciprocals are shown separately, with the crossing direction where the more derived population was maternal represented by arrows on the left, and the other crossing direction by arrows to the right. If no significant asymmetry was detected, then only one arrow is shown representing the results for the pooled hybrid population.

Crosstype	Clade	Germination	Survival	Flower #	Seed #	Pollen	Fitness	Cross
within	App							5x73
within	App			↑*				91x92
within	App	↑*		↑*			↑*	92x73
within	App							5x91
within	East		↓*				↑	85x71
within	Smokies			↑*		↓*		90x86
within	Smokies					↓*		88x86
within	Smokies							88x90
between	Smokies x East			↑*		↓*	↓* ↑	86x71
between	Smokies x East						↓	90x85
between	Smokies x East							88x71
between	Smokies x East							88x85
between	East x App		↓*	↓*			↓*	71x73
between	East x App		↓*					71x92
between	East x App		↓*	↓*				85x5
between	East x App							85x91
between	Smokies x App							90x5
between	Smokies x App		↓*					90x91
between	Smokies x App	↑*	↓*			↓*	↓*	86x73
between	Smokies x App		↓*	↓ ↑*	↓	↓*	↓* ↑	86x92
between	Smokies x App	↓	↓*	↓*			↓*	88x73
between	Smokies x App		↓*	↓ ↑			↓*	88x92
between	Smokies x App		↓*				↓*	88x5
between	Smokies x App		↓*	↓*	↓		↓*	88x91

Grey: $P < 0.075$, black $P < 0.05$, * $P < 0.0125$ (significant after correction for multiple non-independent contrasts)

Figure 1: Location of *C. americanum* populations in the eastern United States used for evaluating RI. Populations shaded according to genetic clade.

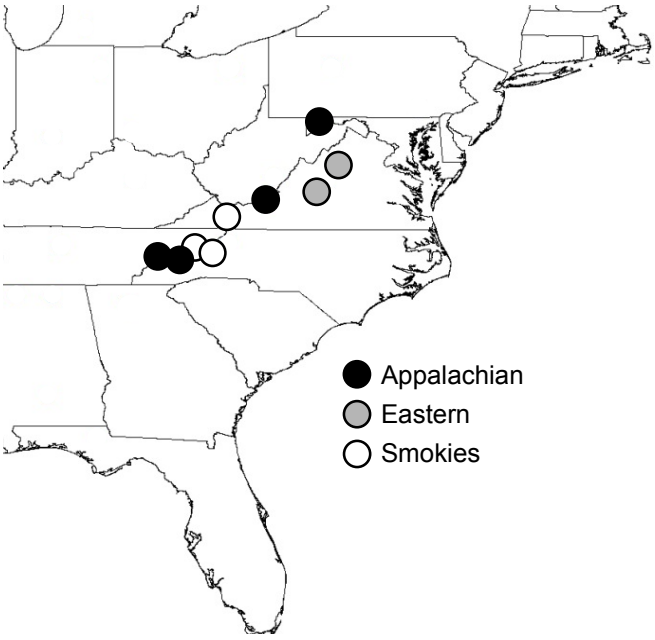


Figure 2: Survival of *C. americanum* hybrids relative to parental populations graphed against chloroplast-genetic distance, calculated as the percent reduction in hybrid survival relative to the parents, such that negative values represent RI. A) Mean hybrid survival for each cross. B) Hybrid survival broken down by crossing direction (indicated by shading of triangles), with only the between-clade crosses shown. r^2 values obtained from regression analysis of relative hybrid survival against chloroplast genetic distance.

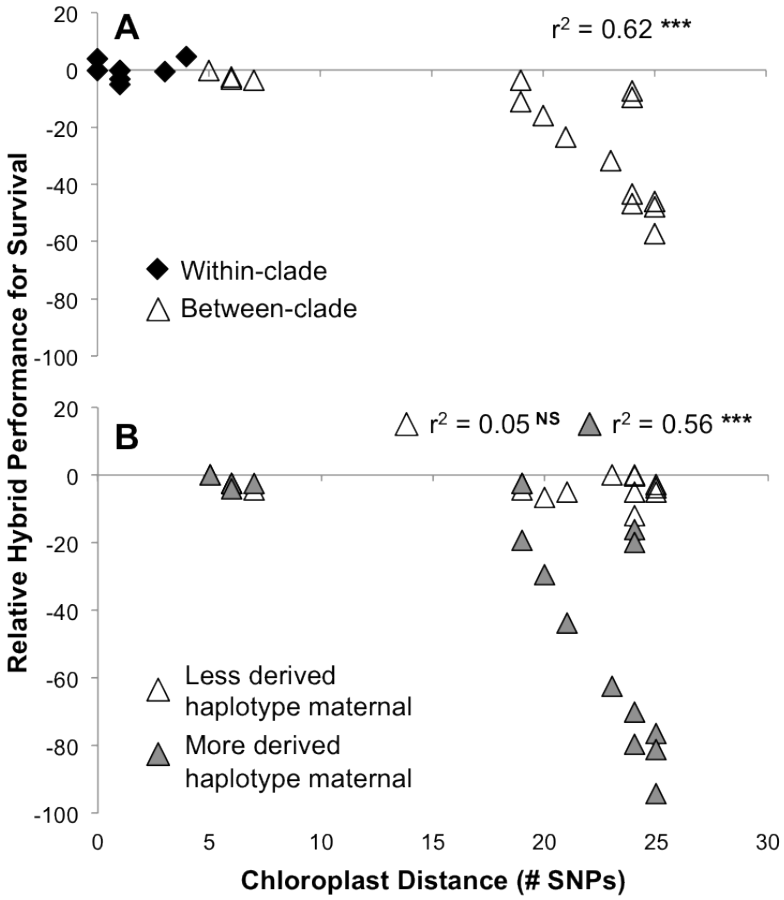


Figure 3: Pollen viability of *C. americanum* hybrids relative to parental populations graphed against chloroplast-genetic distance, calculated as the percent reduction in hybrid pollen viability relative to the parents, such that negative values represent RI. Each point represents a cross mean.

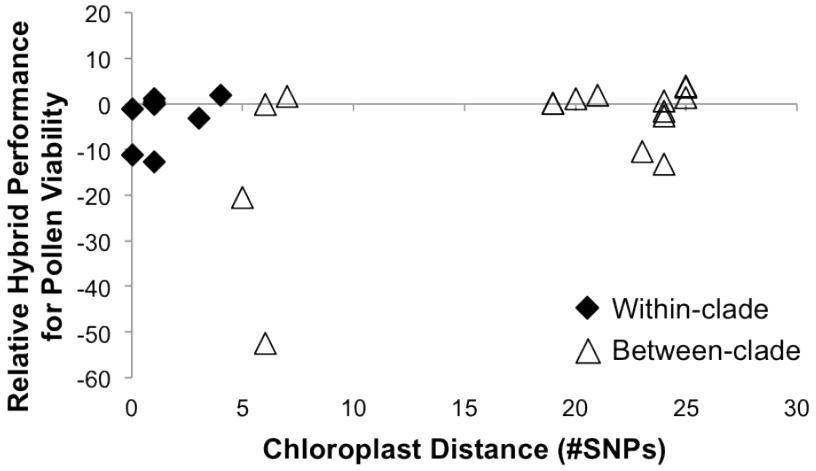
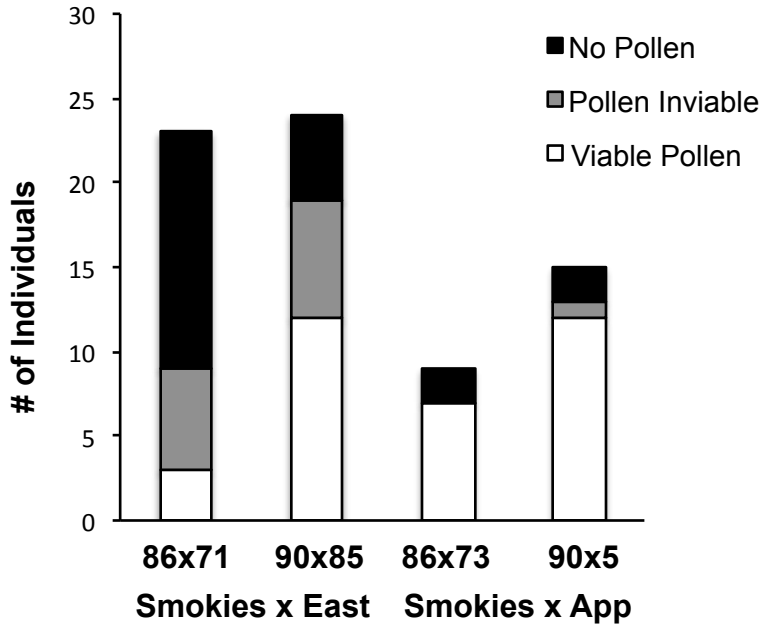


Figure 4: Patterns of pollen sterility in four *C. americanum* between-clade crosses. For each cross, the number of hybrid individuals are shown that produced no pollen, only inviable pollen, or viable pollen. Data is only shown for the direction of the cross where the population with the more derived chloroplast was maternal, as pollen sterility did not occur in the other crossing direction. Population IDs are listed for each cross, as well as the type of between-clade cross to which they belong.



Supplementary Table 1: Information for the *C. americanum* populations in the eastern United States used in the experiment to evaluate RI including geographic location, phylogenetic clade, and chloroplast haplotype (from Chapter 1).

Pop ID	State	Latitude	Longitude	Clade	Haplotype
91	North Carolina	35.5862	-83.0663	Appalachian	A
92	Tennessee	35.6758	-83.5259	Appalachian	B
5	Maryland	39.6137	-79.1158	Appalachian	D
73	Virginia	37.3534	-80.5522	Appalachian	D
71	Virginia	38.3305	-78.4901	East	E/F
85	Virginia	37.7576	-79.1876	East	E
90	North Carolina	35.7669	-82.1636	Smokies	I
86	Virginia	36.6344	-81.5881	Smokies	I
88	Tennessee	35.9833	-82.4989	Smokies	J

Chapter 4:

Biparental chloroplast inheritance rescues cytonuclear incompatibility

This chapter will be formatted as a co-authored manuscript (Barnard-Kubow KB, McCoy MA, and Galloway LF).

Abstract

Although organelle inheritance is predominantly maternal across animals and plants, biparental chloroplast inheritance has arisen multiple times in the angiosperms. Biparental chloroplast inheritance has been hypothesized to have evolved to overcome incompatible plastids in species with cytonuclear incompatibility. Accordingly, biparental chloroplast inheritance has the potential to impact the evolutionary dynamics of cytonuclear incompatibilities, which have been proposed to be some of the earliest genetic incompatibilities to arise in the speciation process. In the current study, we examine the interplay between biparental inheritance and cytonuclear incompatibility in *C. americanum*, a plant species that exhibits both of these traits. By determining patterns of chloroplast inheritance in genetically similar and divergent crosses, and hybrid survival across multiple generations, we evaluate the factors influencing the occurrence of biparental inheritance and its impact on cytonuclear incompatibility. We find substantial biparental inheritance in *C. americanum* independent of the presence of cytonuclear incompatibility, although biparental inheritance increases in the presence of cytonuclear incompatibility. Biparental inheritance also mitigates the effects of cytonuclear incompatibility, leading to increased fitness of F1 hybrids and recovery in the F2 generation. Therefore, even though a strong cytonuclear incompatibility exists among populations of *C. americanum*, supporting the idea that cytonuclear incompatibilities are among the first genetic incompatibilities to arise, biparental inheritance can rescue hybrids from this incompatibility, potentially slowing speciation.

Introduction

Organelle inheritance is predominantly uniparental across plants and animals and is achieved by numerous mechanisms, leading to the hypothesis that uniparental inheritance is likely under positive selection (Birky, 1995; Mogensen, 1996; Birky, 2001). Multiple theories have been proposed for why uniparental inheritance might be selectively advantageous, primarily centered on the idea of minimizing genomic conflict and the spread of selfish organelles (Eberhard, 1980; Hurst, 1992; Burt & Trivers, 2006). However, biparental inheritance of the chloroplast genome has arisen multiple times within the angiosperms, with around 20% of species tested having the potential for biparental inheritance (Corriveau & Coleman, 1988; Zhang *et al.*, 2003). Biparental inheritance appears to be derived from maternal inheritance (Hu *et al.*, 2008). Despite the potential importance of this mechanism of chloroplast inheritance, the evolutionary implications of biparental inheritance have received relatively little study.

The potential selective forces leading to biparental inheritance of the chloroplast have not been as extensively discussed as those contributing to the pervasiveness of uniparental inheritance. The theory most often proposed for why biparental inheritance may be favored in some taxa, is the idea that it has evolved as a mechanism to overcome incompatible plastids in species that have cytonuclear incompatibility (Hu *et al.*, 2008; Zhang & Sodmergen, 2010; Jansen & Ruhlman, 2012). Some evidence for this theory is provided by the occurrence of biparental inheritance in several taxa that exhibit cytonuclear incompatibility, including *Oenothera* and *Pelargonium* (reviewed in Jansen & Ruhlman, 2012). Strength of cytonuclear incompatibility was also found to influence

chloroplast transmission in *Oenothera* (Chiu & Sears, 1993), further supporting the idea of an interaction between cytonuclear incompatibility and chloroplast inheritance.

There has been a growing interest in cytonuclear incompatibilities as they have been found to contribute to reproductive isolation and speciation in plants (e.g. Levy, 1991; Fishman & Willis, 2006; Sambatti *et al.*, 2008; Greiner *et al.*, 2011; Leppala & Savolainen, 2011), yeast (reviewed in Chou & Leu, 2010), and animals (reviewed in Burton *et al.*, 2013).

Several factors predispose cytonuclear incompatibilities to be strong contributors to reproductive isolation. These include the likelihood of cytonuclear coevolution, as the organelle and nuclear genomes must function together to carry out important cellular processes such as photosynthesis and respiration (Rand *et al.*, 2004). Organelles also have a reduced effective population size as a result of their haploid nature and uniparental mode of inheritance. Along with a lack of normal recombination, this reduction in effective population size leads to an increased accumulation of deleterious mutations in the organelle genomes (Birky, 2001). The uniparental inheritance and haploid nature of organelle genomes also results in a greater likelihood of cytonuclear incompatibilities impacting fitness in first generation (F1) hybrids relative to nuclear genetic incompatibilities. Due to these factors, cytonuclear incompatibilities have been proposed to be among the first genetic incompatibilities to arise, potentially impacting the earliest stages of the speciation process (Levin, 2003; Fishman & Willis, 2006; Burton *et al.*, 2013). However, this reasoning relies, in part, on the assumption of uniparental inheritance of the organelle genomes. Little is known regarding how biparental inheritance may impact the dynamics of cytonuclear incompatibility.

Biparental inheritance of the chloroplast has the potential to impact the relative contribution of cytonuclear incompatibility to reproductive isolation. First, the occurrence of biparental inheritance in crosses that result in cytonuclear incompatibility increases the likelihood hybrids inherit a chloroplast that is compatible with the nuclear genome (e.g. Chiu & Sears, 1993; Bogdanova, 2007). Second, biparental inheritance introduces genetic variability among the organelles, allowing selection to occur. For example, in plants, if one chloroplast haplotype is incompatible on the hybrid nuclear background, selection against that haplotype, along with vegetative sorting, could lead to the loss of that haplotype (Birky, 2001). Losing the incompatible haplotype could lead to a recovery in fitness in the recombinant second-generation (F2) hybrids. Both of these mechanisms would enable biparental chloroplast inheritance to act as a rescue mechanism for hybrids suffering a loss of fitness due to cytonuclear incompatibilities, supporting the theory that biparental chloroplast inheritance has evolved as a way for species with cytonuclear incompatibilities to overcome incompatible plastids (Hu *et al.*, 2008; Zhang & Sodmergen, 2010; Jansen & Ruhlman, 2012). Although this theory has been raised multiple times in the literature, it has rarely been tested.

In the current study we examine whether biparental inheritance influences the dynamics of cytonuclear incompatibility using the plant species *Campanulastrum americanum*. Previous studies found a strong cytonuclear incompatibility between genetically divergent populations of *C. americanum*, leading to chlorosis and greatly reduced survival in F1 hybrids (Galloway & Etterson, 2005; Etterson *et al.*, 2007; Chapters 2 and 3). Green and white variegation was also frequently observed in F1 hybrids, suggesting the potential for biparental chloroplast inheritance; variegation is

thought to be due to biparental inheritance and subsequent vegetative sorting of the chloroplast (Birky, 2001), if one of the chloroplasts is incompatible on the hybrid nuclear background (Ureshino *et al.*, 1999; Bogdanova, 2007; Weihe *et al.*, 2009). In addition, studies have found plastid DNA present in the pollen generative cells in all of the 10 species examined in the Campanulaceae (Sears, 1980; Corriveau & Coleman, 1988; Zhang *et al.*, 2003), suggesting the potential for biparental plastid inheritance in *C. americanum*.

Accordingly, *C. americanum* appears to have both cytonuclear incompatibility and biparental inheritance, making this a model system in which to examine the interplay and evolutionary consequences of these two traits. By determining patterns of chloroplast inheritance and survival across multiple hybrid generations, we examine these evolutionary dynamics, addressing the following questions: 1) What is the typical pattern of chloroplast inheritance in *C. americanum*? 2) Does hybridization between genetically divergent lineages trigger or increase the frequency of biparental inheritance? 3) Does biparental inheritance lead to increased fitness of F1 hybrids in genetically-divergent crosses that show cytonuclear incompatibility? 4) Does biparental inheritance and vegetative sorting allow for selection against and ultimately the loss of the incompatible plastid, leading to recovery in recombinant hybrid generations?

Methods

Study System

C. americanum is an autotretraploid, monocarpic herb found in the eastern half of the United States. Individuals are annual or biennial, insect-pollinated, and highly

outcrossing (Galloway *et al.*, 2003). Plants grow as a vegetative rosette, and after vernalization, bolt and flower. *C. americanum* typically grows in disturbed habitats and seeds are passively dispersed, characteristics that likely contribute to its patchy population structure. Chloroplast markers resolve three genetic clades: Western, Appalachian, and Eastern (Chapter 1). These clades differ in their distribution, with the Appalachian and Eastern clades restricted primarily to the Appalachians, and the Western clade spread throughout the range. Populations are generally fixed for chloroplast haplotype (Chapter 1). *C. americanum* is a member of the Campanulaceae, which have been shown to have highly rearranged chloroplast genomes with increased rates of nucleotide substitution (Cosner *et al.*, 2004; Knox, 2014; Chapter 5) in addition to the potential for biparental chloroplast inheritance.

Patterns of Chloroplast Inheritance

F1 Samples and Survival

To examine patterns of chloroplast inheritance in *C. americanum*, F1 hybrid seed was used from crosses carried out in earlier studies (Chapters 2 and 3). Three within-clade crosses were chosen to determine the typical mode of inheritance (AL69xNE59, MI44xIL20, and TN34xMO7; Fig 1). All six parental populations contained chloroplast haplotypes found in the previously characterized Western haplotype clade (Chapter 1), henceforth referred to as haplotype A. Within-clade crosses of *C. americanum* do not show reproductive isolation, and always produce green, viable F1 hybrids (Chapters 2 and 3). Six between-clade crosses were chosen to determine the mode of chloroplast inheritance in genetically-divergent crosses (IN68xVA73, PA52xWV72, IA10xWV72,

AL29xMD5, VA86xTN92, and TN88xNC91; Fig 1). All six crosses were between the more derived Western (haplotype A) and Appalachian (haplotype C) chloroplast clades and have been found to exhibit cytonuclear incompatibility (Chapters 2 and 3). Altogether this resulted in 9 crosses (3 within- and 6 between-clade) each with two crossing directions.

Hybrid seed from the nine crosses were grown under controlled conditions to determine chloroplast inheritance and survival for F1 hybrids. For each crossing direction of each cross, multiple replicates of two seeds each were surface sown onto potting medium (3:1 Promix:Turface) in 2.54 cm x 2.54 cm cells in 9x18 germination flats. Number of replicates ranged from 40-85 based on expected germination rate (effected by age of hybrid seed and reproductive isolation for germination, Chapter 1). Ten replicates of two seeds each were also planted for each parental population. In total, 1150 replicates were planted. Replicates were spread evenly among approximately ten maternal families for each cross. Seed location was fully randomized and germination flats were placed in a growth chamber with a 24°C day/14°C night temperature regime and 12-hour days and were watered daily. Seedlings were scored for germination and non-green phenotypes were recorded (e.g. variegated, white, chlorotic with yellowing due to insufficient production of chlorophyll; Fig 2). Due to lower than expected germination rates in some crosses, additional F1 hybrid seed was planted for a subset of the between-clade crosses to obtain larger sample sizes. Seed was planted and grown as described above, except that replicates contained 15 seeds each.

To determine chloroplast haplotype, seedlings from within-clade crosses and parental populations were harvested whole once they had at least two true leaves. A total

of 480 within-clade cross and 313 parental individuals were harvested. For the between-clade crosses, seedlings were not sacrificed unless they were completely white, as previous work found that white seedlings do not live beyond a week (Barnard-Kubow unpublished). Any seedlings starting to die were also harvested. Once germination had slowed, the surviving plants were moved to 5°C for vernalization to stimulate flowering. The first batch of plants was vernalized for ten weeks, while the second batch was vernalized for seven weeks.

To estimate F1 survival to reproduction, after vernalization a subset of plants was transplanted into conetainers and moved to a greenhouse, where supplemental light increased day length to 16 h. For most of the between-clade crosses, 25 randomly selected plants were transplanted per crossing direction, distributed across maternal families. However, for crosses where low survival was expected due to the high frequency of non-green phenotypes, all surviving plants (up to 70 individuals) were transplanted. Plants not transplanted or that were starting to die were harvested. Smaller plants were harvested whole, while for larger plants, a leaf was harvested that was representative of the overall phenotype. Transplanted plants had rosette leaf samples taken before or within a few days of starting to bolt. For variegated individuals, both green and white leaf tissue samples were taken to test if phenotype (green vs. white) was correlated with chloroplast haplotype. Combining pre- and post-transplant, a total of 978 tissue samples were taken from between-clade hybrid individuals (summing across the first and second plantings). Plants were monitored for survival and flowering.

Vegetative Sorting

Plants that start off heteroplasmic for the organelle genomes (containing multiple haplotypes) are thought to primarily transmit only a single organelle haplotype to the next generation, due to the tendency for one organelle genome to be lost through vegetative sorting as individuals grow (Birky, 2001). To examine whether vegetative sorting contributes to changes in chloroplast haplotype frequency, a leaf subtending a flower was collected for 24 plants that had both haplotypes A and C when juvenile. For comparison, comparable apical leaf tissue samples were also taken from individuals that inherited only haplotype A or C (N=13 and 37 respectively).

F2 Samples and Survival

To make F2 seed, two pollinations were carried out on each between-clade F1 plant. Individuals were randomly crossed to four other plants from the same cross and crossing direction, with each individual serving twice as a pollen recipient and twice as a pollen donor. The number of individuals pollinated ranged from 12 to 28 per cross and crossing direction, with pollinations within families avoided to minimize inbreeding.

Chloroplast haplotype and survival in the F2 generation was evaluated with seed from the between-clade F1 hybrids using the methods detailed above. 40-56 replicates of two seeds each, evenly distributed across 10-14 families, were planted for each cross (476 replicates total). To get an accurate estimate of performance of offspring from F1 mothers with differing chloroplast haplotypes (A, C, AC), F2 seed was chosen from maternal plants to represent the three haplotypes. F2 seed was also only chosen from maternal plants where apical leaf tissue had been sampled to examine vegetative sorting. As a result, the haplotype distribution of maternal plants contributing to the F2 generation

was not always representative of the total haplotype distribution for that cross. For example, cross 88x91 had 13A, 2C, and 5AC individuals flower. However, seeds from 5A, 2C, and 3AC individuals were planted. As in the F1, seeds were scored for germination, non-green phenotypes were recorded, and any seedlings that were completely white or dying were harvested. After germination slowed, seedlings were randomly thinned to one per cell, with a subset of thinned seedlings kept for tissue samples. The germination flats were moved to 5°C for seven weeks to stimulate flowering, and then transplanted to containers and moved to the greenhouse. Plants were monitored for survival and flowering. Plants were scored as surviving if they were still alive after the majority of plants had flowered, even if they never flowered. Again, any plants that started to die were harvested and all surviving plants had rosette leaf samples taken before or within a few days of bolting. A total of 823 leaf tissue samples were collected from between-clade F2 hybrids.

Genotyping Samples for Chloroplast Haplotype

To determine chloroplast haplotypes in F1 and F2 hybrid individuals, a subset of leaf samples was selected for each cross. Up to 62 samples per cross and crossing direction for both the F1 and F2 generation were chosen for DNA extraction to determine chloroplast haplotype. Samples were distributed among maternal families, with most crosses having 8 to 10 families. Eight samples were also chosen from each parental population as a control. For the F1 generation, 15 of 18 combinations of cross and crossing direction had sample sizes greater than 45 (Table 1), with an average of six individuals per family. The remaining three had lower sample sizes due to poor

germination and survival. For the F2 generation, 11 of 12 combinations of cross and crossing direction had samples sizes greater than 40; the exception was 73x68, which had a sample size of 15.

DNA extractions were carried out using a modified CTAB procedure either in single tube or 96 well plate format. Chloroplast haplotype was determined for each sample using Custom Taqman SNP Genotype Assays (Life Technologies), similar to methods used for tracking mitochondrial inheritance (Bentley *et al.*, 2010). Three different assays were designed and used for genotyping based on SNPs in the chloroplast genome. For each assay, a standard curve was first constructed by precisely mixing together various ratios (e.g. 2:98, 5:95, 10:90, 20:80, etc.) of the two parental haplotypes using parental population DNA extractions. These mixtures were run in triplicate on an ABI 7500 fast real-time PCR system (Life Technologies). The average difference in CT value between the two probes was calculated and graphed against the log of the artificial mixture ratios. This comparison resulted in a straight line whose equation could then be used to back calculate ratios obtained from DNA samples to determine the ratio of the two chloroplast haplotypes present in a given sample (Supp. Fig 1). Each DNA extraction was run at least once using the RT-PCR assay. A subset of individuals was run three times to examine inherent variation in the RT-PCR assays for estimating error in the calculated haplotype ratios.

Statistical Analysis

Patterns of Chloroplast Inheritance

The degree of biparental inheritance, and whether it differed among crosses and was impacted by genetic divergence was determined. RT-PCR results were used to calculate the frequency of non-maternal inheritance for each cross as well as for each crossing direction. Individuals were conservatively scored as having non-maternal inheritance when the RT-PCR assay estimated a 10% or greater contribution of the paternal chloroplast based on threshold detection values of 5% when calculating the standard curve and an estimated variance of 5% when running the same sample multiple times. A log-linear analysis assuming a binomial distribution and a logit link was conducted to test the effect of genetic divergence (within-clade versus between-clade crosses) as well as cross (nested within genetic divergence) on degree of non-maternal inheritance (PROC GENMOD, SAS 9.3 SAS Institute, INC. 2011). To test whether inheritance patterns differed if populations with chloroplast haplotype A or C were maternal, a similar log-linear analysis was conducted using only between-clade crosses, with cross, crossing direction, and the interaction of these two factors in the model.

Chloroplast Haplotype and Fitness

To examine the impact of chloroplast haplotype on the fitness of between-clade F1 hybrids, the effect of cross, chloroplast haplotype, and their interaction on F1 hybrid survival was tested using a log-linear analysis assuming a binomial distribution and a logit link (PROC GENMOD, SAS 9.3 SAS Institute, INC. 2011). The data suggested that haplotype A was predominantly incompatible on the hybrid nuclear background, and that this incompatibility varied among crosses. To test this observation, a log-linear analysis determined the effect of cross on F1 hybrid survival for haplotype A individuals. Finally,

a correlation analysis was run to determine the relationship between the probability of individuals in a given cross surviving when having only haplotype A and degree of non-maternal inheritance for that cross (PROC CORR, SAS 9.3, SAS Institute, INC. 2011).

Non-maternal Inheritance and Fitness

To examine the impact of non-maternal inheritance on the fitness of F1 hybrids, we first determined the proportion of surviving F1 hybrid offspring that exhibited maternal versus non-maternal inheritance. Then, the proportion of surviving F2 offspring resulting from non-maternal inheritance was estimated by first determining the mean survival of F2s from maternal plants that had been genotyped as A, C, or AC. Finally, using the number of A, C, or AC F1 hybrids that flowered for each cross, the proportion of F2 offspring that would have come from F1 maternal plants with non-maternal inheritance had all F2 families been planted was calculated. As the data indicated haplotype A was incompatible on the divergent hybrid nuclear background, the impact of non-maternal inheritance on fitness was only determined for the crossing direction where haplotype A was maternal.

Non-maternal Inheritance and Recovery

We examined whether there was a loss of the incompatible haplotype A in the F2 relative to the F1 hybrid generation in the crossing direction where haplotype A was maternal. The frequency of haplotype A for each cross in the F1 and F2 generation was calculated as the number of individuals genotyped with haplotype A plus the number of individuals with both haplotypes multiplied by 0.5, divided by the total number of individuals

assayed. The proportionate loss of haplotype A in the F2 relative to the F1 was then calculated. Positive values indicate a loss of haplotype A in the F2, while negative values indicate an increase in frequency. A correlation analysis was run to determine if there was a relationship between the proportionate loss of haplotype A and the probability of F1 hybrids surviving when only inheriting haplotype A (PROC CORR, SAS 9.3, SAS Institute, INC. 2011).

To examine whether loss of haplotype A leads to concomitant recovery in fitness, we tested whether there was significant recovery in the F2 generation relative to the F1. Recovery was examined by testing for an effect of cross, generation (F1 versus F2), and their interaction on hybrid survival using a log-linear analysis assuming a binomial distribution and a logit link. As the interaction was significant, similar analyses were run for each cross individually, testing for an effect of generation. Finally, a correlation analysis was run to determine if there was a relationship between proportionate loss of haplotype A and extent of recovery in survival.

Results

Patterns of Chloroplast Inheritance

Substantial biparental and paternal chloroplast inheritance was found in all crosses (Table 1; Figs 3, 4), although inheritance was predominantly maternal. Biparental inheritance was confirmed in variegated seedlings where genotyping white and green leaf tissue punches from 36 individuals found white leaf tissue always contained haplotype A and green tissue haplotype C (Supp. Fig 2). The frequency of non-maternal inheritance ranged from 15-51% among combinations of cross and crossing direction with a mean

and standard deviation of 27±12%. Crossing between clades led to an increase in the frequency of non-maternal inheritance relative to when crossing within a clade (31% versus 23%; $\chi^2_1 = 6.82$, p=0.009). There was also an overall effect of crossing direction in between-clade crosses ($\chi^2_1 = 8.71$, p<0.001), with a greater degree of non-maternal inheritance when populations with chloroplast haplotype A were maternal (36% versus 22%; Table 1).

Chloroplast Haplotype and Fitness

Survival of F1 between-clade hybrids varied among chloroplast haplotypes (A, C, or AC; $\chi^2_2 = 305.56$, p<0.001) and among crosses ($\chi^2_5 = 19.88$, p=0.001), but the effect of haplotype was consistent across crosses (haplotype*cross; $\chi^2_{10} = 15.30$, p=0.121). The likelihood of survival was lowest for individuals with haplotype A (12% survival), highest for those with haplotype C (96% survival), and intermediate for those that inherited both A and C haplotypes. The likelihood of surviving with haplotype A varied among between-clade crosses from 0-30% ($\chi^2_5 = 30.96$, p<0.001). In the crossing direction where populations with haplotype A were maternal, the likelihood of surviving was positively correlated with degree of maternal inheritance (r=0.93, p=0.008, Fig 5A).

Non-maternal Inheritance and Fitness

Non-maternal inheritance contributed substantially to the fitness of F1 hybrids in the crossing direction where populations with haplotype A were maternal. In the three crosses where haplotype A was almost entirely inviable (68x73, 52x72, and 10x72; x-axis Fig 5), 92-100% of surviving F1 hybrids were the result of non-maternal inheritance (Fig

6A). While in crosses where around 15% of individuals with haplotype A survived (29x5 and 86x92), 58-68% of surviving F1 hybrids were the result of non-maternal inheritance. The final cross, 88x91, had 30% survival of hybrid individuals with haplotype A and only 32% of surviving F1 hybrids were the result of non-maternal inheritance. These differential effects of non-maternal inheritance on hybrid fitness were also found when examining performance of the F2 hybrid offspring (Fig 6B). For 68x73, 52x72, and 10x72, 96-100% of surviving F2 offspring were estimated to have come from F1 individuals with non-maternal inheritance. This number was also high at 83% for 29x5 and 86x92. However, for 88x91, only 32% of surviving F2 offspring were estimated to come from F1 individuals with non-maternal inheritance.

Vegetative Sorting

Evidence for vegetative sorting was found when genotyping between-clade, variegated plants. A sample of 122 F1 hybrids were scored as variegated shortly after geminating, suggesting an AC haplotype. However, according to genotyping, only 83 had the expected AC haplotype, with the remaining 39 individuals containing only a single haplotype.

Additional evidence for vegetative sorting was found when comparing rosette and apical leaf tissue samples from between-clade hybrids. Of the 24 variegated individuals genotyped as AC as rosettes, two thirds of them were genotyped as containing only a single haplotype when flowering (haplotype C in all but 2 cases), indicating vegetative sorting had occurred (Fig 7). However, the remaining eight individuals retained both haplotypes, indicating this process was not always complete. In comparison, of the 50

individuals genotyped as either A or C when rosettes, all but one retained those genotypes when flowering. The one remaining individual was genotyped as AC when flowering.

Non-maternal Inheritance and Recovery

All six between-clade crosses had a reduction in the frequency of haplotype A in the F2 generation relative to the F1 when focusing on the crossing direction where populations with haplotype A were maternal. The degree of loss was negatively correlated with the likelihood of F1 hybrid survival when inheriting only haplotype A ($r=-0.99$, $p<0.001$, Fig 5B). Crosses in which F1 hybrids with haplotype A survived (29x5, 86x92, and 88x91) experienced a 14-54% loss of haplotype A, while crosses where F1 hybrids with haplotype A were inviable experienced a larger reduction (81-97%). Although reductions in haplotype A were substantial, none of the crosses showed a complete loss of this haplotype.

All six between-clade crosses had improved survival in the F2 generation relative to the F1 when haplotype A was maternal (Fig 8). Across all between-clade crosses there was a significant effect of generation on survival ($\chi^2_1 = 44.73$, $p<0.001$), indicating an overall recovery in the F2 generation relative to the F1. However, the degree of recovery varied among crosses, as indicated by the significant interaction between cross and generation ($\chi^2_5 = 19.48$, $p=0.002$). Three of the crosses exhibited significant recovery in the F2 relative to the F1 (10x72: $\chi^2_1 = 30.46$, $p<0.001$; 52x72: $\chi^2_1 = 17.16$, $p<0.001$; 68x73: $\chi^2_1 = 3.86$, $p=0.05$). Recovery in 29x5 was near significant ($\chi^2_1 = 3.40$, $p=0.065$), while there was only modest recovery for 86x92 ($\chi^2_1 = 2.15$, $p=0.142$) and 88x91 ($\chi^2_1 =$

0.79, p=0.375). There was also a positive, near significant correlation between the degree of loss of haplotype A and degree of recovery in survival (r=0.79, p=0.059, Fig 9).

Discussion

Substantial biparental and paternal chloroplast inheritance (up to 51%) was found in *C. americanum* regardless of genetic divergence. Biparental inheritance fits with previous studies that found the potential for biparental inheritance in all Campanulaceae species screened, as evidenced by the presence of plastid DNA in the pollen generative cells (Sears, 1980; Corriveau & Coleman, 1988; Zhang *et al.*, 2003). However, potential biparental inheritance is not synonymous with realized biparental inheritance, as multiple mechanisms exist to achieve uniparental inheritance beyond exclusion of plastids from pollen generative cells (Mogensen, 1996). Accordingly, studies that have screened for patterns of plastid inheritance in hybrid offspring have found wide ranges of realized biparental plastid inheritance, ranging from 1-2% (e.g. *Nicotiana* and *Iris*, Medgyesy *et al.*, 1986; Cruzan *et al.*, 1993), to up to 90% (e.g. *Oenothera* and *Zanteschedia*, Chiu & Sears, 1980; Snijder *et al.*, 2007). Taxa with high levels of biparental inheritance, such as *Oenothera* and *Zanteschedia*, also tend to show a wide range in inheritance patterns (3-90%) among different crosses, and interestingly, often also exhibit cytonuclear incompatibility (reviewed in Snijder *et al.*, 2007). The patterns of biparental inheritance found for *C. americanum* in this study fit well with the patterns observed in other taxa.

Degree of biparental inheritance in *C. americanum* was influenced by both genetic divergence and cytonuclear incompatibility. Increased biparental inheritance was found in more genetically divergent between-clade crosses relative to within-clade

crosses, which fits with previous studies that found taxonomically wider crosses often lead to a breakdown in maternal inheritance or increased degrees of non-maternal inheritance (e.g. Bogdanova, 2007, Hansen *et al.*, 2007). In the between-clade crosses of *C. americanum* greater biparental inheritance was also found when the incompatible haplotype A was maternal. The degree of biparental inheritance in this crossing direction was strongly correlated with strength of cytonuclear incompatibility. Altogether, these results indicate an interaction between the strength of cytonuclear incompatibility and degree of biparental inheritance in *C. americanum*. This finding fits with previous studies in *Oenothera* and *Pisum* where cytonuclear incompatibility was found to influence plastid transmission (Chiu & Sears, 1993; Bogdanova, 2007).

Biparental chloroplast inheritance in *C. americanum* acted to rescue genetically divergent hybrids from cytonuclear incompatibility. When the incompatible chloroplast A was maternal, biparental inheritance allowed for increased survival of between-clade F1 hybrids, with 30-100% of surviving individuals resulting from non-maternal inheritance. Similarly, other studies have demonstrated biparental chloroplast inheritance can lead to the inheritance of a compatible paternal plastid, allowing for survival in F1 hybrids that would otherwise die (Chiu & Sears, 1993; Ureshino *et al.*, 1999; Bogdanova, 2007). The substantial fitness contributions of biparental inheritance in *C. americanum* cascaded down to the next generation, as F1 individuals with biparental inheritance contributed the majority of surviving F2 offspring in five of the six between-clade crosses.

Biparental inheritance, along with vegetative sorting and selection, also led to a reduction of the incompatible haplotype A in the F2 hybrid generation. However, even in crosses where haplotype A was completely inviable, it was not completely lost. One

reason for this could be incomplete vegetative sorting in heteroplasmic individuals that contain both haplotypes A and C. Typically, vegetative sorting is thought to lead to loss of one of the haplotypes as a heteroplasmic plant grows (Birky, 2001). If sorting results in fixation of the incompatible chloroplast, the plant will most likely die, preventing the transmission of the incompatible haplotype to the next generation. Alternatively, if sorting results in fixation of the compatible haplotype, the plant will likely survive and reproduce, passing on this haplotype to the next generation. Evidence for such vegetative sorting was found in the 14 *C. americanum* individuals that were genotyped as AC as rosettes, but contained only the compatible C haplotype when flowering. However, another eight individuals genotyped as AC as rosettes still contained both haplotypes when flowering, indicating that vegetative sorting was not always complete, which would provide a mechanism by which haplotype A could be transmitted to the F2 generation even when inviable.

Concomitant with a reduction in the frequency of haplotype A in the F2 generation, the between-clade crosses also exhibited an increase in survival. However, this recovery was only statistically significant in some crosses, with crosses that had a greater loss of haplotype A experiencing greater recovery. These results support the idea that biparental inheritance can act as a rescue mechanism for cytonuclear incompatibility by allowing for selection against, and loss of, the incompatible plastid, leading to a recovery in fitness in recombinant hybrid generations.

While biparental inheritance overall mitigated the effects of cytonuclear incompatibility in *C. americanum*, there was variance in the importance of this rescue mechanism among crosses. Crosses in which the cytonuclear incompatibility was weaker

(individuals with haplotype A could survive) had less of a fitness increase due to biparental inheritance and did not exhibit significant recovery in the F2, unlike the three crosses where individuals with haplotype A was completely inviable. The cross with the weakest cytonuclear incompatibility, 88x91, was particularly striking in this regard. Only 32% of its fitness came from individuals with biparental inheritance, and it had the lowest F2 survival with almost no recovery relative to the F1 generation. One potential explanation for the variance in strength of cytonuclear incompatibility and ability of biparental inheritance to act as a rescue effect may be due to secondary contact and gene flow. Two of the crosses that exhibited weaker cytonuclear incompatibility and no significant recovery (86x92 and 88x91) are between populations that are found in the part of *C. americanum's* range where the genetically divergent lineages now co-occur (Chapter 1). Populations 88 and 91, in particular, are separated by only 68km.

Overall our results provide support for the theory that biparental inheritance evolved as a mechanism for overcoming incompatible plastids in species that exhibit cytonuclear incompatibility (Hu *et al.*, 2008; Zhang & Sodmergen, 2010; Jansen & Ruhlman, 2012). However, biparental plastid inheritance also occurs in *C. americanum* in the absence of cytonuclear incompatibility, suggesting that while this mode of inheritance is selectively favored in the presence of cytonuclear incompatibility, its underlying occurrence may be independent of this incompatibility. Taxa that exhibit potential biparental plastid inheritance, such as the Campanulaceae, often also contain highly rearranged plastid genomes, with increased repetitive DNA, and accelerated rates of nucleotide substitution (Jansen & Ruhlman, 2012; Chapter 5). Accordingly, in these taxa, biparental inheritance may be one manifestation of a more general destabilization of the

plastid genome (Jansen & Ruhlman, 2012); a destabilization that may also increase the propensity for cytonuclear incompatibility.

The ability of biparental inheritance to act as a rescue mechanism for cytonuclear incompatibility has the potential to impact the evolutionary dynamics of cytonuclear incompatibilities and their role in the speciation process, particularly since these traits often co-occur (Jansen & Ruhlman, 2012). Previous studies in *C. americanum* demonstrated the presence of a strong cytonuclear incompatibility leading to reduced survival in intraspecific hybrids (Chapters 2 and 3), supporting the idea that cytonuclear incompatibilities may be among the first genetic incompatibilities to evolve (Levin, 2003; Fishman & Willis, 2006; Burton *et al.*, 2013). However, results from the current study demonstrate that biparental inheritance can rescue this cytonuclear incompatibility, reducing its contribution to reproductive isolation and potentially slowing the speciation process. The exact outcome of this interplay, though, depends on the strength of the cytonuclear incompatibility (i.e. strength of selection). When the incompatibility is strong, selection leads to loss of the incompatible haplotype and recovery in the F2. When the incompatibility is weaker, selection is also weaker, and the incompatible chloroplast continues to be maintained in the F2, negatively impacting fitness. Almost paradoxically then, the weaker cytonuclear incompatibility is maintained to a greater extent in recombinant generations, resulting in an overall greater contribution to reproductive isolation.

Acknowledgements

We thank David McCauley and Sara Samoray for assisting us with getting the Taqman SNP assays up and running. We also thank Connor Johnson, Elissa Trieu, and Mohan Nagaraja for assisting with DNA extractions and AnhThu Nguyen for technical advice. Janis Antonovics, Ben Blackman, Edmund Brodie, David Carr, Douglas Taylor, and members of the Galloway and Brodie labs provided helpful discussion and comments. This work was supported by DEB-1210513, a Sigma Xi Grant in Aid of Research, and the ARCS foundation.

References

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 Evolutionary Biology* **14:** 268.

Bentley, K.E., Mandel, J.R. & McCauley, D.E. 2010. Paternal leakage and heteroplasmy of mitochondrial genomes in *Silene vulgaris*: evidence from experimental crosses. *Genetics* **185:** 961-968.

Birky, C.W. 1995. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. *Proceedings of the National Academy of Sciences* **92:** 11331-11338.

Birky, C.W. 2001. The inheritance of genes in mitochondria and chloroplast: laws, mechanisms, and models. *Annual Review of Genetics* **35:** 125-148.

Bogdanova, V.S. 2007. Inheritance of organelle DNA markers in a pea cross associated with nuclear-cytoplasmic incompatibility. *Theoretical and Applied Genetics* **114:** 333-339.

Burt, A. & Trivers, R. 2006. *Genes in conflict : the biology of selfish genetic elements*. Belknap Press of Harvard University Press, Cambridge, Mass.

Burton, R.S., Pereira, R.J. & Barreto, F.S. 2013. Cytonuclear genomic interactions and hybrid breakdown. *Annual Review of Ecology, Evolution, and Systematics, Vol 44* **44:** 281-302.

Chiu, W.L. & Sears, B.B. 1993. Plastome-genome interactions affect plastid transmission in *Oenothera*. *Genetics* **133:** 989-997.

Chou, J.Y. & Leu, J.Y. 2010. Speciation through cytonuclear incompatibility: Insights from yeast and implications for higher eukaryotes. *Bioessays* **32:** 401-411.

Corriveau, J.L. & Coleman, A.W. 1988. Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperm species. *American Journal of Botany* **75:** 1443-1458.

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

Cruzan, M.B., Arnold, M.L., Carney, S.E. & Wollenberg, K.R. 1993. cpDNA inheritance in interspecific crosses and evolutionary inference in Louisiana irises. *American Journal of Botany* **80:** 344-350.

Eberhard, W.G. 1980. Evolutionary consequences of intracellular organelle competition. *The Quarterly Review of Biology* **55:** 231-249.

Etterson, J.R., Keller, S.R. & Galloway, L.F. 2007. Epistatic and cytonuclear interactions govern outbreeding depression in the autotetraploid *Campanulastrum americanum*. *Evolution* **61:** 2671-2683.

Fishman, L. & Willis, J.H. 2006. A cytonuclear incompatibility causes anther sterility in *Mimulus* hybrids. *Evolution* **60:** 1372-1381.

Galloway, L.F., Etterson, J.R. & Hamrick, J.L. 2003. Outcrossing rate and inbreeding depression in the herbaceous autotetraploid, *Campanula americana*. *Heredity* **90:** 308-315.

Galloway, L.F. & Etterson, J.R. 2005. Population differentiation and hybrid success in *Campanula americana*: geography and genome size. *Journal of Evolutionary Biology* **18:** 81-89.

Greiner, S., Rauwolf, U., Meurer, J. & Herrmann, R.G. 2011. The role of plastids in plant speciation. *Molecular Ecology* **20:** 671-691.

Hansen, A.K., Escobar, L.K., Gilbert, L.E. & Jansen, R.K. 2007. Paternal, maternal, and biparental inheritance of the chloroplast genome in *Passiflora* (Passifloraceae): implications for phylogenetic studies. *American Journal of Botany* **94:** 42-46.

Hu, Y.C., Zhang, Q., Rao, G.Y. & Sodmergen 2008. Occurrence of plastids in the sperm cells of Caprifoliaceae: biparental plastid inheritance in angiosperms is unilaterally derived from maternal inheritance. *Plant and Cell Physiology* **49:** 958-968.

Hurst, L.D. 1992. Intragenomic conflict as an evolutionary force. *Proceedings of the Royal Society B-Biological Sciences* **248:** 135-140.

Jansen, R. & Ruhlman, T. 2012. Plastid genomes of seed plants. In: *Genomics of Chloroplasts and Mitochondria*, Vol. 35 (R. Bock & V. Knoop, eds), pp. 103-126. Springer Netherlands.

Knox, E.B. 2014. The dynamic history of plastid genomes in the Campanulaceae sensu lato is unique among angiosperms. *Proceedings of the National Academy of Sciences* **111:** 11097-11102.

Leppala, J. & Savolainen, O. 2011. Nuclear-cytoplasmic interactions reduce male fertility in hybrids of *Arabidopsis lyrata* subspecies. *Evolution* **65:** 2959-2972.

Levin, D.A. 2003. The cytoplasmic factor in plant speciation. *Systematic Botany* **28:** 5-11.

Levy, F. 1991. A genetic analysis of reproductive barriers in *Phacelia dubia*. *Heredity* **67:** 331-345.

Medgysey, P., Pay, A. & Marton, L. 1986. Transmission of paternal chloroplasts in *Nicotiana*. *Molecular and General Genetics* **204:** 195-198.

Mogensen, H.L. 1996. The hows and whys of cytoplasmic inheritance in seed plants. *American Journal of Botany* **83:** 383-404.

Rand, D.M., Haney, R.A. & Fry, A.J. 2004. Cytonuclear coevolution: the genomics of cooperation. *Trends in Ecology & Evolution* **19:** 645-653.

Sambatti, J.B.M., Ortiz-Barrientos, D., Baack, E.J. & Rieseberg, L.H. 2008. Ecological selection maintains cytonuclear incompatibilities in hybridizing sunflowers. *Ecology Letters* **11:** 1082-1091.

Sears, B.B. 1980. Elimination of plastids during spermatogenesis and fertilization in the plant kingdom. *Plasmid* **4:** 233-255.

Snijder, R.C., Brown, F.S. & van Tuyl, J.M. 2007. The role of plastome-genome incompatibility and biparental plastid inheritance in interspecific hybridization in the genus *Zantedeschia* (Araceae). *Floriculture and Ornamnetal Biotechnology* **1:** 150-157.

Ureshino, K., Miyajima, I., Ozaki, Y., Kobayashi, N., Michishita, A. & Akabane, M. 1999. Appearance of albino seedlings and ptDNA inheritance in interspecific hybrids of azalea. *Euphytica* **110:** 61-66.

Weihe, A., Apitz, J., Pohlheim, F., Salinas-Hartwig, A. & Borner, T. 2009. Biparental inheritance of plastidial and mitochondrial DNA and hybrid variegation in *Pelargonium*. *Molecular Genetics and Genomics* **282:** 587-593.

Zhang, Q., Liu, Y. & Sodmergen 2003. Examination of the cytoplasmic DNA in male reproductive cells to determine the potential for cytoplasmic inheritance in 295 angiosperm species. *Plant and Cell Physiology* **44:** 941-951.

Zhang, Q. & Sodmergen 2010. Why does biparental plastid inheritance revive in angiosperms? *Journal of Plant Research* **123:** 201-206.

Table 1: Chloroplast inheritance results for F1 hybrids from within- and between-clade crosses of *C. americanum*. Number of individuals genotyped and proportion of individuals with non-maternal inheritance are given for each cross and crossing direction separately. The maternal population is listed first for a given cross.

Type	Cross	N	Non-maternal	N	Non-maternal
		A1xA2		A2xA1	
Within clade	69x59	53	0.36	47	0.06
	44x20	49	0.31	60	0.10
	34x7	60	0.22	60	0.33
		AxC		CxA	
Between clade	68x73	60	0.47	6	0.17
	52x72	62	0.51	53	0.34
	10x72	62	0.42	23	0.26
	29x5	60	0.37	37	0.19
	86x92	61	0.26	59	0.19
	88x91	59	0.15	59	0.15

Figure 1: Phylogenetic tree and map depicting crosses used to estimate degree and consequences of non-maternal chloroplast inheritance. Maximum likelihood phylogenetic tree with 1000 bootstrap replicates based on chloroplast DNA sequence (Chapter 1). Location of populations used are marked on the map, and colored according to chloroplast clade. Solid represent genetically similar, within-clade crosses, while dashed lines represent genetically divergent, between-clade crosses.

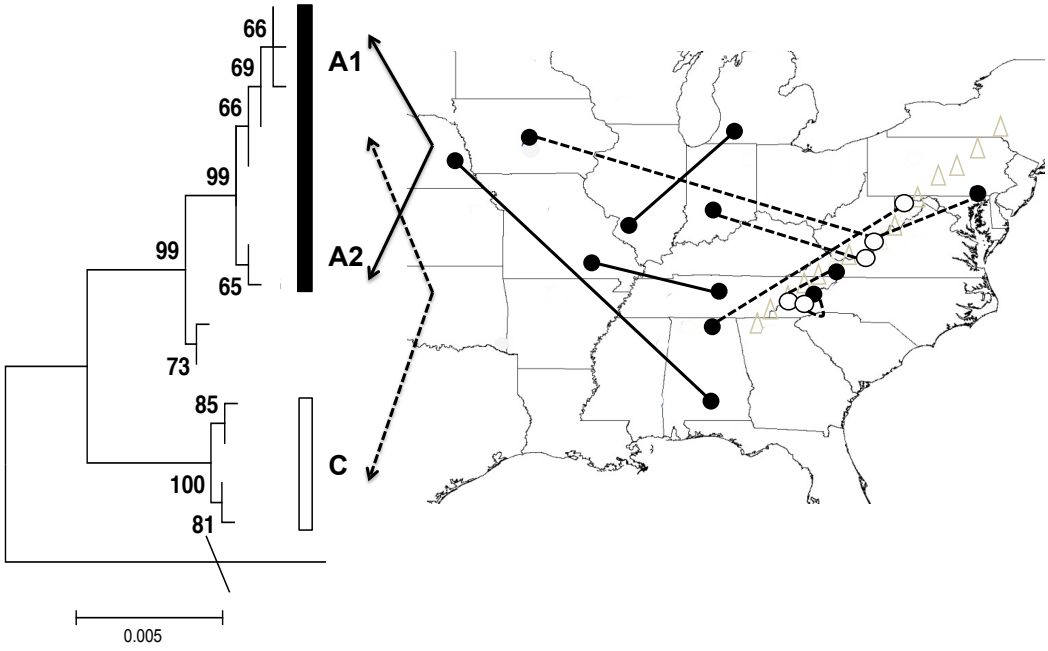


Figure 2: Examples of variegated, white, and chlorotic seedlings observed in genetically divergent *C. americanum* F1 hybrids.

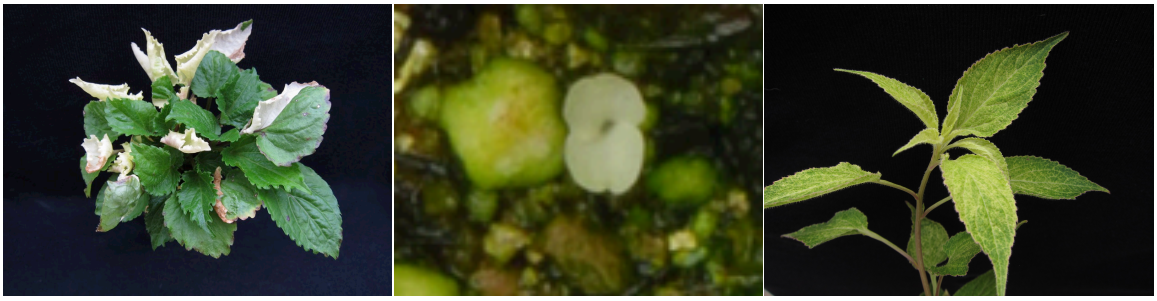


Figure 3: Distribution of chloroplast inheritance in F1 hybrids from within-clade crosses of *C. americanum*. Bars on the far left represent the frequency of individuals with maternal inheritance, while bars on the right represent the frequency of individuals with paternal inheritance. Bars in-between represent the frequency of individuals with varying proportions of the maternal and paternal chloroplasts (biparental inheritance). Crossing direction where A1 populations were maternal (A) and A2 populations were maternal (B).

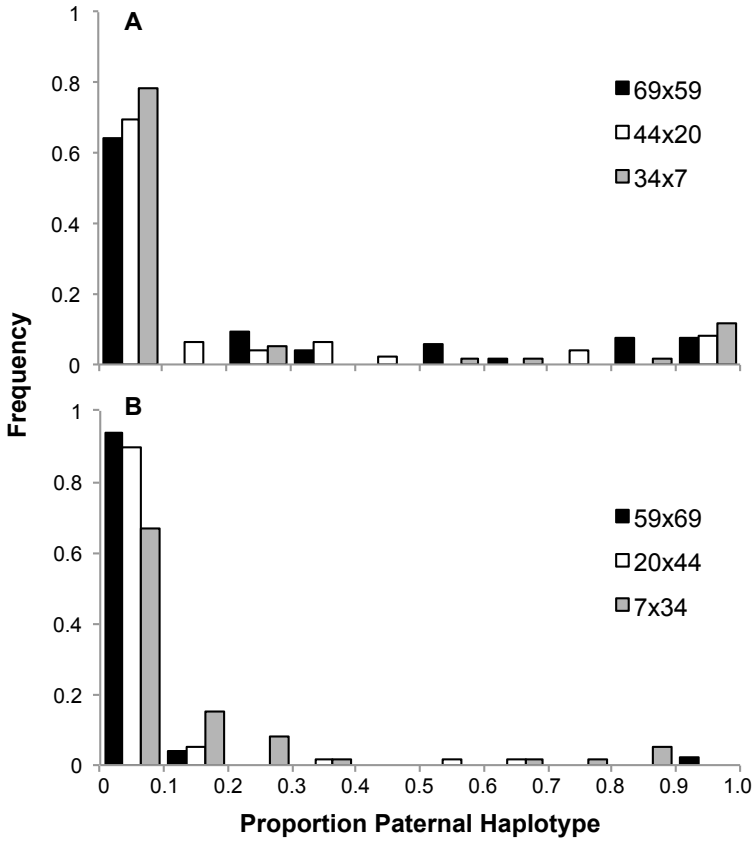


Figure 4: Distribution of chloroplast inheritance in F1 hybrids from between-clade crosses of *C. americanum*. Bars on the far left represent the frequency of individuals with maternal inheritance, while bars on the right represent the frequency of individuals with paternal inheritance. Bars in-between represent the frequency of individuals with varying proportions of the maternal and paternal chloroplasts (biparental inheritance). Crossing direction where populations with chloroplast A were maternal (A) and chloroplast C were maternal (B).

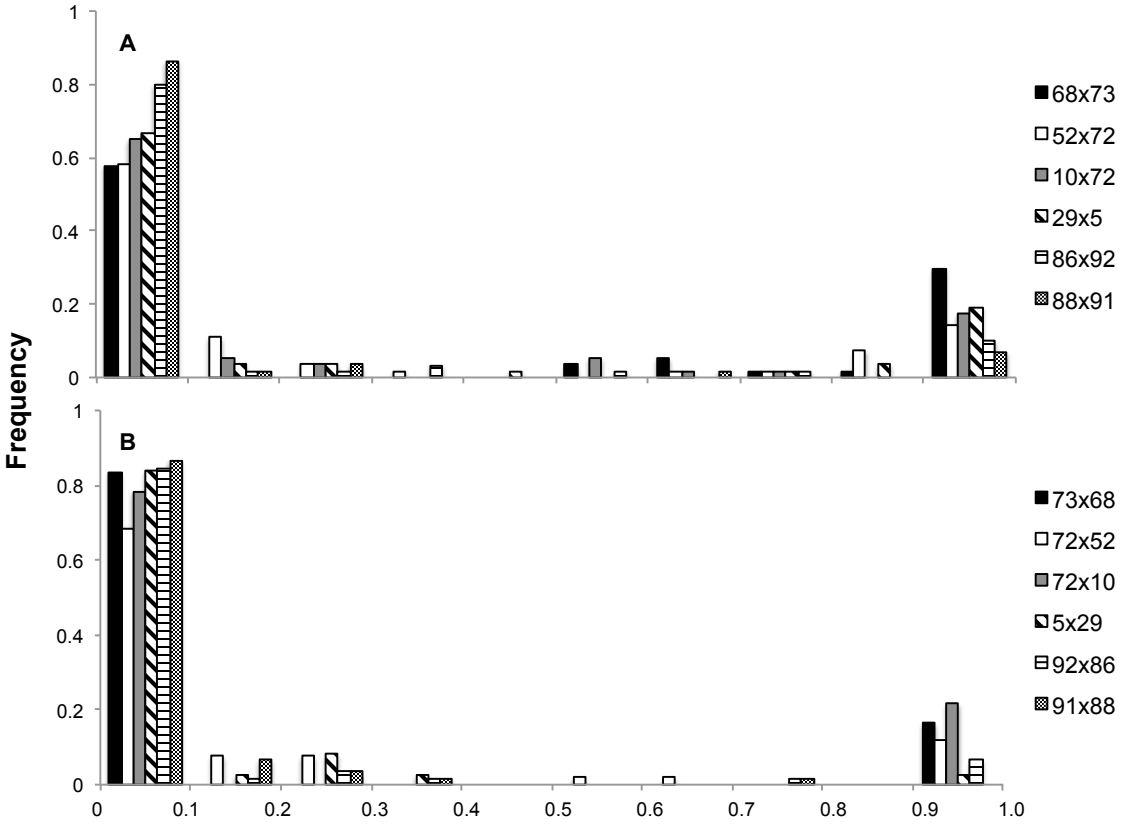


Figure 5: Correlation between the likelihood of between-clade *C. americanum* F1 hybrids surviving when only inheriting haplotype A and either A) the proportion of maternal inheritance in F1 hybrids or B) the proportionate loss of haplotype A in the F2 relative to the F1. Each point represents the mean of a genetically divergent cross. Only data from the crossing direction where populations with chloroplast A were maternal is included.

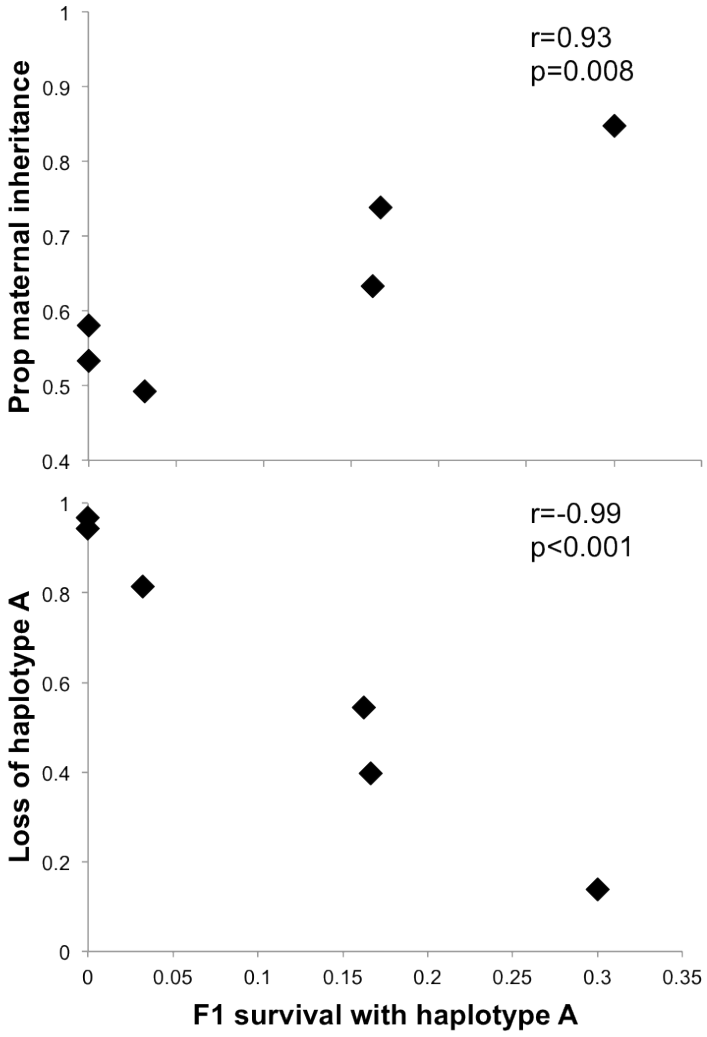


Figure 6: Survival and fitness of each chloroplast haplotype through F2 survival for *C. americanum* F1 hybrids from between-clade crosses. Only data from the crossing direction where populations with chloroplast A were maternal is included. Cross IDs are listed across the bottom of the graphs. A) Number of F1 hybrid individuals that survived or died for each genetically divergent cross, according to chloroplast haplotype. White bars outlined in grey represent the number of individuals that died for each category. B) The estimated proportion of surviving F2 offspring for each genetically divergent cross that originated from F1 progeny containing the alternate chloroplast haplotypes.

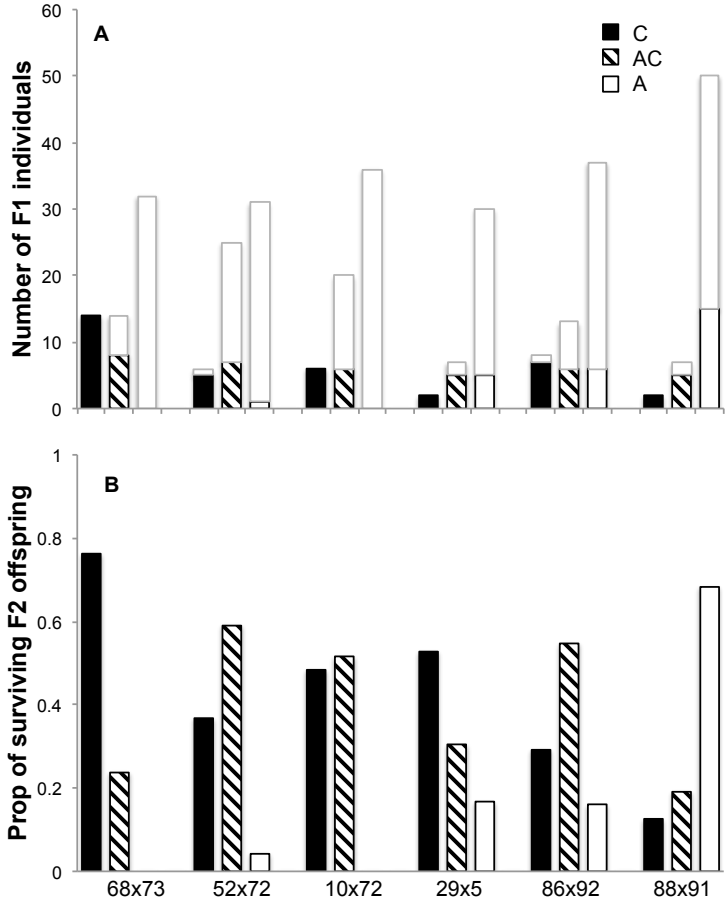


Figure 7: Comparing genotypes of basal and apical leaf tissue samples from F1 hybrids from between-clade crosses of *C. americanum*. Pie charts represent the apical leaf tissue genotypes, while the letters below each graph represent the basal leaf tissue genotypes.

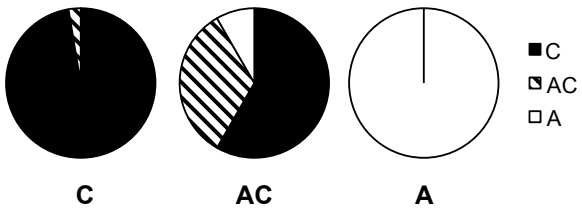


Figure 8: Proportion survival of *C. americanum* F1 and F2 generation hybrids for each between-clade cross. Data shown for the crossing direction where populations with haplotype A were maternal. a: p<0.07; *: p<0.5; ***: p<0.001.

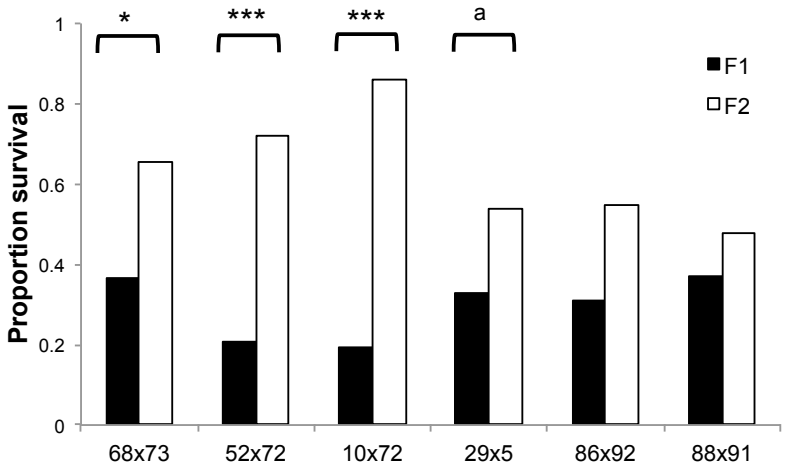
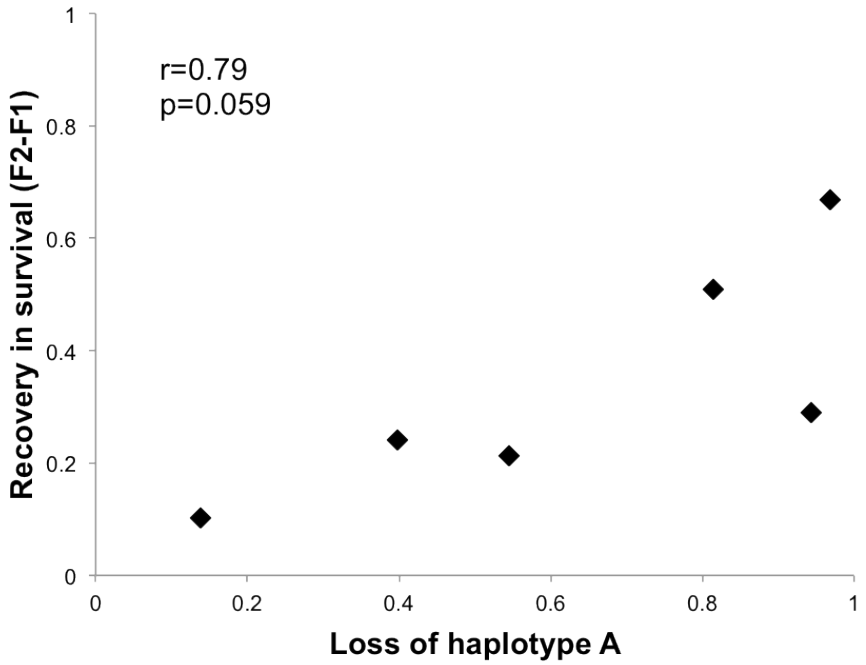


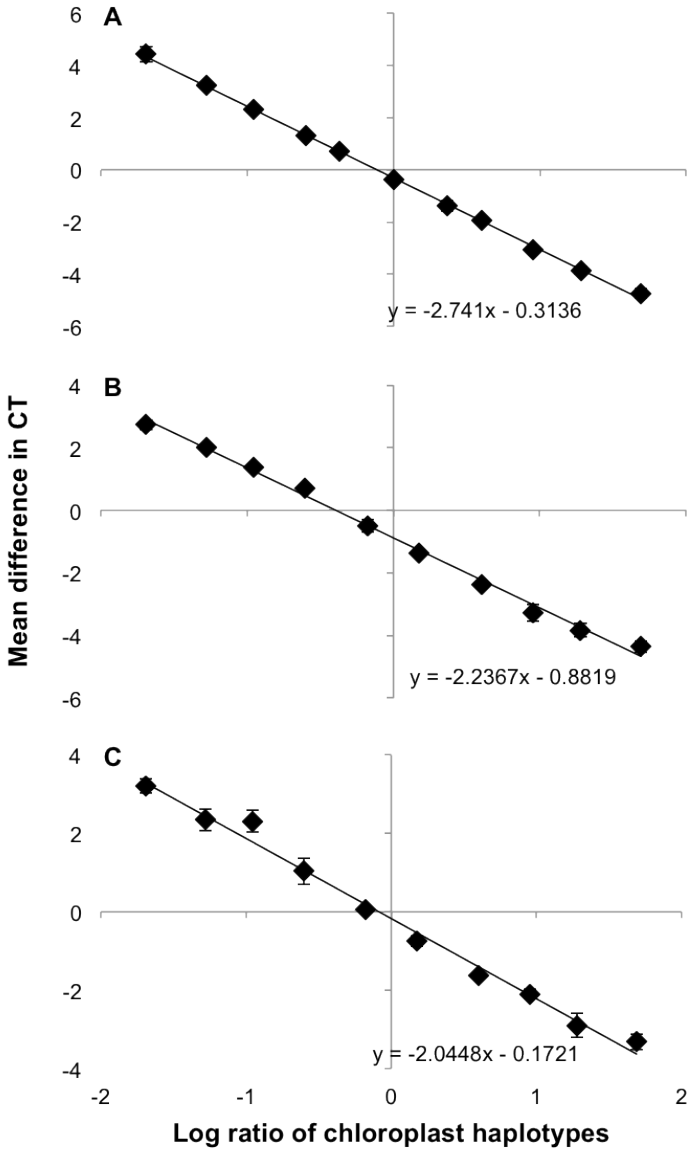
Figure 9: Relationship between the proportionate loss of haplotype A in F2 *C. americanum* hybrids relative to the F1 and degree of recovery in survival. Each point represents a between-clade cross.



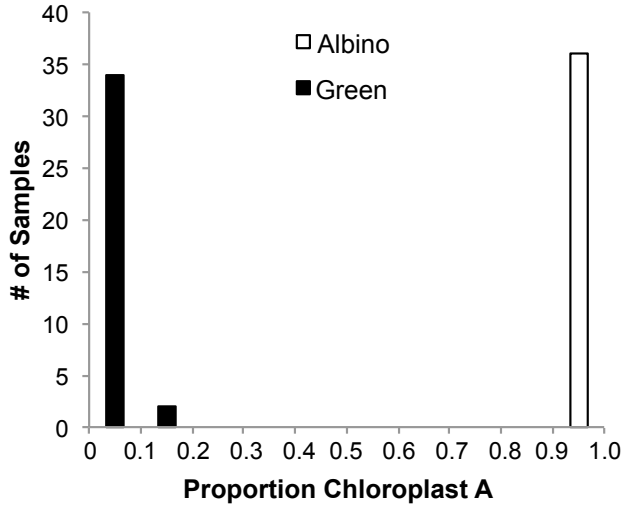
Supplementary Table 1: Geographic location and chloroplast haplotype for the *C. americanum* populations used in within-clade (haplotypes A1 and A2) and between-clade (haplotypes A and C) crosses to examine patterns and consequences of non-maternal chloroplast inheritance. All populations were located in the eastern United States.

Pop ID	State	Latitude	Longitude	Haplotype
5	Maryland	39.6137	-79.1158	C
7	Missouri	37.1328	-91.2772	A2
10	Iowa	42.0728	-93.6725	A
20	Illinois	38.6074	-89.8992	A2
29	Alabama	34.6514	-86.5017	A
34	Tennessee	36.0822	-86.2961	A1
44	Michigan	42.2909	-85.5942	A1
52	Pennsylvania	39.9406	-76.3464	A
59	Nebraska	41.1609	-96.5393	A2
68	Indiana	39.1643	-86.5269	A
69	Alabama	31.8485	-86.6402	A1
72	West Virginia	37.9931	-80.3618	C
73	Virginia	37.3534	-80.5522	C
86	Virginia	36.6344	-81.5881	A
88	Tennessee	35.9833	-82.4989	A
91	North Carolina	35.5862	-83.0663	C
92	Tennessee	35.6758	-83.5259	C

Supplementary Figure 1: Graphs showing standard curves used for calculating chloroplast ratios according to CT values from RT-PCR Taqman SNP Assays. A) Used for crosses 44x20, 34x7, 72x10, 29x5, and 88x91. B) Used for crosses 52x72, 10x72, and 86x92. C) Used for cross 59x69.



Supplementary Figure 2: Genotyping results of white and green leaf tissue punches taken from 36 variegated F1 hybrids from between-clade crosses of *C. americanum*.



Chapter 5:

Correlation between sequence divergence and polymorphism reveals similar evolutionary mechanisms acting across multiple timescales in a rapidly evolving plastid genome

As published: Barnard-Kubow, K.B., Sloan, D.B., and 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 Evolutionary Biology. 14:268

Abstract

Background: Although the plastid genome is highly conserved across most angiosperms, multiple lineages have increased rates of structural rearrangement and nucleotide substitution. These lineages exhibit an excess of nonsynonymous substitutions (i.e., elevated dN/dS ratios) in similar subsets of plastid genes, suggesting that similar mechanisms may be leading to relaxed and/or positive selection on these genes. However, little is known regarding whether these mechanisms continue to shape sequence diversity at the intraspecific level.

Results: We examined patterns of interspecific divergence and intraspecific polymorphism in the plastid genome of *Campanulastrum americanum*, and across plastid genes found a significant correlation between dN/dS and pN/pS (i.e., the within-species equivalent of dN/dS). A number of genes including *ycf1*, *ycf2*, *clpP*, and ribosomal protein genes exhibited high dN/dS ratios. McDonald-Kreitman tests detected little evidence for positive selection acting on these genes, likely due to the presence of substantial intraspecific divergence.

Conclusions: These results suggest that mechanisms leading to increased nucleotide substitution rates in the plastid genome are continuing to act at the intraspecific level. Accelerated plastid genome evolution may increase the likelihood of intraspecific cytonuclear genetic incompatibilities, and thereby contribute to the early stages of the speciation process.

Keywords: Plastid, Selection, dN/dS, pN/pS, Reproductive Isolation, Intraspecific, Chloroplast, Sequence Evolution

Background

The structure of the plastid genome is generally conserved across the angiosperms (Jansen & Ruhlman, 2012), and its nucleotide substitution rates are usually low relative to the nuclear genome (Drouin *et al.*, 2008). However, multiple lineages exhibit extensively rearranged plastid genomes as well as increased rates of nucleotide substitution and elevated dN/dS ratios for some genes (Jansen *et al.*, 2007). While this connection suggests the potential for a common underlying cause of structural instability and increased nucleotide substitution rates (Jansen *et al.*, 2007), it is important to note that while structural instability impacts the plastid genome as whole, the increase in substitution rate appears to vary depending on gene function (Guisinger *et al.*, 2008). In lineages that show increased substitution rates, similar sets of plastid genes have experienced increased substitution rates and elevated dN/dS ratios, suggesting the possibility of common mechanisms or selective regimes acting on these genes in independent angiosperm groups. However, it remains unclear whether these patterns reflect positive selection, relaxed purifying selection, changes in underlying mutation rates, a breakdown in DNA repair mechanisms such as gene conversion (Khakhlova & Bock, 2006), or some combination of these.

Accelerated evolution in the plastid genome of some angiosperm lineages raises the question as to whether the mechanisms responsible continue to operate at the intraspecific level. Numerous studies examining nucleotide substitution rates and potential signatures of selection in the plastid genome, based on variation in dN/dS, have been carried out based on divergence among species (Erixon & Oxelman, 2008; Greiner *et al.*, 2008; Guisinger *et al.*, 2008; Sloan *et al.*, 2012; Weng *et al.*, 2012; Sloan *et al.*,

2014a), but data describing genome-wide intraspecific sequence variation are needed to investigate whether accelerated plastid genome evolution is occurring within species. Estimates of intraspecific polymorphism are also useful for interpreting interspecific divergence, as it allows for distinguishing between the effects of positive selection and relaxed purifying selection on nucleotide substitution rates based on changes in the relative ratio of non-synonymous to synonymous changes before and after selection has acted (Hudson *et al.*, 1987; McDonald & Kreitman, 1991). Accordingly, positive selection is expected to lead to a significantly higher ratio for interspecific divergence than for intraspecific polymorphism.

We examined patterns of sequence divergence and polymorphism in *Campanulastrum americanum* to determine whether similar mechanisms of plastid genome evolution are acting within as well as between species. This species is a good study system in which to address these questions as it is in the Campanulaceae, a family in which the taxa have highly rearranged plastid genomes (Cosner *et al.*, 2004; Knox, 2014), increasing the likelihood of detecting intraspecific accelerated plastid evolution. In particular we sought to answer the following questions. 1) Do a similar set of plastid genes exhibit increased nucleotide substitution rates and elevated dN/dS ratios in *C. americanum* as found in previous studies with other species? 2) Are similar patterns found when examining plastid sequence variation among populations within *C. americanum*? 3) Do we find evidence for positive selection leading to increased substitution rates and elevated dN/dS ratios in these plastid genes?

Methods

Study System

Campanulastrum americanum is a monocarpic herb found in the eastern half of the United States. Individuals are autotetraploid, annual or biennial, and primarily outcrossing (Galloway *et al.*, 2003; Galloway & Etterson, 2005). The Campanulaceae has been shown to have highly rearranged plastid genomes as well as the potential for biparental plastid inheritance (Corriveau & Coleman, 1988; Zhang *et al.*, 2003). Crossing studies in *C. americanum* have found that while inheritance is primarily maternal, biparental and paternal inheritance occurs in roughly 25% of offspring (Barnard-Kubow, unpublished results). However, plastid polymorphism within populations appears relatively low, with genotyping at five loci (including portions of ycf1, rps2, and rps4) finding individuals within a population to be generally fixed for plastid haplotype (Chapter 1). Therefore, while biparental inheritance may complicate full assembly of the plastid genome when using a maternal family, it is unlikely to cause significant error in terms of estimates of polymorphism or in determining the presence/absence of plastid genes.

Sample Material and Library Construction

For sequencing the plastid genome of *C. americanum*, 180 grams of fresh leaf tissue was collected from multiple individuals from a single maternal family (i.e., seeds from a single plant) from a population in Virginia (Table 1). Individuals were germinated in a growth chamber from field-collected seed and grown for several months in the greenhouse with regular watering and fertilization.

For examining within species polymorphism of the plastid genome, 150 grams of fresh leaf tissue was collected and pooled from multiple individuals from four populations of *C. americanum* (VA, MN, OH, and AL), including the same VA population used for the single population plastid sequencing (Table 1). These populations were chosen because they span the geographic range of *C. americanum* and were known to differ genetically based on sequencing of individual chloroplast loci (Chapter 1). VA individuals were transplanted from the field, while MN, OH, and AL individuals were germinated from field-collected seed in a growth chamber. Plants from all four populations were then grown for several months in the greenhouse with regular watering and fertilization. Intact chloroplasts were isolated from the single population (VA) and pooled samples using a combination of differential centrifugation and separation on a sucrose step gradient (Palmer, 1986; Jansen *et al.*, 2005). Chloroplasts were then lysed, and DNA was obtained via a phenol-chloroform extraction and ethanol precipitation. The purity of plastid DNA (cpDNA) was confirmed by restriction digestion.

For each plastid sample, shotgun libraries were constructed with multiplex identifier (MID) tags following standard protocols for sequencing on a Roche 454 GS-FLX platform with Titanium reagents. MID-tagged libraries were sequenced as part of a larger pooled sample. All 454 library construction and sequencing steps were performed at the Genomics Core Facility in the University of Virginia's Department of Biology. A total of 28,694 and 24,552 sequencing reads were obtained from the VA and pooled libraries, respectively. The mean sequence lengths were 335 bp for the VA library and 339 bp for the pooled library. The reads from each library were deposited in NCBI's Short Read Archive [SRX595708 and SRX595709].

Plastid Assembly and Annotation

454 reads for the single population sample were assembled using Roche's GS de novo Assembler v2.3 ("Newbler") using default settings. Initial assembly produced hundred of contigs, however many of these were identified as bacterial or nuclear contamination. By visualizing the remaining contigs in Consed v21 (Gordon *et al.*, 1998) and using information regarding reads that span multiple contigs, 63 of the initial contigs were re-assembled into nine final contigs with a total length of 147.3 kb and an average single copy coverage depth of 20×. For the *ccsA* gene, PCR and Sanger sequencing were used to obtain sequence spanning a gap and complete the full sequence.

DOGMA (Wyman *et al.*, 2004) was used to annotate the protein, transfer RNA (tRNA), and ribosomal RNA (rRNA) genes for each of the contigs. One gene, *clpP*, exhibited high sequence divergence in the first exon. To determine the full sequence of the gene, correctly identify the exon/intron boundaries, and confirm transcription of the gene, *clpP* was amplified from cDNA constructed from an individual from the same VA population. Another gene, *ycf1*, also exhibited high sequence divergence and appeared to have multiple frameshift mutations. However, these frameshift mutations were in long homopolymer or repetitive regions, raising the possibility they were due to 454 sequencing errors. PCR and Sanger sequencing confirmed that the frameshifts were the result of homopolymer-related sequencing errors. The corrected sequence yielded an intact *ycf1* reading frame. The final annotated contig sequences were deposited to GenBank under accession [GenBank:KJ920499-KJ920507].

Interspecific Divergence in cpDNA Sequence

To estimate divergence and dN/dS ratios for plastid coding genes, the following species were used as outgroups and the corresponding gene sequences were obtained from GenBank: *Trachelium caeruleum* in the Campanulaceae [GenBank:NC_010442], and two more distantly related species, *Helianthus annuus* [GenBank:NC_007977] and *Nicotiana tabacum* [GenBank:NC_001879]. Outgroups were chosen to span a range of phylogenetic distances with one, *T. caeruleum*, in the Campanulaceae, and another, *H. annuus*, in the Asterales. Sequences were aligned with MUSCLE (Edgar, 2004) as implemented in Codon Code Aligner v3.5 (CodonCode Corporation). High sequence divergence was observed for both *ycf1* and *ycf2*, necessitating the deletion of large regions of unalignable sequence. An average of 28 regions were removed per outgroup, with deletions averaging 78 bp and ranging up to 465 bp. Therefore, the resulting divergence values for these genes are underestimated. For *clpP*, the *T. caeruleum* sequence was re-annotated using DOGMA to locate the gene in the full chloroplast sequence obtained from GenBank and using homology between *C. americanum*, *H. annuus*, and *N. tabacum* to designate the intron/exon boundaries (Additional file 1). Gene alignments were deposited in Dryad (Barnard-Kubow *et al.*, 2014).

The relative rates of sequence divergence and dN/dS ratio were determined for the protein coding genes using codon-based models of evolution in PAML v4.4 (Yang, 2007). All analyses implemented a constrained topology with *T. caeruleum* and *C. americanum* monophyletic relative to *H. annuus* and *N. tabacum*, as *T. caeruleum* and *C. americanum* are within the same family. Codon frequencies were determined by an F3×4 model. The parameter values for dN/dS and transition/transversion ratio were estimated

from the data with initial values of 0.4 and 2 respectively. Separate dN/dS values were estimated for each branch. Analyses were run on separate concatenations for each of the following sets of protein genes: 1) ATP synthase (*atp*), 2) NADH-plastoquinone oxidoreductase (*ndh*), 3) cytochrome b6/f complex (*pet*), 4) photosystem I (*psa*), 5) photosystem II (*psb)*, 6) large ribosomal subunit (*rpl)*, 7) small ribosomal subunit (*rps*), and 8) RNA polymerase (*rpo*), as well as the following individual protein genes: *ccsA*, *cemA*, *clpP*, *matK*, *rbcL*, *ycf1*, *ycf2*, *ycf3*, and *ycf4*. See Additional file 2: Table S1 for a list of specific genes included in each concatenation. The *psbT* gene was excluded from the analysis because it was multicopy in *C. americanum*, and *petN* was excluded because the *T. caeruleum* copy was unalignable. PAML estimated a dS value of zero for *clpP* on the terminal branch for *C. americanum*, resulting in an undefined dN/dS ratio. For subsequent analyses we estimated dS and dN/dS for *clpP* assuming a single synonymous substitution. All PAML files used were deposited in Dryad (Barnard-Kubow *et al.*, 2014).

We tested for signatures of positive selection (defined by a dN/dS value significantly greater than one) by constraining the dN/dS ratio to one for the terminal branch leading to *C. americanum* for any genes where this branch had an initial estimated dN/dS ratio greater than one. Separate dN/dS values were estimated for each of the remaining branches. Likelihood ratio tests were used to compare the constrained and unconstrained analyses and determine if the estimated dN/dS ratios were significantly greater than one (Yang, 1998). We applied a Bonferroni correction factor of 17 to account for multiple comparisons (17 genes/concatenations). To further examine sequence divergence and the potential for positive selection in one gene that showed a high dN/dS ratio, *clpP*, maximum likelihood trees were constructed for *clpP* intronic and

exonic sequence separately using *Arabidopsis thaliana* as an outgroup. Trees were constructed using baseml in PAML v4.4 (Yang, 2007) with a fixed topology and a GTR model of evolution based on results from jModelTest v2.1.5 (Guindon & Gascuel, 2003; Darriba *et al.*, 2012).

Intraspecific Polymorphism in cpDNA Sequence

To identify within species polymorphism and estimate pN/pS (within species equivalent of dN/dS), 454 reads from the pooled multiple-population *C. americanum* sample were mapped to the assembled contigs from the VA sample using Roche's GS Reference Mapper v2.5.3. The mapped reads had an average single copy coverage depth of 25×. A Perl script was written to use annotated gene locations and SNP information from the "HCDiff" mapping file to extract all high-confidence SNPs and identify them as genic/intergenic, exonic/intronic, and non-synonymous/synonymous (Additional File 3). Total numbers of non-synonymous and synonymous SNPs were tallied for each set of concatenated or individual genes, using the same concatenation groupings as used when estimating dN/dS. To estimate pN, pS, and the pN/pS ratio, the nonsynonymous and synonymous polymorphism counts were divided by the number of nonsynonymous and synonymous sites determined by PAML in our dN/dS analyses.

Polymorphism and Divergence

We determined whether similar sets of genes have an elevated dN/dS and pN/pS ratio by running a correlation analysis on the log transformed dN/dS and pN/pS ratios from each gene or concatenation (PROC CORR, SAS 9.3, SAS Institute, INC. 2011). Only genes or

concatenations that had three or more SNPs were included in the correlation analysis (n=11).

Additionally, to test whether genes with elevated dN/dS ratios have been under positive selection versus relaxed purifying selection, McDonald-Kreitman (M-K) tests (McDonald & Kreitman, 1991) were run using Fisher's exact test for each set of concatenated or individual genes (PROC FREQ, SAS 9.3, SAS Institute, INC. 2011). Pairwise divergence data between *C. americanum* and each of the three outgroup species (*T. caeruleum*, *H. annuus*, and *N. tabacum*) were obtained using a Perl script to extract all pairwise SNPs from the four species gene alignments used in the plastid divergence analysis above (Additional File 4). We applied a Bonferroni correction factor of 51 to account for multiple comparisons (17 genes/concatenations x 3 outgroup species).

Results

Plastid Genome Assembly

Sequencing of purified cpDNA from *C. americanum* produced high depth coverage of the plastid genome, which was assembled into nine contigs, totaling 147.3 kb. A complete assembly of *C. americanum's* plastid genome was unattainable due to the presence of a large number of repetitive regions. These findings fit with those from the plastid genome of *T. caeruleum*, another member of the Campanulaceae, which also contains an unusually high level of repeats (Haberle *et al.*, 2008). The plastid genomes of the Campanulaceae family have also been found to contain many inversions (Cosner *et al.*, 2004; Haberle *et al.*, 2008; Knox, 2014). Likely this propensity for structural instability explains why mapping *C. americanum* reads to *T. caeruleum's* plastid genome

was not helpful in further assembly of *C. americanum's* plastid genome. The potential for biparental inheritance of the plastid genome and subsequent heteroplasmy may have further contributed to difficulties in assembly when using a single maternal family.

Several genes commonly found in other angiosperms appear to be missing or presumed non-functional in *C. americanum*. No evidence was found for *accD*, even when searching against the raw reads, while *infA* is likely a pseudogene due to multiple internal stop codons. Only a 50 bp fragment remains of *rpl23*, though this fragment is present in at least two locations. The *rps16* intron has also been lost. These three genes and the *rps16* intron have also been lost or are presumed non-functional in *T. caeruleum* (Haberle *et al.*, 2008), suggesting these losses occurred prior to the divergence of these two species within the Campanulaceae. The loss of *accD* fits with evidence for this gene having been transferred to the nuclear genome in the Campanulaceae (Rousseau-Gueutin *et al.*, 2013). The *accD*, *rpl23* and *infA* genes have also been independently lost from the plastid genome in multiple other angiosperm lineages (Zurawski & Clegg, 1987; Millen *et al.*, 2001; Delannoy *et al.*, 2011).

Several gene duplications have also occurred, which appear unique to *C. americanum*, though similar in pattern to those observed in *T. caeruleum (Haberle et al., 2008)*. There has been a 300 bp partial duplication of *psbB* and a partial duplication of the beginning of *rrn16* upstream of the full *rrn16* gene. The *ndhF* gene has experienced multiple partial duplications including two identical 100 bp duplications as well as a separate 70 bp duplication. Several tRNAs were also duplicated, with two tandem copies of *trnM-CAU*, two copies of *trnS-GCU*, and three copies of both *trnfM-CAU* (two in

tandem) and *trnL-CAA* spread throughout the genome. Duplications of tRNAs were also found in *T. caeruleum* where *trnI-CAU* is present in two copies (Haberle *et al.*, 2008).

One plastid gene, *psbT*, has undergone multiple duplication events, leading to three full-length copies in the *C. americanum* genome. One of these copies is highly conserved and retains the ancestral amino acid sequence when compared to *T. caeruleum*, *H. annuus*, and *N. tabacum*, while the other two copies have accumulated multiple amino acid changes. Again, a similar phenomenon was observed in *T. caeruleum*, where a different photosystem II gene, *psbJ,* is present in three copies (Haberle *et al.*, 2008).

Interspecific Divergence in cpDNA Sequence

The dN/dS ratio varied widely across *C. americanum's* plastid genes, suggesting these genes are experiencing different selective regimes (Fig 1, Table 2). The photosynthesis genes exhibited evidence of strong purifying selection, as indicated by their low dN/dS ratios (Table 2), while *ycf1*, *ycf2*, *clpP*, and the small subunit ribosomal protein genes had elevated dN/dS ratios close to or above one, suggesting the possibility of relaxed purifying selection, positive selection, or a mixture of both (Table 2, Fig 1). These genes also varied in the extent to which changes in dN or dS led to the elevated dN/dS ratios, suggesting that the selective regime leading to the elevated ratios may not be consistent across genes. Relative to other genes, *ycf1* and *ycf2* exhibited both an elevated dN and dS, while the small subunit ribosomal protein genes had only an elevated dN (Fig 1) The *clpP* gene exhibited a moderate increase in dN as well as a greatly reduced dS (Fig 1). The estimated dS of zero for *clpP* is likely to be a statistical anomaly due to the short sequence length providing a limited number of sites at which synonymous substitutions

can occur. In addition, the branch leading to *T. caeruleum* and *C. americanum* exhibited an accelerated substitution rate and a dN/dS ratio greater than one for *clpP* (Additional file 5: Table S2). Therefore, the high dN/dS ratio found for *clpP* in *C. americanum* does not appear to be an artifact of the low estimated dS for this gene.

The high dN/dS ratio suggests that positive selection may be acting on *clpP*. Further support for positive selection was identified when comparing branch lengths inferred from *clpP* intronic and exonic sequence. The exonic tree showed a greatly increased branch length (> 7×) on the branch leading to *C. americanum* and *T. caeruleum* relative to the intronic tree (Fig 2). In contrast, the branch length on the branches leading to *H. annuus* and *N. tabacum* were shorter in the exonic tree relative to the intronic tree. These results are similar to those found in the tribe Sileneae (Caryophyllaceae), where evidence for positive selection on *clpP* was also observed when comparing intron and exon tree branch lengths (Erixon & Oxelman, 2008). The increased length and high dN/dS ratio (1.23) found for the branch leading to *C. americanum* and *T. caeruleum* suggest that altered selection was likely acting on *clpP* prior to the split of these two species (Additional file 5: Table S2).

Intraspecific Polymorphism in cpDNA Sequence

We detected a total of 174 high confidence SNPs in 62853 bp of total protein coding sequence from the *C. americanum* plastid genome. As with divergence, the pN/pS ratio also varied strongly across plastid genes, suggesting that they are continuing to experience differing selective regimes at the within-species level (Fig 1, Table 2). Similar to the divergence results, the photosynthetic genes appear to be under purifying selection,

while *ycf1*, *ycf2*, and the small subunit ribosomal genes had signatures of relaxed purifying or positive selection as evidenced by their elevated pN/pS ratios with values greater than one (Table 2, Fig 1). The elevated pN/pS ratios for these genes were primarily due to a higher pN, with the small subunit ribosomal genes in particular having a pN that is at least twice as high as for any other single gene or concatenation (Fig 1). On the other hand, *clpP* had a pN/pS ratio of zero, but that is based on only a single identified SNP in *C. americanum* (Fig 1, Table 2).

The patterns of polymorphism also suggested the existence of structural variation in the plastid genome within *C. americanum*. The initial polymorphism data indicated multiple non-synonymous SNPs within the first exon of *clpP*. Further examination of the pooled sequence data found evidence for a duplication of the first exon of *clpP* that did not appear to be present in the single population plastid assembly. Primers were then designed to amplify the first exon from either the full copy or duplication of *clpP*. The full copy first exon was amplified in all four populations used for sequencing the plastid genome (VA, AL, OH, and MN), while the duplication was only amplified in AL, OH, and MN, indicating the duplication does not exist in the VA population. These findings suggest the duplication event occurred since the divergence of the *C. americanum* populations. Sequencing and alignment of the partial duplication and full copy of *clpP* (deposited in Dryad [24]) recovered the nonsynonymous SNPs found in the initial polymorphism data, indicating they were artifacts caused by mapping the partial duplication to the full VA *clpP* copy.

Polymorphism and Divergence

Overall the dN/dS and pN/pS ratios are correlated across *C. americanum's* plastid genes (Fig 3), suggesting that similar selective pressures are acting on the same genes across multiple time scales. The dN/dS ratios on the unrooted branch connecting *C. americanum* and *T. caeruleum* to *H. annuus* and *N. tabacum* are also correlated with *C. americanum's* dN/dS ratios (R^2=0.88, $p < 0.001$), further supporting similar selective pressures acting across time scales (Additional file 5: Table S2). While *clpP* seemed to deviate from this general pattern in that it showed a strongly elevated dN/dS ratio but a pN/pS ratio of zero, there is limited confidence in this estimate due to the short sequence length of *clpP* and the ratio being based only upon a single synonymous SNP.

In general, the M-K tests found no significant difference between the number of non-synonymous and synonymous SNPs when comparing the polymorphism and divergence data. The one exception was the set of the small subunit ribosomal proteins when using either *H. annuus* or *N. tabacum* as an outgroup. These analyses had a significant difference before, but not after, Bonferroni correction. In both of these comparisons, there was evidence for purifying selection, as the divergence among species had a lower ratio of non-synonymous to synonymous changes than the polymorphism within *C. americanum*. When using *T. caeruleum* as an outgroup, however, there was no longer a significant difference between the polymorphism and divergence SNP ratios for the small subunit ribosomal proteins.

Discussion

Previous studies have found the Campanulaceae *sensu lato* to have relatively unstable plastid genomes characterized by a high frequency of inversions, the presence of

repetitive regions, as well as gene duplications (Cosner *et al.*, 2004; Haberle *et al.*, 2008; Knox, 2014). Fitting these earlier findings, assembly and annotation of *C. americanum's* plastid genome revealed the presence of a number of repetitive regions as well as gene duplications. Further evidence for instability is suggested by a variable partial duplication of *clpP* among populations of *C. americanum*.

While the photosynthetic genes exhibited evidence for strong purifying selection, elevated nucleotide substitution rates and dN/dS ratios were found for *ycf1*, *ycf2*, *clpP*, and the small subunit ribosomal genes in *C. americanum*. The *ycf1* gene was recently found to be involved in protein translocation (Kikuchi *et al.*, 2013), while *ycf2* is essential for cell viability but of unknown function (Drescher *et al.*, 2000), and *clpP* codes for a protease subunit. The PAML analyses used average dN/dS across the entire gene (or gene concatenation), making this a conservative test for positive selection (dN/dS significantly greater than 1). If only a subset of codons within a gene/concatenation were under positive selection, the analyses used would have been unlikely to detect this signature. Accordingly, it is possible that some of the genes that exhibited evidence of purifying selection may have had positive selection at a subset of sites. At the same time, the genes with elevated average dN/dS ratios may have even higher dN/dS ratios at specific sites.

These genes with an elevated dN/dS ratio also differed in whether this was due to underlying changes in dN or dS, suggesting that the selective regime leading to the elevated ratios may not be consistent across genes. As synonymous substitutions are neutral, changes in dS are likely to reflect changes in the underlying mutation rate, potentially due to problems with DNA repair, whereas rates of non-synonymous substitutions are impacted not only by the underlying mutation rate, but also selection.

Therefore changes in dN can also give insight into changes in selection. Relative to other genes, *ycf1* and *ycf2* exhibit both an elevated dN and dS, though the increase in dN was greater than dS (2.5× and 1.5× higher respectively), suggesting that the underlying mutation rate, as well as potentially the selective regime, has been altered in the these genes. In contrast, the small subunit ribosomal protein genes had only an elevated dN (Fig 1), strongly suggesting a change in selective regime, and not underlying mutation rate has led to the elevated dN/dS ratio in these genes. The *clpP* gene exhibits a moderate increase in dN as well as a greatly reduced dS. However, the uncertainties regarding the estimate of dS for this gene make it difficult to come to any conclusions regarding the underlying causes of the elevated dN/dS ratio.

A similar set of plastid genes (including *ycf1*, *ycf2*, *clpP*, and ribosomal protein genes) have increased nucleotide substitution rates and elevated dN/dS ratios in other taxa (Jansen *et al.*, 2007), including species within the tribe *Sileneae* (Erixon & Oxelman, 2008; Sloan *et al.*, 2012; Sloan *et al.*, 2014a), the genus *Oenothera* (Erixon & Oxelman, 2008; Greiner *et al.*, 2008), and the Geraniaceae (Guisinger *et al.*, 2008; Weng *et al.*, 2012). Increased nucleotide substitution rates have also been associated with increased structural variability (Jansen *et al.*, 2007), similar to our findings in *C. americanum.* These similarities suggest the possibility of a common evolutionary mechanism, whether adaptive or non-adaptive (Sloan *et al.*, 2014a). Our results suggest that this mechanism is continuing to operate at very recent time scales because we detect similar accelerated plastid evolution at the within-species level.

Almost all of the M-K tests were non-significant, indicating not only a lack of support for positive selection acting on the genes with an elevated dN/dS ratio, but also

no support for purifying selection acting on the genes with a low dN/dS ratio. This lack of significance is probably due to the low power of the M-K tests we used, as a result of two factors. First, for several of the plastid genes/concatenations there are low levels of polymorphism (Table 2), which restricts the power of the M-K test. Second, one of the four *C. americanum* populations sequenced for this study (VA) has a plastid genome that is highly divergent from the other three. Therefore, most of the polymorphisms observed in *C. americanum* are likely fixed within populations and old enough that they have been subject to significant selection. The M-K test compares neutrally arising variation (within species polymorphism) to fixed differences after selection has acted (between species divergence) (McDonald & Kreitman, 1991). If selection has acted on within species polymorphism, this will reduce the power of the M-K test to detect a significant difference between the polymorphism and divergence data. Therefore, when there is substantial intraspecific divergence, such as in *C. americanum*, the nature of the polymorphism data makes it difficult to definitively distinguish between the contributions of positive selection and relaxed purifying selection to the increased nucleotide substitution rates and dN/dS ratios observed in some plastid genes. Future studies using increased sampling could allow for more sensitive tests for positive selection. For example, the recent sequencing of dozens of chloroplast genomes within the Campanulaceae *sensu lato* (Knox, 2014) raises the future possibility of using phylogenetic tests to look for site-specific positive selection.

Accelerated plastid evolution may be an important contributor to the development of reproductive isolation and subsequent speciation. The nuclear and plastid genome are likely to be co-evolved as they must interact closely to carry out essential functions, such

as photosynthesis (Rand *et al.*, 2004). Increased nucleotide substitution rates, altered selective regimes, and increased structural variation have the potential to lead to rapid local co-evolution of these genomes, leading to an increased likelihood for cytonuclear genetic incompatibilities when crossing between populations. Cytonuclear incompatibilities are proposed to be among the first genetic incompatibilities to arise (Levin, 2003; Fishman & Willis, 2006) and are increasingly thought to play an important role in the speciation process as they have been implicated in contributing to reproductive isolation in plants, yeast, and animals (Ellison & Burton, 2008; Sambatti *et al.*, 2008; Chou & Leu, 2010; Leppala & Savolainen, 2011).

Strong reproductive isolation is found between populations of *C. americanum,* and the asymmetrical pattern of this breakdown along with observations of chlorosis and variegation (Fig. 4) suggest cytonuclear incompatibilities contribute to this isolation (Galloway & Etterson, 2005; Etterson *et al.*, 2007). Positive and or relaxed purifying selection on *ycf1*, *ycf2*, *clpP* and the small subunit ribosomal genes, as well as the general instability of the plastid genome, may contribute to cytonuclear incompatibility and reproductive isolation in *C. americanum*. The small subunit ribosomal genes in particular are interesting candidates for intraspecific cytonuclear incompatibilities due to their highly elevated level of pN and the fact that nuclear-encoded subunits of organelle ribosomes have been found to exhibit evidence of compensatory substitutions in response to rapid evolution of organelle genomes (Barreto & Burton, 2013; Sloan *et al.*, 2014b).

Similar to *C. americanum*, other independent lineages of angiosperms have increases in nucleotide substitution rate and elevated dN/dS for a subset of plastid genes (Jansen *et al.*, 2007; Jansen & Ruhlman, 2012). Could accelerated plastid genome

evolution also contribute to cytonuclear incompatibility and reproductive isolation in these lineages? Cytonuclear incompatibilities are known from interspecific crosses in *Oenothera* (reviewed in Greiner *et al.*, 2011) and in *Pelargonium* (Metzlaf *et al.*, 1982; Weihe *et al.*, 2009), both of which exhibit similar patterns of accelerated plastid evolution. However, cytonuclear incompatibilities and reproductive isolation are rarely examined at the intraspecific level. Further work in examining the relationship between accelerated plastid evolution, cytonuclear incompatibility, and reproductive isolation at the intraspecific level would allow for a more general conclusion as to whether accelerated evolution and positive selection on plastid genes could help drive the early stages of the speciation process.

Conclusions

We found increased nucleotide substitution rates when examining intraspecific polymorphism in the plastid genome of *C. americanum*. In addition, there was a significant correlation between the dN/dS and pN/pS ratios across plastid genes. These results provide evidence that mechanisms leading to increased nucleotide substitution rates in the plastid genome are continuing to operate at recent evolutionary timescales and may, therefore, be contributing to the early stages of the speciation process through the development of intraspecific cytonuclear incompatibilities and reproductive isolation.

Availability of Supporting Data

The data sets supporting the results of this article are available in the Dryad repository, [doi:10.5061/dryad.d143r, http://datadryad.org/review?wfID=36174&token=aaae9cb7-2577-4c7f-817f-0910a55105a1].

Additional Files

Additional file 3.txt: Perl script used for extracting intraspecific SNPs.

Additional file 4.txt: Perl script used for extracting interspecific SNPs. (Additional files 3 and 4 not included in dissertation)

Acknowledgements

We thank Doug Taylor for his assistance in planning the project, John Chuckalovcak for his 454 sequencing efforts, and Martin Wu for his support in carrying out the sequencing. We also thank Janis Antonovics and three anonymous reviewers for constructive feedback on earlier drafts of this manuscript. This work was supported by the National Science Foundation (DEB-0922214) and the ARCS Foundation. This article was published in part thanks to funds provided by the University of Virginia Library Open Access Fund.

References

Barnard-Kubow, K.B., Sloan, D.B. & Galloway, L.F. 2014. Data from: Correlation between sequence divergence and polymorphism reveals similar evolutionary mechanisms acting across multiple timescales in a rapidly evolving plastid genome. *Dryad Digital Repository* **doi:10.5061/dryad.d143r**.

Barreto, F.S. & Burton, R.S. 2013. Evidence for compensatory evolution of ribosomal proteins in response to rapid divergence of mitochondrial rRNA. *Molecular Biology and Evolution* **30:** 310-314.

Chou, J.Y. & Leu, J.Y. 2010. Speciation through cytonuclear incompatibility: Insights from yeast and implications for higher eukaryotes. *Bioessays* **32:** 401-411.

Corriveau, J.L. & Coleman, A.W. 1988. Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperm species. *American Journal of Botany* **75:** 1443-1458.

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

Darriba, D., Taboada, G.L., Doallo, R. & Posada, D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9:** 772-772.

Delannoy, E., Fujii, S., des Francs-Small, C.C., Brundrett, M. & Small, I. 2011. Rampant gene loss in the underground orchid *Rhizanthella gardneri* highlights evolutionary constraints on plastid genomes. *Molecular Biology and Evolution* **28:** 2077-2086.

Drescher, A., Ruf, S., Calsa, T., Carrer, H. & Bock, R. 2000. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. *Plant Journal* **22:** 97-104.

Drouin, G., Daoud, H. & Xia, J. 2008. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Molecular Phylogenetics and Evolution* **49:** 827-831.

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

Ellison, C.K. & Burton, R.S. 2008. Interpopulation hybrid breakdown maps to the mitochondrial genome. *Evolution* **62:** 631-638.

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.

Etterson, J.R., Keller, S.R. & Galloway, L.F. 2007. Epistatic and cytonuclear interactions govern outbreeding depression in the autotetraploid *Campanulastrum americanum*. *Evolution* **61:** 2671-2683.

Fishman, L. & Willis, J.H. 2006. A cytonuclear incompatibility causes anther sterility in *Mimulus* hybrids. *Evolution* **60:** 1372-1381.

Galloway, L.F., Etterson, J.R. & Hamrick, J.L. 2003. Outcrossing rate and inbreeding depression in the herbaceous autotetraploid, *Campanula americana*. *Heredity* **90:** 308-315.

Galloway, L.F. & Etterson, J.R. 2005. Population differentiation and hybrid success in *Campanula americana*: geography and genome size. *Journal of Evolutionary Biology* **18:** 81-89.

Gordon, D., Abajian, C. & Green, P. 1998. Consed: A graphical tool for sequence finishing. *Genome Research* **8:** 195-202.

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

Greiner, S., Rauwolf, U., Meurer, J. & Herrmann, R.G. 2011. The role of plastids in plant speciation. *Molecular Ecology* **20:** 671-691.

Guindon, S. & Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52:** 696-704.

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. *Proceedings of the National Academy of Sciences* **105:** 18424-18429.

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. *Journal of Molecular Evolution* **66:** 350-361.

Hudson, R.R., Kreitman, M. & Aguade, M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116:** 153-159.

Jansen, R. & Ruhlman, T. 2012. Plastid genomes of seed plants. In: *Genomics of Chloroplasts and Mitochondria*, Vol. 35 (R. Bock & V. Knoop, eds), pp. 103-126. Springer Netherlands.

Jansen, R.K., Raubeson, L.A., Boore, J.L., DePamphilis, C.W., Chumley, T.W., Haberle, R.C., Wyman, S.K., Alverson, A.J., Peery, R., Herman, S.J., Fourcade, H.M., Kuehl, J.V., McNeal, J.R., Leebens-Mack, J. & Cui, L. 2005. Methods for obtaining and analyzing whole chloroplast genome sequences. *Molecular Evolution: Producing the Biochemical Data, Part B* **395:** 348-384.

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., Depamphilis, C.W., Leebens-Mack, J., Muller, 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. *Proceedings of the National Academy of Sciences* **104:** 19369-19374.

Khakhlova, O. & Bock, R. 2006. Elimination of deleterious mutations in plastid genomes by gene conversion. *The Plant Journal* **46:** 85-94.

Kikuchi, S., Bedard, J., Hirano, M., Hirabayashi, Y., Oishi, M., Imai, M., Takase, M., Ide, T. & Nakai, M. 2013. Uncovering the protein translocon at the chloroplast inner envelope membrane. *Science* **339:** 571-574.

Knox, E.B. 2014. The dynamic history of plastid genomes in the Campanulaceae sensu lato is unique among angiosperms. *Proceedings of the National Academy of Sciences* **111:** 11097-11102.

Leppala, J. & Savolainen, O. 2011. Nuclear-cytoplasmic interactions reduce male fertility in hybrids of *Arabidopsis lyrata* subspecies. *Evolution* **65:** 2959-2972.

Levin, D.A. 2003. The cytoplasmic factor in plant speciation. *Systematic Botany* **28:** 5-11.

McDonald, J.H. & Kreitman, M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351:** 652-654.

Metzlaf, M., Pohlheim, F., Börner, T. & Hagemann, R. 1982. Hybrid variegation in the genus *Pelargonium*. *Current genetics* **5:** 245-249.

Millen, R.S., Olmstead, R.G., Adams, K.L., Palmer, J.D., Lao, N.T., Heggie, L., Kavanagh, T.A., Hibberd, J.M., Giray, 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.

Palmer, J.D. 1986. Isolation and structural-analysis of chloroplast DNA. *Methods in Enzymology* **118:** 167-186.

Rand, D.M., Haney, R.A. & Fry, A.J. 2004. Cytonuclear coevolution: the genomics of cooperation. *Trends in Ecology & Evolution* **19:** 645-653.

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.

Sambatti, J.B.M., Ortiz-Barrientos, D., Baack, E.J. & Rieseberg, L.H. 2008. Ecological selection maintains cytonuclear incompatibilities in hybridizing sunflowers. *Ecology Letters* **11:** 1082-1091.

Sloan, D.B., Alverson, A.J., Wu, M., Palmer, J.D. & Taylor, D.R. 2012. Recent acceleration of plastid sequence and structural evolution coincides with extreme mitochondrial divergence in the angiosperm genus *Silene*. *Genome Biology and Evolution* **4:** 294-306.

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.

Sloan, D.B., Triant, D.A., Wu, M. & Taylor, D.R. 2014b. Cytonuclear interactions and relaxed selection accelerate sequence evolution in organelle ribosomes. *Molecular Biology and Evolution* **31:** 673-682.

Weihe, A., Apitz, J., Pohlheim, F., Salinas-Hartwig, A. & Borner, T. 2009. Biparental inheritance of plastidial and mitochondrial DNA and hybrid variegation in *Pelargonium*. *Molecular Genetics and Genomics* **282:** 587-593.

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

Wyman, S.K., Jansen, R.K. & Boore, J.L. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **20:** 3252-3255.

Yang, Z.H. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution* **15:** 568-573.

Yang, Z.H. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24:** 1586-1591.

Zhang, Q., Liu, Y. & Sodmergen 2003. Examination of the cytoplasmic DNA in male reproductive cells to determine the potential for cytoplasmic inheritance in 295 angiosperm species. *Plant and Cell Physiology* **44:** 941-951.

Zurawski, G. & Clegg, M.T. 1987. Evolution of higher-plant chloroplast DNA-encoded genes: implications for structure-function and phylogenetic studies. *Annual Review of Plant Physiology and Plant Molecular Biology* **38:** 391-418.

Table 1: Location information for the populations used in this study

Population	Latitude	Longitude
VA (Virginia, USA)	37.35495	-80.55415
AL (Alabama, USA)	34.65048	-86.51643
MN (Minnesota, USA)	44.81650	-93.30758
OH (Ohio, USA)	41.11472	-81.51806

Table 2: Sequence divergence and polymorphism values for *Campanulastrum americanum* protein coding plastid genes.

Gene	dN/dS	pN/pS	# Polymorphisms	Gene Length
Photosynthesis				
atp	0.131	0.031	11	4929
ndh	0.248	0.324	17	10263
pet	0.137	0.649	3	2289
psa	0.031	0.000	2	4929
psb	0.079	0.173	6	6360
rbcl	0.102	0.605	3	1425
ycf3	0.093	NA	1	504
ycf4	0.605	0.000	1	552
Transcription and Translation				
rpl	0.326	0.204	14	2757
rpo	0.354	0.428	33	10230
rps	0.915	2.135	36	4788
Other				
ccsA	0.509	0.290	2	918
cemA	0.578	NA	1	684
clpP	5.412*[a]	0.000	1	555
matK	0.406	0.257	6	1491
ycf1	1.293	1.144	11	3042
ycf2	1.086	1.528	26	7137
All Genes	0.430[b]	0.518		

* Indicates significance before Bonferroni correction ($p < 0.05$). a: dN/dS ratio for *clpP* was estimated by calculating dN/dS as if there had been one synonymous SNP. b: mean dN/dS ratio was calculated without including *clpP*. NA: pN/pS ratio was inestimable due to pS being zero.

Figure 1: Sequence divergence (A) and polymorphism (B) for *Campanulastrum americanum* protein coding plastid genes. Sequence divergence and polymorphism as estimated by the number of substitutions per site in the terminal branch leading to *C. americanum* or within the species, respectively. Black and white bars indicate substitutions at synonymous and non-synonymous sites, respectively.

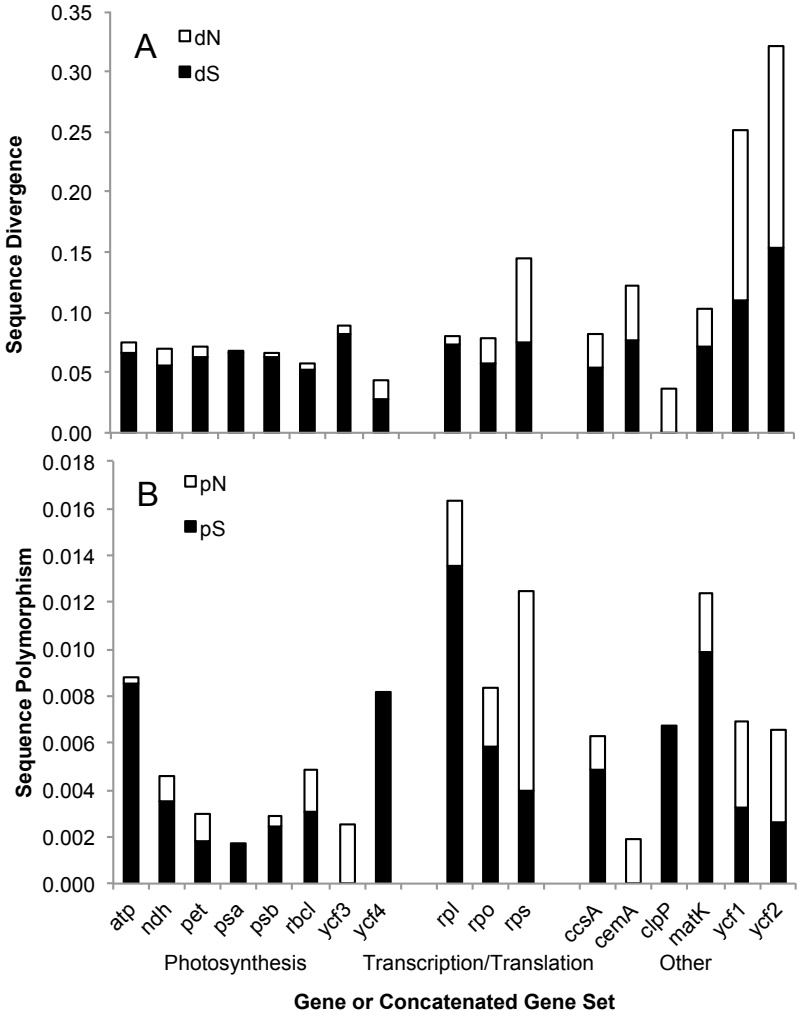


Figure 2: Trees for *clpP* intronic (A) and exonic (B) sequence. Trees with maximum likelihood branch length estimates using *Arabidopsis thaliana* as an outgroup.

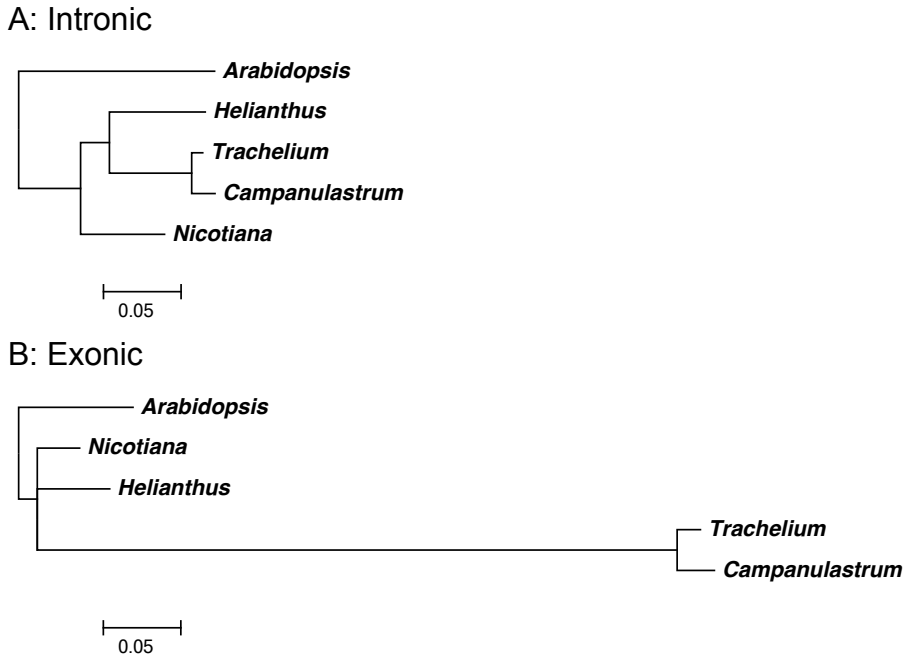


Figure 3: Relationship between the dN/dS and pN/pS ratios for *Campanulastrum americanum* protein coding plastid genes. Genes or concatenations with less than three SNPs, including *clpP*, are not shown. The three points with both an elevated dN/dS and pN/pS ratio are labeled.

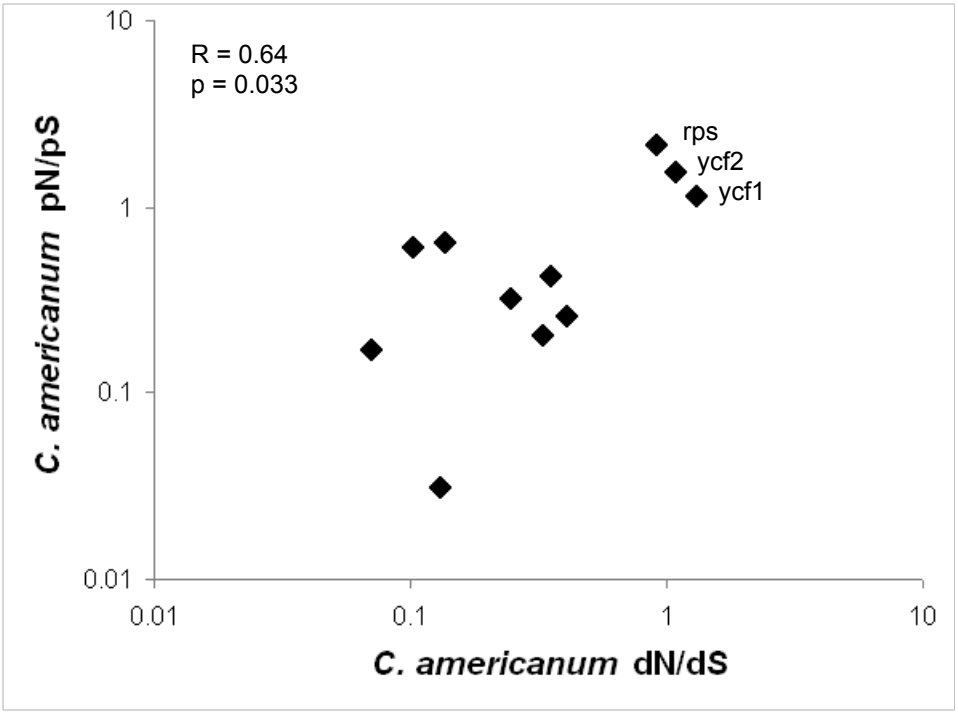


Figure 4: Intraspecific first-generation (F1) hybrid of *Campanulastrum americanum* showing variegation. This variegation is representative of what is found when crossing between populations with high levels of genetic divergence, such as the divergence between Virginia and remaining three populations used in this study.

Additional file 1.pdf: Sequence of *clpP* for *Trachelium caeruleum* with re-annotated intron/exon boundaries. Exons in capital letters and introns in lowercase.

ATGCCAGTTGGTGTTCCGAAAGTCCCACCTATGCCACTTGAACCTGAATCTGAAAA
AAATCAGAAAAAGAAACCTAAACAAGGATTCCTCGAAAAATCCCTCAAATTCCACAA
AAGCCTAGCACGAGAAAAAGAATTCATAGACGAAGAAGAAGTAGATGAAGAAGAAG
TAGATGAAGAAGAAGTAGATGAAGTATATGAAGTATATGAAGAAGAAGTAGAAAAAG
TAGAAACAAAAAAAAAAGGAAAAAGACGGAGAAGAAGAGGAGGAACAAGAGGAGGA
AGAGGAACCTTGGGTTGACTTATAgtgcgacttgtcagatatattggctcatatgggatttccccgttctctccc
cgattgagagatcctctatttcgcccaagaaagattaattgaatcatccaaaatttggagcgtgaagtccaattagatacatttt
ttaggggcattcaaattactattagccgttcgaataatttatggttggcttggttggacgaataaattgaagtatccaggctccgt
ttaaaaaaaccaattggtaatctatctaccattactccttatatcataaacggattcctagttagaaagaaatcttatttgtaaata
gaaatgatctgaaactgttctgttaatcaatcgagtaatatgcatgaagacctggaatatttgccgaaatgcctgaattgaga
aaaaagaagtggtaatgggagtatttgttctatatgtgcaaatcaaaagcgggtggatctttacccggagtagagtataaac
ctaaaaagattcagattaacgaggtccatttaggaacaagcaaatgacatcgtgatttgaattggatctcgatgaaagacta
tatcaatgaaaagtgaattcgataagtttcattaattctttactcgtctttattgaaaatcgaatcaaaatgagaagtccgaaag
agcattctatgaaatccgaaaggggattggaatctatacattgatttttttcgcaaaattttggaaccgtatgcgtcaaaaggc
gcctgtacggttcctaaggaatagaatttgaccctaatCGAACGACTTTATCGAGACAGAGCACTTTTT
TTATTCAAAGAGCTGGATAAGGAGCTCGCGAATACACTTGTGGGTCTTATGGTATTT
CTCAATATAGAGGATAATACCAAAGAACAATTTTTATTTATCAACTCTCCTGGTGGAT
CACTAGTGTATGGAATATCTGTGCATGATGTTAGCCGACTGGTGCGACCAGATGTA
CATACACTAGGCATGGGAGTAGCCGCTTCAATGGCCGCTTTCATCCTGTCCGGAG
GAGCACAAACCAAACGTCTAGCATTCCCTCACGTTtggcgccaatgaggttttatttcagagaaaaa
agaagactatgccttcgccatatgaaatatgaatagtaagtaatataagtaataatagcatggcactttgaattcgatatatga
aagtttttgattgttttttcaaagcattagattatgtatcgagagagtagtatgagataaaaggtatttatgatttttcttatctatcgg
gagtccagttcagcgtcacaaacttttttgctttcccaccggagatctagtaacaatatttatgttatgaacgagtgaaaaaaa
aaaaaaaaaaaaaaagattctttttttccttagtttatttaatcaaataaaaaagcaactttgggattgcttaatcatagacaaaaa
aagaaatccatatataaagcaacggagccatcatagtagttttgaactcccacgaaaggaagggtggtaatttgatcattta
ctgatcgaggtctaatagatcctatttttatctttctttgatagagggtaaggatcaatttgattgtagagccgtatgcaatgcaca
aaagatgcctgtacggttgttcaattctatctttttttcttcttattcttttatctttctttcttttatcttccctttatcaggcgaactagaaga
acctttttattatatcatcAGGGTAATGATTCATCAACCAAAATGTGTTTTTGCAAAGAATCGCAT
CCCGATTGACGTAGGCCTGGACGGAGAAGAAGTGACAAAATTACGTAACTACGTCA
TAGGACATTATGCACAAAGATCGCGCAGGTCTATAGCGATGGTAGTCGCCGACCTG
AAAAGACATACTTATATGACACCAACAGAGGCCCGAACTTATGGAATTGTTGATTCT
ATAGCGGCTGACTGGGAGGTCTACTAAATCCATGATTTGA

Additional file 2.csv: Table S1: Designation of which individual genes were included in concatenations. Also includes list of genes analyzed individually. NA: genes excluded from analysis, see text for details.

	Gene Concatenations								
Gene	atp	ndh	pet	psa	psb	rpl	rps	rpo	Individual
atpA	x								
atpB	x								
atpE	x								
atpF	x								
atpH	x								
atpI	x								
ndhA		x							
ndhB		x							
ndhC		x							
ndhD		x							
ndhE		x							
ndhF		x							
ndhG		x							
ndhH		x							
ndhI		x							
ndhJ		x							
ndhK		x							
petA			x						
petB			x						
petD			x						
petG			x						
petL			x						
petN			NA						
psaA				X					
psaB				X					
psaC				X					
psaI				X					
psaJ				X					
psbA					x				
psbB					x				
psbC					x				
psbD					x				
psbE					x				
psbF					x				
psbH					x				
psbI					x				

psbJ	x				
psbK	x				
psbL	x				
psbM	x				
psbN	x				
psbT	NA				
psbZ	x				
rpl14		x			
rpl16		x			
rpl2		x			
rpl20		x			
rpl22		x			
rpl32		x			
rpl33		x			
rpl36		x			
rps11			x		
rps12			x		
rps14			x		
rps15			x		
rps16			x		
rps18			x		
rps19			x		
rps2			x		
rps3			x		
rps4			x		
rps7			x		
rps8			x		
rpoA				x	
rpoB				x	
rpoC1				x	
rpoC2				x	
ccsA					x
cemA					x
clpP					x
matK					x
rbcl					x
ycf1					x
ycff2					x
ycf3					x
ycf4					x

Additional file 5.csv: Table S2: dN/dS ratios for all four species used in the divergence analyses. dN/dS ratios from the codeml analyses for the terminal branches leading to each species as well as for the internal branch leading to the common ancestor of *T. caeruleum* and *C. americanum*.

			dN/dS		
Gene	N. tabacum*	H. annuus	T. caeruleum	C. americanum	internal branch
Photosynthesis					
atp	0.073	0.078	0.099	0.131	0.260
ndh	0.136	0.144	0.243	0.248	0.223
pet	0.042	0.082	0.135	0.137	0.064
psa	0.019	0.040	0.103	0.031	0.043
psb	0.034	0.029	0.086	0.079	0.072
rbcl	0.115	0.154	0.056	0.102	0.045
ycf3	0.057	0.020	0.000	0.093	0.111
ycf4	0.082	0.258	2.000	0.605	0.219
Transcription and Translation					
rpl	0.151	0.172	0.212	0.326	0.470
rpo	0.133	0.339	0.318	0.354	0.253
rps	0.094	0.095	0.462	0.915	0.719
Other					
ccsA	0.201	0.182	0.354	0.509	0.355
cemA	0.243	0.247	0.552	0.578	0.697
clpP	0.061	0.400	0.927	5.412a	1.230
matK	0.324	0.334	0.701	0.406	0.427
ycf1	0.285	0.500	0.753	1.293	0.550
ycf2	0.577	0.730	0.688	1.086	1.869

*: *N. tabacum* is the outgroup in the analysis, and therefore the reported values are a combination of the terminal branch and the divergence on the internal branch leading to the Asterales.

a: dN/dS ratio for *clpP* was estimated by calculating dN/dS as if there had been one synonymous SNP.