Rethinking phylogenetics using Caryophyllales (angiosperms), *matK* gene and *trnK* intron as experimental platform

Sunny Sheliese Crawley

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Biological Sciences

Khidir W. Hilu
Eric P. Beers
Carla V. Finkielstein
Jill C. Sible

December 2, 2011
Blacksburg, Virginia

Keywords: (phylogeny, missing data, caryophyllids, *trnK* intron, *matK*, RNA editing, gnetophytes)

Rethinking phylogenetics using Caryophyllales (angiosperms), *matK* gene and *trnK* intron as experimental platform

Sunny Sheliese Crawley

ABSTRACT

The recent call to reconstruct a detailed picture of the tree of life for all organisms has forever changed the field of molecular phylogenetics. Sequencing technology has improved to the point that scientists can now routinely sequence complete plastid/mitochondrial genomes and thus, vast amounts of data can be used to reconstruct phylogenies. These data are accumulating in DNA sequence repositories, such as GenBank, where everyone can benefit from the vast growth of information. The trend of generating genomic-region rich datasets has far outpaced the expasion of datasets by sampling a broader array of taxa. We show here that expanding a dataset both by increasing genomic regions and species sampled using GenBank data, despite the inherent missing DNA that comes with GenBank data, can provide a robust phylogeny for the plant order Caryophyllales (angiosperms). We also investigate the utility of *trnK* intron in phylogeny reconstruction at relativley deep evolutionary history (the caryophyllid order) by comparing it with rapidly evolving *matK*. We show that *trnK* intron is comparable to *matK* in terms of the proportion of variable sites, parsimony informative sites, the distribution of those sites among rate classes, and phylogenetic informativness across the history of the order. This is especailly useful since *trnK* intron is often sequenced concurrently with *matK* which saves on time and resources by increasing the phylogenetic utility of a single genomic region (rapidly evolving *matK*/*trnK*). Finally, we show that the inclusion of RNA edited sites in datasets for phylogeny reconstruction did not appear to impact resolution or support in the Gnetales indicating that edited sites in such low proportions do not need to be a consideration when building datasets. We also propose an alternate start codon for *matK* in *Ephedra*

based on the presense of a 38 base pair indel in several species that otherwise result in pre-mature stop codons, and present 20 RNA edited sites in two Zamiaceae and three Pinaceae species.

Acknowledgements

I would like to give my most sincere thanks to Dr. Hilu for his mentorship during the course of my graduate studies. He has always been supportive and encouraging and has provided me with so many wonderful opportunities both in the lab and through numerous meetings throughout the years. Thank you for encouraging me to take a semester off to teach Plant Taxonomy and helping that endeavor to be a success. I have enjoy debating our work together and look forward to continuing collaborations. I also thank each member of my committee, Dr. Jill Sible for always listening and providing encouraging feedback along with sage scientific advice, Dr. Carla Finkielstein for all her help with the expression work and always encouraging me to do my best, and Dr. Eric Beers for teaching me plant biology and asking critical questions to help me keep on track.

I would also like to thank everyone that has come and gone through the Hilu lab during my tenure. Thanks to Dr. Michelle Barthet-Parker for laying the groundwork of a wonderful and exciting project and setting a great example of how to succeed in graduate school. Thanks to Dr. Sheena Friend-Elliot for taking me under your wing when I first arrived, for getting through all of our classes together both scientific and pedagogical, for troubleshooting the research, for being my traveling companion to the various meetings, for being a wonderful friend and so much more. To Stephanie Voshell for your undying enthusiasm for research and doing the last minute leg-work in my place. To Atia for always having a smile on your face and a joke to share (chicken cookie?). Thanks to Chelsea Black for being a great friend and our PCR miracles. Thank you to the many undergraduate students that have helped with the research through the years, specifically Shelli Newman my "sister" in science, Adrianna Feriolli and the countless hours of PCR and tree building, Dipan Oza my bioinformatics buddy, Emily Steele and the mega-PCR sessions, Keenan Moukarzel for figuring out new programs with me, and Brittany for taking on the yeast expression project.

Thanks to all the other Derring Hall 5th floor graduate students over the years, Gana, Chen, Catherine, Justin, Nassiba, Shakur, Sarah, and everyone else for your friendship and support. Thanks to all those that have helped me to be a better teacher, specifically Shelli Fowler, Karen DePauw, Art Buikema, Mary Schaeffer, Catherine Sarmadi, Sharon Sible, Chiquita Thomas, Carla Tyler, and Jason Yolitz for teaching me about pedagogy and providing feedback on my teaching through the years.

Thank you to my parents Thomas and Laurel Drysdale and my wonderful in-laws Patti and Joe Crawley for their example, support, and encouragement. Thanks to my husband Adam Crawley for enduring the countless hours spent on school work instead of playing and our son Marcus for tolerating my tired days.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Current approaches to phylogenetic reconstruction

In recent years that has been a call to assemble a tree of life that will reconstruct the evolutionary origins of all living organisms (http://www.phylo.org/atol/). This initiative has changed the landscape of phylogeny reconstruction as large-scale analyses are being conducted to get a more detailed picture of the tree of life. Additionally, DNA sequencing technology has improved and the ease with which large amounts of DNA data are being sequenced has greatly increased. This has lead to phylogenetic studies that are being based upon increasingly information rich datasets (Qiu et al., 2006; Moore et al., 2007; Jansen et al., 2007; Turmel et al., 2009; Wang et al., 2009; Moore et al., 2010; Jansen et al., 2011; Soltis et al., 2011). In fact, many studies are now based on complete or nearly complete chloroplast genome data (Leebens-Mack et al., 2005; Qiu et al., 2006; Jansen et al., 2007; Moore et al., 2007; Turmel et al., 2009; Moore et al., 2010).

While studies using greater amounts of sequence data per species are increasing, the number of species used in these studies is constrained by the cost and time required for such large-scale sequencing. However, it has been documented that taxon sampling can greatly impact the nature of the tree and its robustness, and therefore increase in sampling is essential for an accurate picture of the tree of life of any particular group (Hillis, 1996; Graybeal, 1998; Zwickl and Hillis, 2002; Soltis et al., 2011). In addition, the call for understanding the details of the tree of life (terminal branches) requires substantial increase in species representation and thus the current mega-sequence approach is not efficient.

Objectives:

1. Explore the impact of missing data on phylogenetic reconstruction of the Caryophyllales

2. Compare the phylogenetic utility of *trnK* intron to *matK* and contrast them both with phylogenies generated using other genomic regions for the Caryophyllales
3. Determine to what extent the *matK* gene is edited as an RNA transcript within the Gnetales

Impact of missing data on phylogeny reconstruction

DNA sequence repositories, such as GenBank, offer a valuable source of data for phylogenetic studies. Yet these data are often underutilized due to the presence of partial sequences and data lacking for some proportion of the taxa of interest. As of 2008, there were just under 100 million sequences deposited in GenBank (http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html) offering a rich source of information available to systematists for use in phylogeny reconstruction. GenBank data can be used to increase the number of genomic regions being sampled for a given number of species and/or to increase the number of species included in a study for a given number of genomic regions. Since researchers have disparate objectives when generating sequence data, the data deposited in GenBank do not always span the exact same proportion of any given genomic region across phylogenetic groups. Nor do all species have sequences available for the same genomic regions. Thus, using GenBank sequence data to expand genomic/taxonomic sampling for phylogeny reconstruction can lead to datasets punctuated to different degrees with missing data. While there have been a handful of simulation studies addressing the role of missing data in phylogenetic reconstruction (Wiens, 1998, 2003a, 2003b, 2005, 2006), there is a noticeable lack of empirical studies that have explored this topic (Agnarsson and May-Collado, 2008; Burleigh et al., 2009).

I am using the flowering plants (angiosperm) order Caryophyllales to explore the impact of missing data on phylogenetic reconstruction. I have mined the GenBank for sequence data to expand a dataset of caryophllid species ~13 folds, and to increase the genomic data sampled ~3 fold, without regard to the amount of missing data incorporated as a result, to evaluate its impact on phylogeny reconstruction. I have generated a dataset of

51 caryophyllid taxa for *matK* and *trnK* intron sequences that is nearly free of missing data (0.3%). This dataset has been expanded in two ways, by increasing taxon sampling to 652 species for *matK*/*trnK* intron (38.4% missing data) or by increasing the number of genomic regions used to include *rbcL, atpB,* and *ndhF* (46.3% missing data). This latter dataset has then been systematically reduced by deleting species completely lacking in a given number of genomic regions to generate three new datasets. For the first, the dataset must include at least a partial sequence for 3 of the 5 regions to remain in the dataset, for the second, 4 of the 5 regions must be present, and for the third all 5 regions must have some portion of each genomic region to be retained in the dataset.

Genomic regions

Genomic data used in plant phylogeny reconstruction is dominated by sequences within the plastid genome (Savolainen et al., 2000; Soltis et al., 2003; Soltis et al., 2004; Jansen et al., 2007; Moore et al., 2007; Soltis et al., 2011). Each of the five regions used here (*matK*, *trnK* intron, *rbcL*, *atpB*, and *ndhF*) are encoded in the chloroplast genome and can be grouped into three classifications based on rate of evolution. The most rapidly evolving are *matK* and *trnK* intron, followed by *ndhF* with a moderate rate of evolution, with *rbcL* and *atpB* in the final category as slowly evolving genes.

The slowly evolving genes *rbcL* and *atpB* have long been used in plant phylogeny reconstruction, especially at deeper historic levels (Hasebe et al., 1994; Conti et al., 1996; Fay et al., 1997; Källersjö et al., 1998; Lledo et al., 1998; Savolainen et al., 2000; Soltis et al., 2003; Anderson et al., 2005; Davis et al., 2005). The *rbcL* gene codes for the large subunit of RuBisCo, an essential enzyme in photosynthesis, while the *atpB* gene encodes the beta subunit of the plastid ATPase (Hasebe et al., 1994; Woessner et al., 1986). The gene *ndhF* has been shown to evolve at a moderate rate, more rapidly than *rbcL* and *atpB*, but not as rapidly as *matK* and *trnK* intron (Kim and Jansen, 1995; Alverson et al., 1999; Applequist and Wallace, 2001). The *ndhF* gene encodes a subunit of the plastid NADH dehydrogenase complex (Kim and Jansen, 1995).

The rapidly evolving *matK* gene has become one of the most widely used genomic regions in plant phylogeny reconstruction since it was promoted by work done in our lab (Hilu and Liang, 1997; Hilu and Alice, 1999; Hilu et al., 1999; Hilu et al., 2003; Barthet and Hilu, 2007; Hilu et al., 2008). This gene is nested within the *trnK* (tRNA Lys ^{UUU}) intron of the large single copy region of the plastid genome (Sugita et al., 1985). It codes for a putative group II intron maturase, an essential splicing factor responsible for splicing the intron in which it resides and possibly several other introns within the plastid genome (Vogel et al., 1997; Vogel et al., 1999; Barthet and Hilu, 2008). The gene has an unusual mode and tempo of evolution, especially when compared with other genes used in molecular phylogenetics. Despite being a protein coding gene *matK* exhibits 3 times higher rate of nucleotide substitution and 6 times higher rate of amino acid substitution than the chloroplast gene *rbcL* (Johnson and Soltis, 1995; Müller et al., 2006). It has been shown that sequence information from *matK* is at least as effective as using 3-11 genes combined, and in one instance provided more phylogenetic information than all the genes found in the chloroplast inverted repeat region (IR) plus 9 chloroplast and 2 nuclear genes. Its resolution has been demonstrated both at shallow and deep historic levels (Graham and Olmstead, 2000; Zanis et al., 2002; Hilu et al., 2003; Brockington et al., 2009).

In most cases, when amplifying the full *matK* gene, universal primers that sit within the *trnK* exons are used, thus generating sequences of the *trnK* intron as a by-product of the sequencing process. The *trnK* intron however, has been used mainly in studies addressing relationships at the familial level (Hu et al., 2000; Young and dePamphilis, 2000; Edwards and Gadek, 2001; Lavin et al., 2001; Wilson, 2004; Ronsted et al., 2005). It has been shown recently that the *trnK* intron evolves at the same rate as *matK* (Hilu et al., 2008) and therefore has the potential to be very useful in phylogeny reconstruction above the family level. However, most researchers do not utilize the information from this intron and thus invaluable data are lost. To evaluate the utility of this intron in phylogeny reconstruction at the ordinal level I have constructed a dataset of caryophyllid species with *matK* and *trnK* intron sequences. The regions are each analyzed as

partitioned datasets and then in combination, under methods of Maximum Parsimony (MP), Maximum Likelihood (ML), and Bayesian Inference (BI).

The plant order Caryophyllales

The Caryophyllales are one of the major lineages of angiosperms and contain a wide variety of life forms including carnivorous, desert plants, and plants with different photosynthetic pathways (C3, C4, CAM). This order was traditionally recognized based on several synapomorphies, including curved embryo, P3 type of sieve element plastid, perisperm, and free central placentation (Bittrich, 1993; Behnke, 1994). Cronquist recognized 12 families in the order while Thorne listed 15 members to the Caryophyllales (Cronquist and Thorne, 1994). Over the past 15 years many more families have been added to the group, which is currently comprised of 11,510 species, grouped into 811 genera, and 34 families (Stevens, 2008; APG III, 2009).

The order has been divided into two major clades, the core and non-core Caryophyllales along with several well-defined subclades (Albert et al., 1992; Nandi et al., 1998; Brockington et al., 2009). Several families have proven difficult to place within the order and some families do not appear as natural biological units (polyphyletic) while other clades are very well defined (Savolainen et al., 2000; Cuénoud et al., 2002; Hilu et al., 2003; Soltis et al., 2003; Brockington et al., 2009). The existence of a robust phylogeny for the order (Brockington et al., 2009), yet the uncertainty that remains for a few families makes this group an ideal choice for exploring the way we reconstruct phylogenies.

RNA editing of Gnetales matK

Molecular phylogeny reconstruction is most often based on DNA sequence data based on the assumption that genomic DNA is the source of stored historical information (Bowe and dePamphilis, 1996; Kumar and Filipski, 2008). The discovery of RNA editing however, changed the paradigm that DNA sequences can be directly translated into

amino acid sequence due to the differences detected in the cDNA transcript (Bowe and dePamphilis, 1996). This has raised questions about the effect RNA edited sites will have on phylogeny reconstruction, since the patterns of RNA editing might evolve differently in different groups of organisms (Bowe and dePamphilis, 1996). A study looking at mitochondrial *coxII* and *coxIII* genes concluded that there was little difference between phylogenies based on DNA or cDNA sequences, but that the two types of data should not be combined (Bowe and dePamphilis, 1996). More recently, a study of *matK* sequences in fern species conducted phylogenetic analyses on a dataset that considered known edited sites along with unedited DNA sequences (Duffy et al., 2009). We chose to examine the impact of RNA editing of the *matK* gene on phylogeny reconstruction in the Gnetales.

The plant order Gnetales (gymnosperms) consists of 96 species grouped into 3 genera, *Welwitschia* (Welwitchiaceae), *Gnetum* (Gnetaceae), and *Ephedra* (Ephedraceae). This order has historically been very difficult to place phylogenetically due to shared characters with both flowering plants and conifers (Carlquist, 1996; Friedman and Carmichael, 1996; Donoghue and Doyle, 2000). Recent molecular studies have placed them as sister to angiosperms, sister gymnosperms, sister to seed plants, or sister to pines within the gymnosperms (Bowe et al., 2000; Chaw et al., 2000; Donoghue and Doyle, 2000; Magallón and Sanderson, 2002; Burleigh and Mathews, 2004; Braukmann et al., 2009; Mathews, 2009). They have also been shown to possess a higher proportion of non-polar amino acids than other plants with respect to the *matK* gene (Barthet and Hilu, 2008). This resulted in the prediction of a transmembrane domain found solely within Gnetales (Barthet, 2006). These distinct traits of the *matK* gene in the Gnetales provide an excellent platform for further experimentation on the impact of RNA editing on phylogeny reconstruction for this group. We chose to explore the possibility of RNA editing of the *matK* transcript in order to determine if there are other unique characteristics of this gene within the Gnetales.

RNA editing is a mechanism employed by cells to alter an RNA message before translation of a protein product (Tillich et al., 2006). Generally this editing involves a C

to U edit in either the first or second codon position, and has been shown in almost all land plant chloroplasts studied to date (Freyer et al., 1997; Tillich et al., 2006). Specific RNA editing of the *matK* transcript has been shown in Barley (Vogel et al., 1997), maidenhair fern *Adiantum* (Wolf et al., 2003), rice (Inada et al., 2004) and in the gymnosperm cycad species *Cycas taitungensis* (Chen et al., 2011). We have sequenced DNA and cDNA sequences of *matK* for 16 members of the Gnetales and 7 other gymnosperm species to look for evidence of RNA editing and address the impact of such editing on phylogeny reconstruction.

References:

Agnarsson, I. and May-Collado, L. J., 2008: The phylogeny of Cetartiodactyla: The importance of dense taxon sampling, missing data, and the remarkable promise of cytochrome b to provide reliable species-level phylogenies. *Molecular Phylogenetics and Evolution,* 48: 964-985.

Albert, V. A., Williams, S. E., and Chase, M. W., 1992: Carnivorous Plants: Phylogeny and Structural Evolution. *Science,* 257: 1491 - 1495.

Alverson, W. S., Whitlock, B. A., Nyffeler, R., Bayer, C., and Baum, D. A., 1999: Phylogeny of the core Malvales: evidence from *ndhF* sequence data. *American Journal of Botany,* 86: 1474-1486.

Anderson, C. L., Kåre, B., and Friis, E. M., 2005: Dating phylogenetically basal eudicots using *rbcL* sequences and multiple fossil reference points. *American Journal of Botany,* 92: 1737-1748.

APG III, 2009: An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society,* 161: 105-121.

Applequist, W. L. and Wallace, R. S., 2001: Phylogeny of the Portulacaceous Cohort Based on *ndhF* Sequence Data. *Systematic Botany,* 26: 406-419.

Barthet, M. M., 2006: Expression and function of the chloroplast-encoded gene *matK*, Biological Sciences, Virginia Polytechnic Institute and State University, Blacksburg. Pages pp.

Barthet, M. M. and Hilu, K. W., 2007: Expression of *matK*: Functional and Evolutionary Implications. *American Journal of Botany,* 94: 1402-1412.

Barthet, M. M. and Hilu, K. W., 2008: Evaluating evolutionary constraint on the rapidly evolving gene *matK* using protein composition. *J Mol Evol,* 66: 85-97.

Behnke, H.-D., 1994: Sieve-Element Plastids: Their Significance for the Evolution and Systematics of the Order. *In* Behnke, H.-D. and Mabry, T. J. (eds.), *Caryophyllales: Evolution and Systematics*. Berlin, Germany: Springer Verlag, 87 - 121.

Bittrich, V., 1993: Introduction to Centrospermae. *In* Kubitzki, K., Rohwer, J. G., and Bittrich, V. (eds.), *The families and genera of vascular plants, vol. II, Magnoliid, hamamelid, and caryophyllid families*. Berlin, Germany: Springer Verlag, 13 - 19.

Bowe, L. M. and dePamphilis, C. W., 1996: Effects of RNA editing on gene processing on phylogenetic reconstrcution. *Mol Biol Evol,* 13: 1159-1166.

Bowe, L. M., Coat, G., and dePamphilis, C. W., 2000: Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and Gnetales closest relatives are conifers. *Proc Natl Acad Sci U S A,* 97: 4092-4097.

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

Brockington, S. F., Alexandre, R., Ramdial, J., Moore, M. J., Crawley, S., Dhingra, A., Hilu, K., Soltis, D. E., and Soltis, P. S., 2009: Phylogeny of the Caryophyllales Sensu Lato: Revisiting Hypotheses on Pollination Biology and Perianth Differentiation in the Core Caryophyllales. *International Journal of Plant Sciences,* 170: 627-643.

Burleigh, J. G. and Mathews, S., 2004: Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. *American Journal of Botany,* 91: 1599-1613.

Burleigh, J. G., Hilu, K. W., and Soltis, D. E., 2009: Inferring Phylogenies with Incomplete Data Sets: A 5-Gene, 567-Taxon Analysis of Angiosperms. *BMC Evolutionary Biology,* 9: 61.

Carlquist, S., 1996: Wood, bark, and stem anatomy of Gnetales: a summary. *International Journal of Plant Sciences,* 157: S58-S76.

Chaw, S.-M., Parkinson, C. L., Cheng, Y., Vincent, T. M., and Palmer, J. D., 2000: Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. *Proc Natl Acad Sci U S A,* 97: 4086-4091.

Chen, H., Deng, L., Jiang, Y., Lu, P., and Yu, J., 2011: RNA Editing Sites Exist in Protein-Coding Genes in the Chloroplast Genome of Cycas Taitungensis. *Journal of Integrative Plant Biology*: no-no.

Conti, E., Litt, A., and Sytsma, K. J., 1996: Circumscription of Myrtales and their relationships to other rosids: evidence from *rbcL* sequence data. *American Journal of Botany,* 83: 221-233.

Cronquist, A. and Thorne, R. F., 1994: Nomenlatural and Taxonomic History. *In* Behnke, H.-D. and Mabry, T. J. (eds.), *Caryophyllales: Evolution and Systematics*. Berlin, Germany: Springer Verlag, 5 - 25.

Cuénoud, P., Savolainen, V., Chatrou, L. W., Powell, M. P., Grayer, R. J., and Chase, M. W., 2002: Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL, atpB,* and *matK* DNA sequences. *American Journal of Botany,* 89: 132 - 144.

Davis, C. C., Webb, C. O., Wurdack, K. J., Jaramillo, C. A., and Donoghue, M. J., 2005: Explosive Radiation of Malpighiales Supports a Mid-Cretaceous Origin of Modern Tropical Rain Forests. *The American Naturalist,* 165: E36-E65.

Donoghue, M. J. and Doyle, J. A., 2000: Seed plant phylogeny: Demise of the anthophyte hypothesis? *Current Biology,* 10: R106-R109.

Duffy, A. M., Kelchner, S. A., and Wolf, P. G., 2009: Conservation of selection on *matK* follwoing an ancient loss of its flanking intron. *Gene,* 438: 17-25.

Edwards, K. J. and Gadek, P. A., 2001: Evolution and biogeography of *Alectryon* (Sapindaceae). *Molecular Phylogenetics and Evolution,* 20: 14 - 26.

Fay, M. F., Cameron, K. M., Prance, G., T., Lledo, M. D., and Chase, M. W., 1997: Familial relationships of *Rhabdodendron* (*Rhabdodendraceae*): plastid *rbc*L sequences indicate a caryophyllid placement. *Kew Bulletin,* 54: 923 - 932.

Freyer, R., Kiefer-Meyer, M.-C., and Kossel, H., 1997: Occurrence of plastid RNA editing in all major lineages of land plants. *Proc Natl Acad Sci U S A,* 94: 6285-6290.

Friedman, W. E. and Carmichael, J. S., 1996: Double Fertilization in Gnetales: Implications for Understanding Reproductive Diversification among Seed Plants. *International Journal of Plant Sciences,* 157: S77-S94.

Graham, S. W. and Olmstead, R. G., 2000: Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *American Journal of Botany,* 87: 1712-1730.

Graybeal, A., 1998: Is it better to add taxa or characters to a difficult phylogenetic problem? *Systematic Biology,* 47: 9-17.

Hasebe, M., Omori, T., Nakazawa, M., Sano, T., Kato, M., and Iwatsuki, K., 1994: *rbcL* gene sequences provide evidence for the evolutionary lineages of leptosporangiate ferns. *Proc Natl Acad Sci U S A,* 91: 5730-5734.

Hillis, D. M., 1996: Inferring complex phylogenies. *Nature,* 383: 130-131.

Hilu, K. W. and Liang, H., 1997: The *matK* Gene: Sequence Variation and Application in Plant Systematics. *American Journal of Botany,* 84: 830 - 839.

Hilu, K. W. and Alice, L. A., 1999: Evolutionary Implications of *matK* Indels in Poaceae. *American Journal of Botany,* 86: 1735 - 1741.

Hilu, K. W., Alice, L. A., and Liang, H., 1999: Phylogeny of Poaceae inferred from *matK* sequences. *Annals of the Missouri Botanical Garden,* 86: 835-851.

Hilu, K. W., Borsch, T., Müller, K., Soltis, D. E., Soltis, P. S., Savolainen, V., Chase, M. W., Powell, M. P., Alice, L. A., Evans, R., Sauquet, H., Neinhuis, C., Slotta, T. A. B., Jens, G. R., Campbell, C. S., and Chatrou, L. W., 2003: Angiosperm phylogeny based on *matK* sequence information. *American Journal of Botany,* 90: 1758 - 1776.

Hilu, K. W., Black, C., Diouf, D., and Burleigh, J. G., 2008: Phylogenetic signal in *matK* vs. *trnK*: a case study in early diverging eudicots (angiosperms). *Molecular Phylogenetics and Evolution,* 48: 1120-1130.

Hu, J.-M., Lavin, M., Wojciechowski, M. F., and Sanderson, M. J., 2000: Phylogenetic Systematics of the Tribe Millettieae (Leguminosae) Based on Chloroplast

trnK/matK Sequences and its Implications for Evolutionary Patterns in Papilionoideae. *American Journal of Botany,* 87: 418 - 430.

Inada, M., Sasaki, T., Yukawa, M., Tsudzuki, T., and Sugiura, M., 2004: A systematic search for RNA editing sites in pea chloroplasts: an editing event causes diversification from the evolutionarily conserved amino acid sequence. *Plant Cell Physiol,* 45: 1615-1622.

Jansen, R. K., Cai, Z., Raubeson, L. A., Daniell, H., dePamphilis, C. W., Leebens-Mack, J., Müller, K. F., Guisinger-Bellian, M., Haberle, R. C., Hansen, A. K., Chumley, T. W., Lee, S.-B., Peery, R., McNeal, J. R., Kuehl, J. V., and Boore, J. L., 2007: Analysis of 81 Genes from 64 Plastid Genomes Resolves Relationships in Angiosperms and Identifies Genome-Scale Evolutionary Patterns. *PNAS,* 104: 19369-19374.

Jansen, R. K., Saski, C., Lee, S.-B., Hansen, A. K., and Daniell, H., 2011: Complete Plastid Genome Sequences of Three Rosids (*Castanea, Prunus, Theobroma*): Evidence for At Least Two Independent Transfers of *rpl22* to the Nucleus. *Mol Biol Evol,* 28: 835-847.

Johnson, L. A. and Soltis, D. E., 1995: Phylogenetic inference in Saxifragaceae sensu stricto and *Gilia* (Polemoniaceae) using *matK* sequences. *Annals of the Missouri Botanical Gardens,* 82: 149 - 175.

Kim, K.-J. and Jansen, R. K., 1995: *ndhF* sequence evolution and the major clades in the sunflower family. *Proc Natl Acad Sci U S A,* 92: 10379-10383.

Källersjö, M., Farris, J. S., Chase, M. W., Bremer, B., Fay, M. F., Humphries, C. J., Petersen, G., Seberg, O., and Bremer, K., 1998: Simultaneous parsimony jackknife analysis of 2538 *rbc*L DNA sequences reveals support for major clades of green plants, land plants, seed plants, and flowering plants. *Plant Systematics and Evolution,* 213: 259-287.

Kumar, S. and Filipski, A., 2008: Molecular Phylogeny Reconstruction, *Encyclopedia of Life Science (ELS)*. Chichester: John Wiley & Sons, Ltd.

Lavin, M., Pennington, R. T., Klitgaard, B. B., Sprent, J. I., Cavalcante de Lima, H., and Gasson, P. E., 2001: The Dalberdioid Legumes (Fabaceae): Delimitation of a Pantropical Monophyletic Clade. *American Journal of Botany,* 88: 503 - 533.

Leebens-Mack, J., Raubeson, L. A., Cui, L., Kuehl, J. V., Fourcade, M. H., Chumley, T. W., Boore, J. L., Jansen, R. K., and dePamphilis, C. W., 2005: Identifying the Basal Angiosperm Node in Chloroplast Genome Phylogenies: Sampling One's Way Out of the Felsenstein Zone. *Molecular Biology and Evolution,* 22: 1948-1963.

Lledo, M. D., Crespo, M. B., Cameron, K. M., Fay, M. F., and Chase, M. W., 1998: Systematics of Plumbaginaceae Based upon Cladistic Analysis of rbcL Sequence Data. *Systematic Botany,* 23: 21-29.

Magallón, S. and Sanderson, M. J., 2002: Relationships among seed plants inferred from highly conserved genes: sorting conflicting phylogenetic signal among ancient lineages. *American Journal of Botany,* 89: 1991-2006.

Mathews, S., 2009: Phylogenetic relationships among seed plants: persistent questions and the limits of molecular data. *American Journal of Botany,* 96: 228-236.

Moore, M. J., Bell, C. D., Soltis, P. S., and Soltis, D. E., 2007: Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc Natl Acad Sci U S A,* 104: 19363-19368.

Moore, M. J., Soltis, P. S., Bell, C. D., Burleigh, J. G., and Soltis, D. E., 2010: Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *PNAS,* 107: 4623-4628.

Müller, K. F., Borsch, T., and Hilu, K. W., 2006: Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: contrasting *matK*, *trnT-F* and *rbcL* in basal angiosperms. *Molecular Phylogenetics and Evolution,* 41: 99 - 117.

Nandi, O. I., Chase, M. W., and Endress, P. K., 1998: A Combined Cladistic Analysis of Angiopserms Using rbcL and Non-Molecular Data Sets. *Annals of the Missouri Botanical Garden,* 85: 137 - 214.

Qiu, Y.-L., Li, L., Wang, B., Chen, Z., Knopp, V., Groth-Malonek, M., Dombrovska, O., Lee, J., Kent, L., Rest, J., Estabrook, G. F., Hendry, T. A., Taylor, D. W., Testa, C. M., Ambros, M., Crandall-Stotler, B., Duff, R. J., Stech, M., Frey, W., Quandt, D., and Davis, C. C., 2006: The deepest divergences in land plants inferred from phylogenomic evidence. *PNAS,* 103: 15511-15516.

Ronsted, N., Law, S., Thronton, H., Fay, M. F., and Chase, M. W., 2005: Molecular Phylogenetic evidence for monophly of *Fritillaria* and *Lilium* (Liliaceae; Liliales) and the infrageneric classification of Fritillaria. *Molecular Phylogenetics and Evolution,* 35: 509 - 527.

Savolainen, V., Chase, M. W., Hoot, S. B., Morton, C. M., Soltis, D. E., Bayer, C., Fay, M. F., DeBruijn, A. Y., Sullivan, S., and Qiu, Y.-L., 2000: Phylogenetics of Flowering Plants Based on Combined Analysis of Plastid *atpB* and *rbcL* Gene Sequences. *Systematic Biology,* 49: 306 - 362.

Savolainen, V., Fay, M. F., Albach, D. C., Backlund, A., van der Bank, M., Cameron, K. M., Johnson, S. A., Lledo, M. D., Pintaud, J. C., Powell, M., Sheahan, M. C., Soltis, D. E., Soltis, P. S., Weston, P., Whitten, W. M., Wurdack, K. J., and Chase, M. W., 2000: Phylogeny of the eudicots: a nearly complete familial analysis based on *rbcL* gene sequences. *Kew Bulletin,* 55: 257-309.

Soltis, D. E., Senters, A. E., Zanis, M. J., Kim, S., Thompson, J. D., Soltis, P. S., Ronse De Craene, L. P., Endress, P. K., and Farris, J. S., 2003: Gunnerales are sister to other core eudicots: implications for the evolution of pentamery. *American Journal of Botany,* 90: 461-470.

Soltis, D. E., Albert, V. A., Savolainen, V., Hilu, K., Qiu, Y.-L., Chase, M. W., Farris, J. S., and Stefanovic, S., 2004: Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. *TRENDS in Plant Science,* 9: 477-483.

Soltis, D. E., Smith, S. A., Cellinese, N., Wurdack, K. J., Tank, D. C., Brockington, S. F., Refulio-Rodriguez, N. F., Walker, J. B., Moore, M. J., Carlsward, B. S., Bell, C. D., Latvis, M., Crawley, S., Black, C., Diouf, D., Xi, Z., Rushworth, C. A., Gitzendanner, M. A., Sytsma, K. J., Qiu, Y.-L., Hilu, K. W., Davis, C. C., Sanderson, M. J., Beaman, R. S., Olmstead, R. G., Judd, W. S., Donoghue, M. J., and Soltis, P. S., 2011: Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany,* 98: 704-730.

Stevens, P. F., 2008: Angiosperm Phylogeny Website

Sugita, M., Shinozaki, K., and Sugiura, M., 1985: Tobacco Chloroplast $tRNA^{Lys}$(UUU) gene contains a 2.5-kilobase-pair intron: An open reading frame and a conserved boundary sequence in the intron. *Proc Natl Acad Sci U S A,* 82: 3557 - 3561.

Tillich, M., Lehwark, P., Morton, B. R., and Maier, U. G., 2006: The evolution of chloroplast RNA editing. *Mol Biol Evol,* 23: 1912-1921.

Turmel, M., Gagnon, M.-C., O'Kelly, C. J., Otis, C., and Lemieux, C., 2009: The Chloroplast Genomes of the Green Algae *Pyramimonas, Monomastix,* and *Pycnococcus* shed new light on the evolutionary history of Prasinophytes and the origin of the secondary chloroplasts of Euglenids. *Mol Biol Evol,* 26: 632-648.

Vogel, J., Hubschmann, T., Borner, T., and Hess, W. R., 1997: Splicing and intron-internal RNA editing of trnK-matK transcripts in barley plastids: support for MatK as an essential splicing factor. *Journal of Molecular Biology,* 270: 179 - 187.

Vogel, J., Borner, T., and Hess, W. R., 1999: Comparative analysis of splicing of the complete set of chloroplast group II introns in three higher plant mutants. *Nucleic Acids Res,* 27: 3866 - 3874.

Wang, H., Moore, M. J., Soltis, P. S., Bell, C., Brockington, S. F., Alexandre, R., Davis, C. C., Latvis, M., Manchester, S. R., and Soltis, D. E., 2009: Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proc Natl Acad Sci U S A,* 106: 3853-3858.

Wiens, J. J., 1998: Does Adding Characters with Missing Data Increase or Decrease Phylogenetic Accuracy? *Systematic Biology,* 47: 625-640.

Wiens, J. J., 2003a: Incomplete Taxa, Incomplete Characters, and Phylogenetic Accuracy: Is There a Missing Data Problem? *Journal of Vertebrate Paleontology,* 23: 297-310.

Wiens, J. J., 2003b: Missing Data, Incomplete Taxa, and Phylogenetic Accuracy. *Systematic Biology,* 52: 528-538.

Wiens, J. J., 2005: Can Incomplete Taxa Rescue Phylogenetic Analyses from Long-Branch Attraction? *Systematic Biology,* 54: 731-742.

Wiens, J. J., 2006: Missing data and the design of phylogenetic analyses. *Journal of Biomedical Informatics,* 39: 34-42.

Wilson, C., 2004: Phylogeny of *Iris* based on chloroplast *matK* gene and *trnK* intron sequence data. *Molecular Phylogenetics and Evolution,* 33: 402 - 412.

Wolf, P. G., Rowe, C. A., Sinclair, R. B., and Hasebe, M., 2003: Complete nucleotide sequence of the chloroplast genome from a leptosporangiate fern, *Adiantum capillus-veneris* L. *DNA Research,* 10.

Woessner, J. P., Gillham, N. W., and Boynton, J. E., 1986: The sequence of the chloroplast *atpB* gene and its flanking regions in *Chlamydomonas reinhardtii*. *Gene,* 44: 17-28.

Young, N. D. and dePamphilis, C. W., 2000: Purifying selection detected in the plastid gene *matK* and flanking ribozyme regions within a group II intron of nonphotosynthetic plants. *Molecular Biology and Evolution,* 17: 1933 - 1941.

Zanis, M. J., Soltis, D. E., Soltis, P. S., Mathews, S., and Donoghue, M. J., 2002: The root of the angiosperms revisited. *Proc Natl Acad Sci U S A,* 99: 6848-6853.

Zwickl, D. J. and Hillis, D. M., 2002: Increased Taxon Sampling Greatly Reduces Phylogenetic Error. *Systematic Biology,* 51: 588-598.

Chapter 2: Impact of missing data, gene choice and taxon sampling on phylogenetic reconstruction: the Caryophyllales (angiosperms)

This work has been published in Plant Systematics and Evolution with the following list of authors: Sunny S. Crawley and Khidir W. Hilu. Permission was obtained from Springer to reprint this material here.

Abstract

Density of taxon sampling and number/kind of characters are central to achieving the ultimate goals in phylogenetic reconstruction: tree robustness and improved accuracy. In molecular phylogenetics, DNA sequence repositories such as GenBank are potential sources for expanding datasets in two dimensions, taxa and characters, to the level of "supermatrices". However, the issue of missing characters/genomic regions is generally considered a major impediment to this endeavor. We used here the angiosperm order Caryophyllales to systematically address the impact of missing data when expanding taxon sampling and number of characters in phylogenetic reconstruction. Our analyses show that expansion of taxon sampling by ~13 fold resulted in improved phylogenetic assessment of the Caryophyllales despite up to 38% missing data. Expanding number of characters in the dataset by allowing for up to 100 fold increase in amount of missing data and inclusion of entries with about 40% missing genomic regions did not negatively impact tree structure or robustness, but to the contrary improved both. These results are timely with the ongoing efforts to achieve a detailed assessment of the tree of life.

Introduction

Advances in DNA sequencing technologies have resulted in an abundance of sequence data, creating a wealth of potential sources of molecular information for phylogenetic reconstruction. Systematists are now challenged to find effective and efficient sampling strategies to resolve phylogenies. In plants, studies using sequences at, or close to, the whole plastid genome scale from a relatively small number of taxa have resolved many historically problematic angiosperm backbone relationships (Leebens-Mack et al. 2005;

Qiu et al. 2006; Jansen et al. 2007; Moore et al. 2007; Turmel at al. 2009; Jansen et al. 2011). Although this approach may establish reliable backbone phylogenies, it has been demonstrated that phylogenetic accuracy and tree robustness benefit from increasing both taxon sampling and number of characters (Hillis 1996; Graybeal 1998; Rannala et al. 1998; Zwickl and Hillis 2002; Hillis et al. 2003; Soltis et al. 2004; Brockington et al. 2009; Wang et al. 2009). To expand taxon sampling at the whole genome level considerable investments in time, labor, and expense are required.

Sequence data have been accumulating at a staggering rate in various databases such as GenBank where over 98 million sequence entries had been submitted by 2008 compared to ~3 million 10 years ago (GenBank 2009). However, it is not clear if sampling opportunistically from the wealth of publically available sequences, creating sparsely distributed "supermatrix" datasets with varied amounts of missing data, can achieve results similar to whole genome scale analyses.

In most multigene phylogenetic studies, regions with abundant partial sequences and/or taxa missing whole genomic regions are excluded to avoid any potential negative impact on phylogenetic reconstruction (Hilu et al. 2003; Philippe et al. 2004; Pryer et al. 2004). Wiens (2003a) noted that it is a common practice in many studies to exclude taxa based on the proportion of their missing data, although this practice is not always explicitly stated. Such approaches reduce both number of characters used and taxa sampled.

The impact of missing data in molecular phylogenetics is not well explored, particularly with empirical data. Earlier studies in this regard were concerned mainly with combining fossil information and morphological characters for extant species (Gauthier 1986; Donoghue et al. 1989; Huelsenbeck 1991; Novacek 1992; Wiens and Reeder 1995; Wilkinson 1995; Gao and Norell 1998; Kearney and Clark 2003). Most of these studies concluded that the inclusion of missing data increased the number of most parsimonious trees, leading to greater ambiguity in the strict consensus tree. Using mostly simulated data, Wiens (1998, 2003a, b, 2005, 2006) concluded that generally increasing number of characters despite missing data is more likely to increase phylogenetic accuracy and that

the number of missing characters is not as important as the total number of characters analyzed. In contrast, Agnarsson and May-Collado (2008) asserted that the inclusion of missing data could compromise phylogenetic accuracy. In a recent empirical study, Burleigh et al. (2009) analyzed a 5-gene data matrix with 27.5% missing data and showed that adding genomic regions despite missing data increases support for phylogenetic relationships and may lead to new phylogenetic hypotheses. That study, however, did not examine the impact of varied proportions of missing data, incomplete representation of genomic regions, and taxon sampling on phylogenetic reconstruction.

We used here the order Caryophyllales to address the interplay between expanding number of genomic regions/characters and taxon density while allowing for varying degrees of missing data. The order is a major angiosperm clade, containing approximately 11,155 species from 34 families (APG III 2009; Stevens 2010) and encompassing diverse life forms such as carnivorous and desert plants and extreme variation in floral morphology (Kubitzki et al. 1993). Traditionally, Caryophyllales were circumscribed by P3 type of sieve element plastid, curved embryo, free central placentation (sometimes basal), and perisperm synapomorphies, (Bittrich 1993; Behnke 1994). Subsequent molecular evidence provided additional support for the monophyly of the Caryophyllales (Chase et al. 1993; Savolainen et al. 2000; Hilu et al. 2003; Soltis et al. 2003). Two recent studies have made much progress in resolving the backbone phylogenetic structure within the order (Cuénoud et al. 2002; Brockington et al. 2009). The analyses of Cuénoud et al. (2002) used nuclear 18S rDNA and plastid *rbcL*, *atpB*, and partial *matK* sequences from 26 to 127 taxa, providing an overall structure for Caryophyllales, but leaving some relationships unresolved (i.e. the placement of Rhabdodendraceae). Brockington et al. (2009) used sequences from nine plastid genes (*atpB*, *matK*, *ndhF*, *psbB*, *psbT*, *pbsN*, *rbcL*, *rpoC2*, and *rps4*), two nuclear genes (18S and 26S), and the entire plastid inverted repeat from 36 taxa, and the resulting phylogeny provided a strongly supported backbone for most of the major clades of Caryophyllales. The analysis of Brockington et al. (2009) is an archetypal example of a study that resolved a classically difficult phylogenetic problem by sequencing a large number of loci from a relatively small number of taxa. However, it is also the latest in a large

number of molecular systematic studies of Caryophyllales taxa (see Meimberg et al. 2001; Cuénoud et al. 2002; Nyffeler 2002; Kadereit et al. 2003; Edwards et al. 2005; Müller and Borsch 2005; O'Quinn and Hufford 2005; Fior et al. 2006; Fior and Karris 2007; Nyffeler 2007; Sanchez and Kron 2008), which have together produced an abundance of sequence data.

In this study, we examine the efficacy of a "supermatrix" approach to resolve the Caryophyllales backbone phylogeny, and explore tradeoffs between adding more taxa and/or genes at the cost of greater fragmentation of the character matrix by incomplete gene sequences and genomic region representation. New sequence data and GenBank sequences for 5 plastid genomic regions, *matK*, *trnK* intron, *atpB*, *rbcL*, and *ndhF*, were analyzed. These regions differ in mode and tempo of evolution in being coding and non-coding regions, and vary in rate of amino acid substitution, transition/transversion ratio, and presence and frequency of gaps. We are measuring phylogenetic robustness by number of nodes resolved and bootstrap support (sensu Källersjö et al. 1999), and phylogenetic accuracy by comparing the resulting trees to an existing robust phylogeny for the Caryophyllales (Brockington et al. 2009).

We used the Brockington et al. (2009) tree as a reference because it is the most robust tree available for the Caryophyllales in terms of resolution and support. The Brockington et al. (2009) study represents all the Caryophyllales families recognized by APG II (APG 2003) and is based on rigorous phylogenetic analyses of partitioned and combined data sets of various gene combinations totaling 42,006 characters. Further, the backbone topology of the Brockington et al. (2009) phylogeny is supported by previous studies of the Caryophyllales (Rettig et al. 1992; Downie and Palmer 1994; Downie et al. 1997; Cuénoud et al. 2002). The Brockington et al. (2009) phylogeny provides improved bootstrap support for some difficult to place lineages such as the Rhabdodendraceae and Simmondsiaceae.

In addition to the robustness of the Brockington et al. (2009) tree, structural synapomorphies exist for some major lineages, supporting the phylogenetic structure of

the tree. For example, within core caryophyllids, a clade comprising Aizoaceae, Phytolaccaceae, Nyctaginaceae, Gisekiaceae, Molluginaceae, Portulacaceae, Didiereaceae, Basellaceae, and Cactaceae is defined by a distinctive P plastid characteristic (Behnke 1994). This clade, which has been termed the "globular inclusion clade," was recovered in previous studies (Giannasi et al. 1992; Rettig et al. 1992; Downie et al. 1997; Cuénoud et al. 2002). Two subclades are nested within this clade: the raphide subclade, characterized by the presence of raphide crystals and an involucre, and succulents clades characterized by CAM metabolism (Judd et al. 2008). The non-core clade is characterized by secretory cells containing plumbagin, stalked, vascularized, gland-headed hairs, basal placentation and starchy endosperm (Judd et al. 2008), although some characters have been lost in one or more lineages. In the non-core group, one of the two lineages is circumscribed mostly by the carnivorous biology (carnivorous clade) whereas the other one is supported by ovaries with a single basal ovule and usually indehiscent fruits (Judd et al. 2008). Therefore, the robustness of the Brockington et al. (2009) tree, its congruence with previous phylogenetic studies on the Caryophyllales, and the presence of structural and chemical synapomorphies for the recovered clades renders this phylogenetic tree as the optimal choice for a comparative phylogenetic analysis.

Materials and Methods

Taxon sampling and genomic regions

In total, we included sequences from 652 species, representing 33 of the 34 Caryophyllales families recognized by APG III (2009). Six core eudicot species from the families Vitaceae, Dilleniaceae, Berberidopsidaceae, and Santalaceae were used as outgroups. We generated 50 new, complete or partial sequences for *matK* and *trnK* intron (both the 5' and 3' regions). Additionally, we sampled GenBank sequences for *matK* and *trnK* intron and added three other plastid regions (*rbcL*, *atpB*, and *ndhF*) commonly used in plant systematic studies (see Olmstead et al. 1992; Clark et al. 1995; Hoot et al. 1995; Smith et al. 1997; Källersjö et al. 1998; Applequist and Wallace 2001; Smissen et al. 2002; Li 2008; Olmstead et al. 2009). The *matK* gene (~1550 base pairs – bp) is nested

within the *trnK* intron (~700 bp), splitting the intron into 5' and 3' regions. We will collectively refer to these 5' and 3' regions as the *trnK* intron to distinguish them from the *matK* open reading frame (ORF). Both *matK* and *trnK* intron are rapidly evolving genomic regions (Hilu et al. 2003; Müller et al. 2006). In contrast, *rbcL* (~1350 bp) and *atpB* (~1400 bp) are regarded as slowly evolving genes (Wolf 1997; Soltis et al. 2000) while *ndhF* (~2100 bp) has an intermediate rate of evolution (Alverson et al. 1999).

For newly generated sequences, genomic DNA was either extracted from plant material collected in the field or from plants grown in the greenhouse and stored at –80°C. Additional DNA samples were obtained from various sources. Information on species used, their taxonomic affiliations, sources of plant material, and GenBank numbers for all sequences are listed in Appendicies B - D.

Generation of new matK and trnK intron sequences

Genomic DNA extraction followed the CTAB method of (Doyle and Doyle 1990). To generate new sequences for the entire *matK* ORF and *trnK* intron, the region was amplified in three overlapping sections using a combination of external primers located in the *trnK* exons and internal *matK* primers. Partial sequences were completed using DNA samples from the same exact species. In some cases these sequences were not completed from the same individual but we do not expect this to impact phylogenetic study at this deep level. The following primers were newly designed for this study: CaryomatK291F (5' GGATTTGCAGTCATTGTGG 3'), CaryomatK467R (5' GTAGANCTTTAGAACCAAG 3'), Caulo1100F (5' GCATCCCATTAGTAARCCG 3'), PImatK1326R (5' TCTAGCACAAGAAAGTCGAAGT 3'), and AsteromatK500R (5' CCAAGTTTGAGAAGCGATGACCC 3'). Additionally, we used previously published primers: trnK3914Fdi (Johnson and Soltis 1995), corematK1 (Barthet and Hilu 2007), TOmatK480F (Hilu et al. 2003), and MG1 (Liang and Hilu 1996).

The DNA was amplified on either a PTC-100 or a PTC-200 Peltier Thermal Cycler (MJ Research, Waltham, Massachusetts). The 25 µL PCR reactions contained 1.0 µL of 20

mM primer, 0.2 µL of 5,000 U/mL Taq polymerase, 2.5 µL thermopol I 10x buffer with MgSO4 (New England BioLabs, Ipswich, Massachusetts), and 0.5 µL of 10 mM dNTPs (Promega, Madison, Wisconsin). The thermocycling profile for amplification was: 95°C, 50°C, 72°C for 3 minutes each, followed by 30 cycles of 95°C for 30 s, 50°C for 1:00 min, and 72°C for 4:00 mins with a final elongation step of 20 mins at 72°C. PCR products were cleaned using Qiagen's QIAQuick gel extraction kit (Qiagen, Valencia, California). Sequences were generated either at the VBI core sequencing facility (Applied Biosystems 3730 Automated DNA Sequencer) or at the Duke University sequencing lab (Applied Biosystems 3730 XL DNA Analyzer) using a Big Dye Terminator Cycle Sequencing Ready Reaction Kit (ABI, Foster City, California).

Dataset assembly and alignment

To address the questions of the impact of taxon sampling and amount of missing data (including both ambiguous, "N", and missing, "?", cells) on phylogenetic structure, four primary datasets were generated. The first two address the impact of increasing taxonomic sampling despite missing data: 1) a dataset comprised of almost complete sequences (0.3% missing data) of entire *matK* ORF and *trnK* intron for 51 taxa (MT-51); and 2) a dataset in which the MT-51 was expanded to include an additional 601 complete and partial GenBank sequences for *matK* ORF and *trnK* intron (MT-652). The latter dataset contains 38% missing data. The next two datasets address the impact of increasing the number of genomic regions for a fixed number of taxa while allowing for missing data: 3) this dataset included complete and partial *matK* ORF and *trnK* intron sequences for 136 taxa with 21% missing data (MT-136); and 4) an expanded MT-136 dataset that included complete and partial sequences of *atpB*, *rbcL*, and *ndhF* for the same species when possible and a "placeholder" from the same genus when necessary, resulting in five genomic regions and 46% missing data (5GR-136). Additionally, distribution of missing data for each family and the major clades in the caryophyllids was calculated to determine if there is correlation between amount of missing data and topological differences. In this case we determined the number of missing characters for

each family as well as major clades and calculated the proportion of their contribution to the total amount of missing characters in the dataset.

To address the question of the impact of eliminating taxa due to lack of whole genomic regions in phylogenetic reconstruction, we used the 5-genomic region dataset (5GR-136) and applied different levels of constraint. Dataset 4a included 98 taxa in which each taxon had at least a partial sequence for 3 of the 5 genomic regions (3/5GR); dataset 4b included only taxa represented by at least a partial sequence for 4 of the 5 genomic regions, which resulted in a reduction in taxon sampling to 48 (4/5GR); and dataset 4c comprised of taxa that contained at least a partial sequence for all 5 genomic regions, resulting in a 15 taxon dataset (5/5GR).

Sequences in all datasets were manually adjusted for gaps using QuickAlign (Müller and Müller 2003). Gaps were inserted at the cost of two or more substitutions (Kelchner 2000; Borsch et al. 2003). In the case of *matK* and *ndhF*, the sequences were translated into amino acids and that alignment was used as an additional guide for the insertion of gaps. In the *matK* ORF, out of frame gaps are confined to the very end of the gene and did not present a problem in homology assessment (Hilu and Alice 1999; Hilu et al. 2003). To evaluate our manual alignment we used the program MUSCLE (Edgar 2004) in CIPRES. Our alignment of all datasets where consistent with those obtained with MUSCLE, except for the MT-652. In most instances, differences were either due to our effort to maintain an intact open reading frame for the protein coding genes, or to avoid violating the rules of the cost of gap insertion (Kelchner 2000; Borsch et al. 2003). In the case of the MT-652 dataset that harbors a large amount of missing data caused by partial sequences, the number and placement of gaps inserted by MUSCLE distorted the alignment of some sectors. The alignments for all coding genes were unambiguous and, thus, were used in their entirety. For the *trnK* intron, we excluded regions of the alignment where homology assessment was uncertain, as well as any poly-A/T tracts. Gaps were not included as characters in the data analyses to eliminate variables relating to 1) extreme differences in number of gaps among the five genomic regions used, and 2)

to be able to provide an unbiased comparison with the phylogenetic tree obtained by Brockington et al. (2009) where gaps were excluded from the analyses.

Analyses

Phylogenetic analyses were performed using maximum parsimony (MP) and maximum likelihood (ML). For the MP analyses, the search for the optimal tree was performed using a ratchet heuristic (Nixon 1999) implemented in PAUP* (Swofford 2003). Each ratchet analysis was set to 10 random addition cycles for 200 replicates at a starting weight of 2, with 25% up-weighting in each replicate. A single tree was retained after each replicate search, a strict consensus tree was generated for each individual analysis, and tree scores were recorded as defined in the settings. The settings for the ratchet search were written using PRAP2 (Müller 2004). Confidence estimates for MP analyses were subsequently obtained by performing 1000 bootstrap (BS; Felsenstein 1985) replicates, each consisting of 10 random sequence addition replicates with TBR branch swapping in PAUP* and saving 1 tree per replicate. To assess the potential impact of increased amounts of invariable characters in expanded datasets on BS support (Felsenstein 2004; Freudenstein and Davis 2010), we also calculated jackknife (JK) support values in PAUP* to compare with BS. Support for nodes measured by BS and JK were very comparable, and nodes receiving >50 were the same (See Figs. A.1, A.2, A.3a, and A.3b). Therefore, increase in number of invariable characters does not appear to have an impact on BS support in this study and, consequently, only BS support values will be discussed.

Maximum likelihood analyses were executed using RAxML under the GTR+CAT model (Stamatakis et al. 2008). The data were submitted to the CIPRES portal (www.phylo.org), a maximum likelihood search option selected, outgroups defined, and nonparametric bootstrapping of 1000 replicates executed, and all other settings were defined by default. To compare differences in support across trees with different taxon representation, we used PAUP* to prune from the bootstrap trees taxa that were not shared by pairs of data matrices in question. For example, we pruned 613 and 97 taxa,

respectively, from the MT-652 and 5GR-136 trees when each was compared with the Brockington et al. (2009) 40-taxon tree. Our dataset differed from the Brockington et al. (2009) dataset only in the substitution of *Dillenia* for *Hibbertia* (both Dilleniaceae) in the outgroup, and the lack of *Physena* (Physenaceae) in the ingroup.

Results

The Caryophyllales as currently envisioned (Brockington et al. 2009; this study) are divided into two major lineages, core and non-core Caryophyllales (Figs. 2.1 and 2.2). The later forms two clades, one contains mostly carnivorous families (Nepenthaceae, Droseraceae, Ancistrocladaceae, Dioncophyllaceae and Drosophyllaceae) and another contains the families Frankeniaceae, Tamaricaceae, Polygonaceae, and Plumbaginaceae, which we will refer to here as the "FTPP" clade. In the core Caryophyllales, a grade of Rhabdodendraceae, Simmondsiaceae, and Asteropeiaceae emerge as sister to three clades that are referred to as the "AAC" clade (Amaranthaceae, Achatocarpaceae, and Caryophyllaceae), the "succulents" clade (Portulacaceae, Halophytaceae, Basellaceae, Didiereaceae, and Cactaceae), and the "raphide" clade (Nyctaginaceae, Phytolaccaceae, Sarcobataceae, Gisekiaceae, and Aizoaceae). Two historically difficult to place families, Limeaceae and Stegnospermataceae, appear sister to the succulents + *Mollugo* and raphide clades (Figs. 2.1 and 2.2).

Taxon Density and Impact of Missing Data

We will focus here on the results of the analyses of the MT-51 and MT-652 datasets comprising characters from the *matK* ORF and *trnK* intron and constituting ~13 fold expansion in taxon sampling and ~100 fold increase in missing data. Although the number of variable characters increased with expansion of taxon sampling (1,683 to 2,296), the percent variable characters decreased (72% to 67%), probably due to the insertion of additional indels and increase in missing data (Table 2.1). More importantly, both number and percentages of parsimony informative (PI) characters increased (1,209;

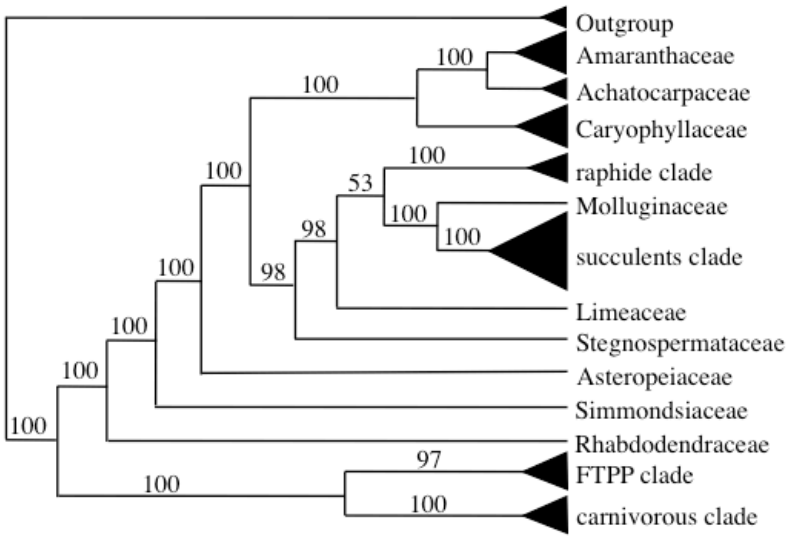

Figure 2.1 Summary of the ML tree of the Caryophyllales based on total evidence (plastid IR region plus eleven other genomic regions) from Brockington et al. (2009; Fig. 1). Percent bootstrap values greater than 50% are noted on branches.

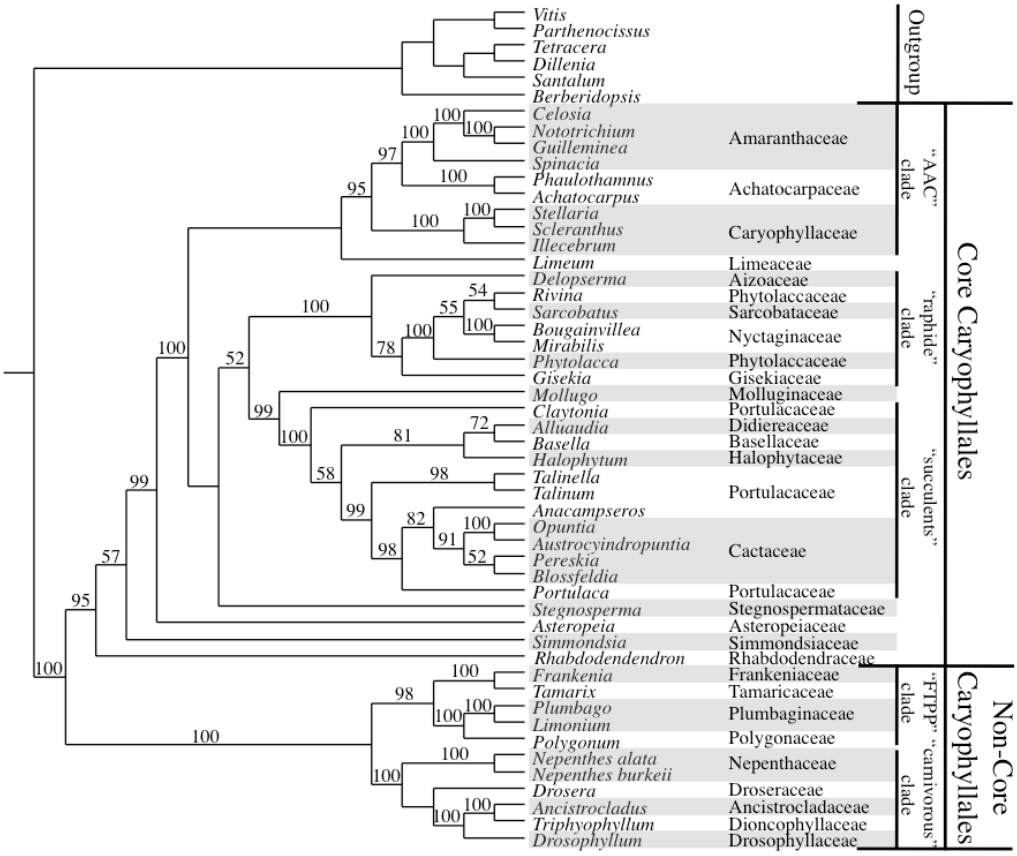

Figure 2.2 ML tree based on the *matK*/*trnK* intron dataset for 51 Caryophyllales taxa (0.3% missing data). Percent bootstrap values greater than 50% are noted on branches. Note the misplacement of the Limeaceae compared to the Brockington et al. (2009) tree (Fig. 2.1).

Table 2.1 Character numbers and maximum parsimony statistics for each of the datasets analyzed.

Dataset Name[a]	Total # of Characters	# Variable (%)	#PI (%)[b]	Tree Length	CI[c]	RI[d]	RC[e]
MT-51	2349	1683 (71.65%)	1209 (71.84%)	6004	0.478	0.560	0.268
MT-652	3443	2296 (66.69%)	1898 (82.67%)	16254	0.284	0.862	0.245
MT-136	2349	1829 (77.86%)	1439 (78.68%)	8221	0.406	0.694	0.282
5GR-136	7234	4046 (55.93%)	2900 (71.68%)	15554	0.430	0.668	0.287
3/5GR	7234	3960 (54.74%)	2791 (70.48%)	14547	0.446	0.627	0.280
4/5GR	7220	3742 (51.83%)	2458 (65.69%)	11788	0.499	0.538	0.269
5/5GR	7153	2674 (37.38%)	1539 (57.56%)	5059	0.702	0.608	0.427

[a] MT=*matK*/*trnK* intron; 5GR=five genomic regions used (*atpB*/*rbcL*/*ndhF*/*matK*/*trnK* intron); 3/5GR, 4/5GR, and 5/5GR=datasets constrained for genomic regions; 51, 652, and 136=number of taxa in each dataset. The number of taxa in the last three datasets is 98, 48, and 15, respectively.

[b] PI (%) = number of parsimony informative characters, percent refers to the percent of PI characters that are variable

[c] CI = Ensemble Consistency Index

[d] RI = Ensemble Retention Index

[e] RC = Ensemble Rescaled Consistency Index

72% to 1,898; 83%). The shortest trees for MT-51 were of 6,004 steps with an ensemble consistency index (CI) of 0.478 and an ensemble retention index (RI) of 0.560. In contrast, the trees based on the MT-652 dataset were 16,254 steps long with CI and RI values of 0.284 and 0.862, respectively (Table 2.1).

The Caryophyllales were recovered with 100% BS support in all cases. The backbone topology of the ML tree obtained from the MT-51 dataset was congruent with the Brockington et al. (2009) total evidence tree (Figs. 2.1 and 2.2) with only one exception. The MT-51 based tree differed in the placement of the Limeaceae, where they emerged as sister to the AAC clade as opposed to being sister to the raphide and succulents + *Mollugo* clades (Figs. 2.1 and 2.2). All major nodes received 95% to 100% BS support (mean 99%) except for the placement of Limeaceae, Stegnospermataceae, and Simmondsiaceae, which received <50% BS (Figs. 2.1, 2.2, and 2.3a). The MT-51 MP strict consensus tree recovered a polytomy of six nodes consisting of Amaranthaceae + Achatocarpaceae (91% BS), Caryophyllaceae, raphide clade (100% BS), succulents clade + *Mollugo* (77% BS), Stegnospermataceae, and Limeaceae. Additionally, Rhabdodendraceae and Simmondsiaceae appeared in a clade instead of a grade but received less than <50% BS support, and Droseraceae were sister to Nepenthaceae (54% BS) instead of Nepenthaceae basal (Fig. A.1).

Expanding the dataset to 652 taxa using GenBank sequences resulted in MP strict consensus and ML trees with overall topologies identical to that of Brockington et al. (2009), correcting the placements of the Caryophyllaceae and Limeaceae found in the MT-51 tree (Figs. 2.2, 2.4, and Fig. A.2). The Limeaceae/Stegnospermataceae and Rhabdodendraceae/Simmondsiaceae grades resolved in Brockington et al. (2009) and MT-51 analyses emerged as two clades in the MP analysis, but with <50% BS support for each relationship (Figs. 2.1, 2.2, 2.4, and Fig. A.2). The BS support for the major clades remained above 92% except for the problematic lineages Limeaceae, Stegnospermataceae, and Simmondsiaceae (Figs. 2.3a and 2.4). Pruning the MT-652 ML tree to be comparable to the Brockington et al. (2009) tree in taxon composition did not impact the topology of the remaining lineages.

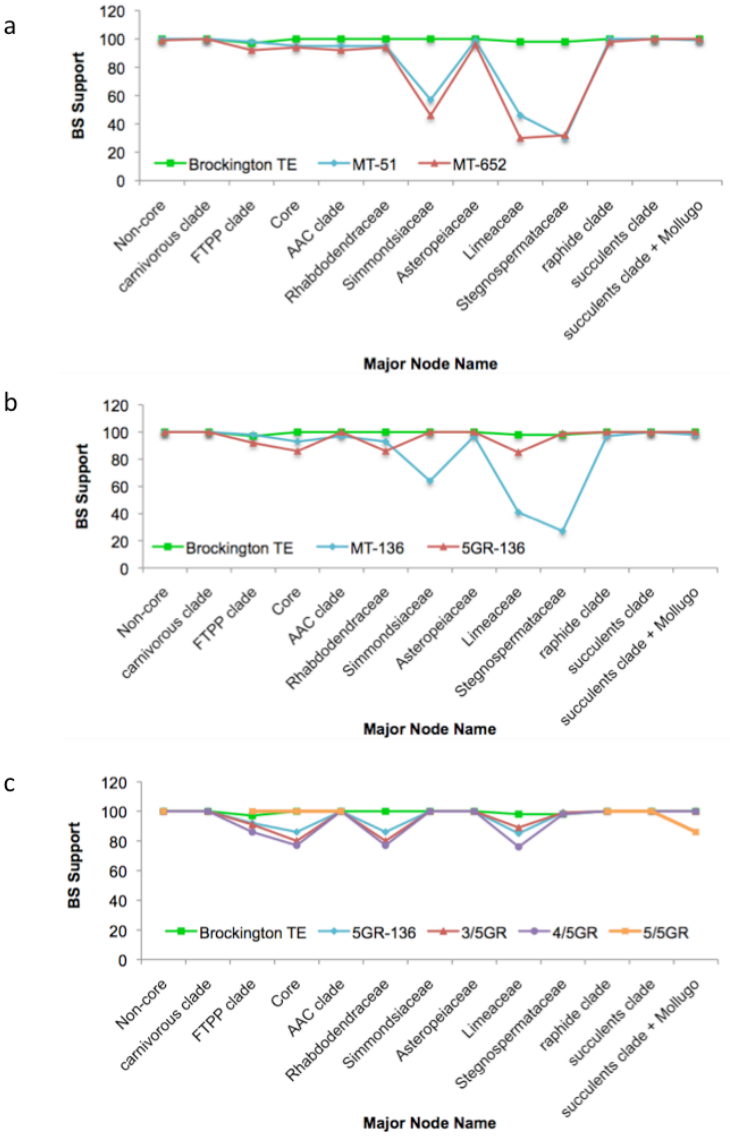

Figure 2.3 Comparison of bootstrap support for the major Caryophyllales nodes in the ML analyses. **a** Support for major nodes obtained using *matK*/*trnK* intron with limited taxon sampling (MT-51), expanded taxon sampling (MT-652), and in the Brockington et al. (2009) study. **b** Support for major nodes using two genomic regions (*matK*/*trnK* intron; MT-136), five genomic regions (5GR-136), and the Brockington et al. (2009) study. **c** Support obtained from the unconstrained five genomic regions (5GR-136), each of the three constrained datasets (3/5GR, 4/5GR, and 5/5GR), and the Brockington et al (2009) study.

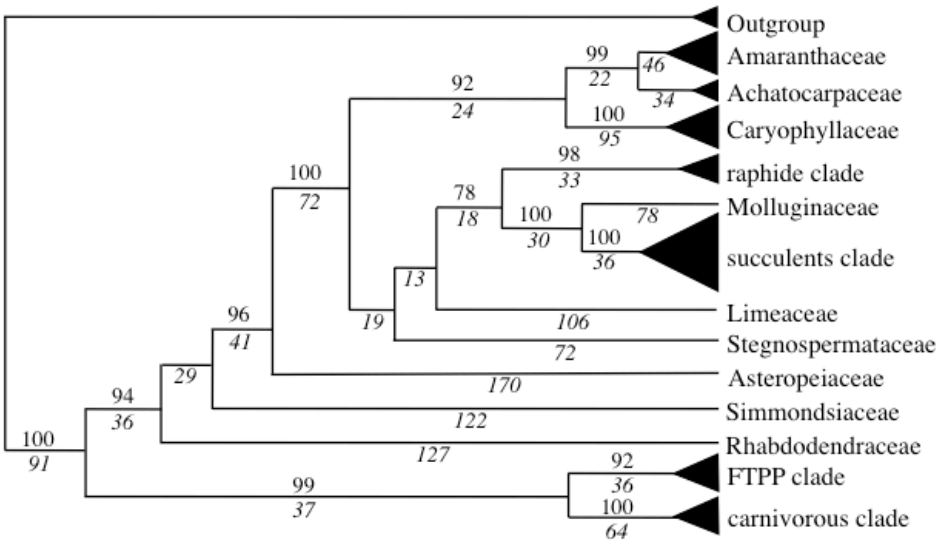

Figure 2.4 Summary of the ML tree based on *matK*/*trnK* intron data with expanded taxon sampling (652 taxa with 38% missing data). Percent bootstrap values greater than 50% are noted on branches and branch lengths are noted below the branches in italics.

Expanding Number of Genomic Regions

Expanding the 136-taxon *matK*/*trnK* intron dataset (MT-136) to include *rbcL*, *atpB*, and *ndhF* (5GR-136) increased the number of total characters from 2,349 to 7,234 and missing data from 21.2% to 46.3%. Similarly, the number of variable characters increased from 1,829 to 4,046 and the number of PI characters doubled (1,439 to 2,900). However, percentages of variable and PI characters decreased (78% to 56% and 79% to 72% respectively), probably due to the inclusion of more slowly evolving genomic regions. The most parsimonious trees increased in length from 8,821 to 15,554 steps despite maintaining a constant number of taxa. The CI value increased while the RI value decreased, albeit the changes in both cases were slight (0.406 to 0.430 and 0.694 to 0.668, respectively; Table 2.1).

The topology of the MT-136 ML tree was congruent with that of the MT-51 tree described above with similar levels of support with the exception of the placement of Limeaceae (Figs. 2.2 and 2.5). The addition of three genomic regions (5GR-136) resulted in an ML tree with a backbone topology identical to that of Brockington et al. (2009), correcting most of the topological inconsistencies detected in the MT-136 and MT-51 trees (Figs. 2.1, 2.2, 2.5, 2.6; and Figs. A.3a, A.3b, and A.3c). Considerable increase in BS support was evident for the placement of the problematic lineages Simmondsiaceae, Stegnospermataceae, and Limeaceae where it went from <50% to 86%, 99%, and 85%, respectively (Fig. 2.3b, 2.5, 2.6). Pruning the 5GR-136 tree to match the Brockington et al. (2009) tree did not impact patterns of divergence among remaining lineages. The MP analysis of the 5GR-136 dataset resulted in a strict consensus tree topology very similar to that of the ML tree but with reduced robustness due to a basal polytomy in the order and a decline in BS support (Figs. A.4a and A.4b).

Constraining Number of Genomic Regions

The focus here will be on the results obtained from applying sequential constraints on number of genomic regions required per taxon using the 5GR-136 dataset. As taxa

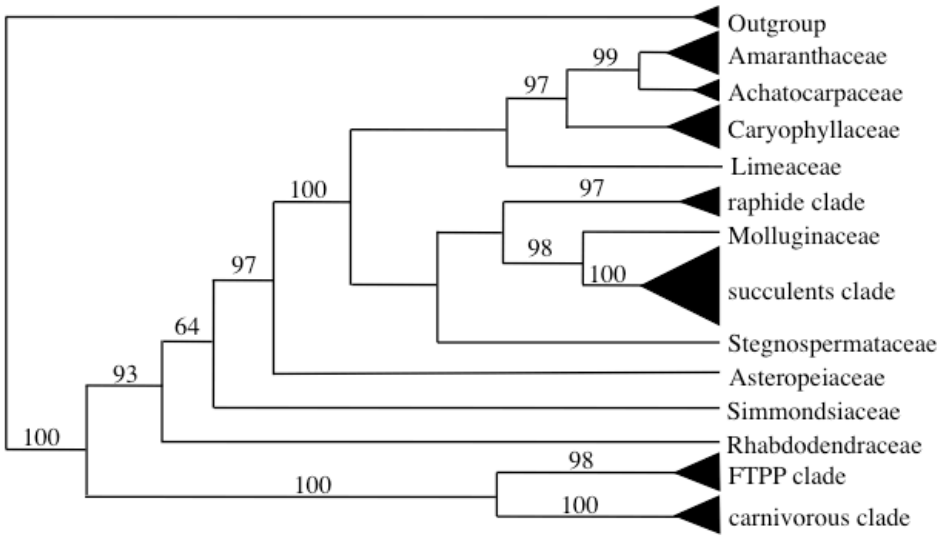

Figure 2.5 Summary of the ML tree based on *matK*/*trnK* intron data for 136 taxa (MT-136; 21% missing data). Percent bootstrap values greater than 50% are noted on branches.

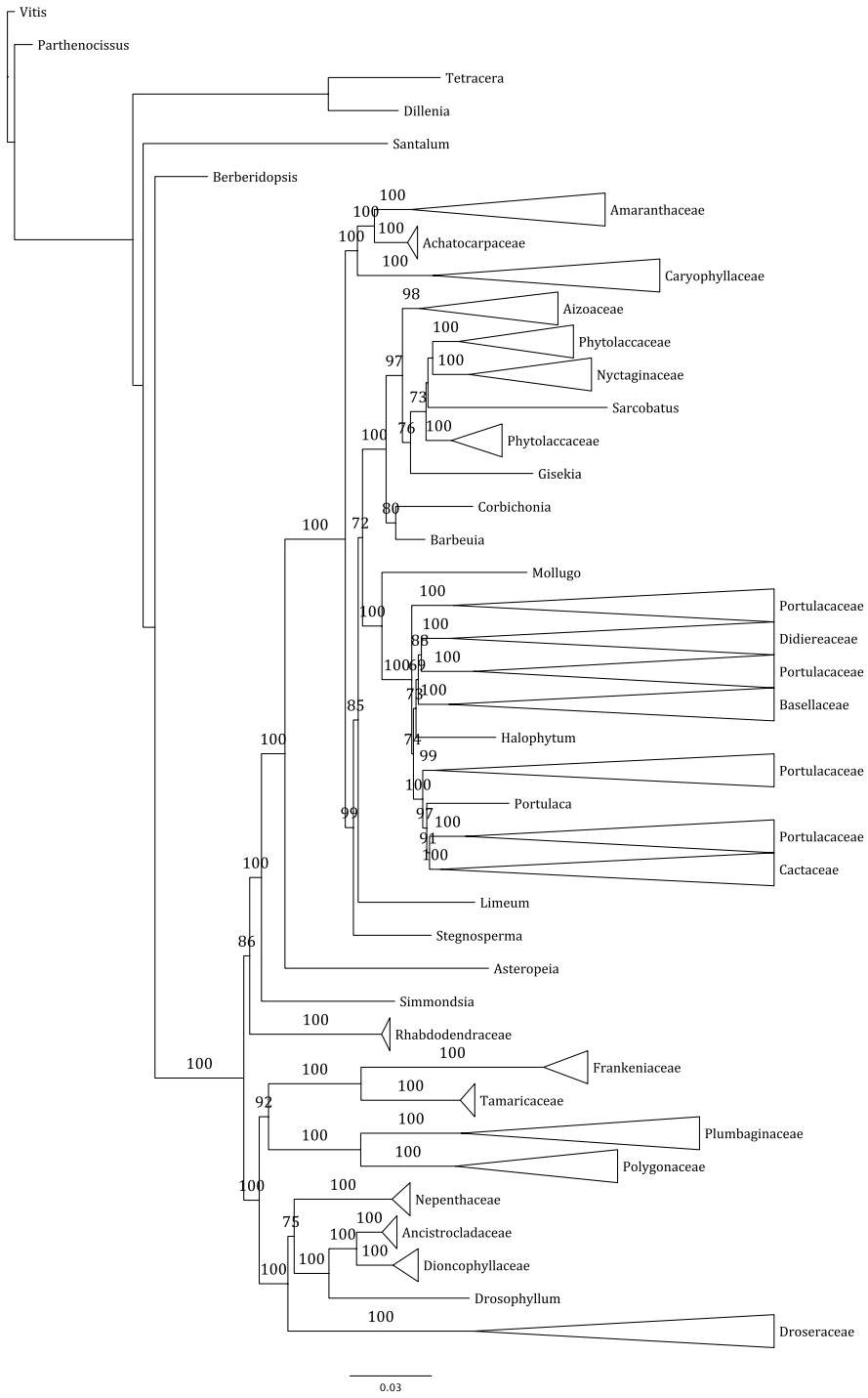

Figure 2.6 Family level detail of the ML tree based on the dataset of five genomic regions (*rbcL*, *atpB*, *ndhF*, *matK*, and *trnK* intron) for 136 taxa (5GR-136; 46% missing data). Percent bootstrap values greater than 50% are noted on branches. Note the substantial increase in bootstrap support for the placement of the problematic lineages Simmondsiaceae, Limeaceae, and Stegnospermataceae compared with the MT-136 (Fig. 2.5). The tree depicting full details can be found as Figs. A5a, A5b, and A5c.

lacking a given number of genomic regions were excluded, the percent of missing data declined from 46.3% (5GR-136) to 35.6% in the 3/5GR (98 taxa) dataset, 17.1% in the 4/5GR (48 taxa) dataset, and 2.2% for the 5/5GR (15 taxa) dataset (Table 2.2). Correspondingly, the number and percent of variable characters decreased from 4,046 (56%) in 5GR-136 to 2,674 (37%) in 5/5GR and the PI characters from 2,900 (72%) to 1,539 (58%) (Table 2.1). The CI values increased (0.430 to 0.702) whereas the RI values fluctuated (Table 2.1). Tree length decreased with reduction in number of taxa following the application of these constraints (15,554 to 5,059; Table 2.1).

The ML trees obtained from the analyses of the original 5GR-136 dataset and its three constrained sub-datasets were identical in backbone topology to the Brockington et al. (2009) tree (Fig. 2.7); the 5/5GR-based tree was lacking in several lineages due to the constraints applied. However, support for some of the major nodes decreased with the increase in constraints (Figs. 2.3c, and 2.7). The exception was the 5/5GR-based tree where support increased, probably due to the drastic reduction in the number of taxa (from 136 to 15) and the consequent exclusion of some entire lineages such as the carnivorous clade, Rhabdodendraceae, Simmondsiaceae, and Asteropeiaceae.

Discussion

One of the prominent questions in molecular phylogenetics is the balance between desirable taxon density and the optimal number of characters needed to increase phylogenetic robustness and accuracy (Hillis 1996; Zwickl and Hillis 2002; Hillis et al. 2003; Kearney and Clark 2003; Rokas and Carroll 2005; Agnarsson and May-Collado 2008; Burleigh et al. 2009). Empirical and simulation studies have provided conflicting results (Rosenberg and Kumar 2001; Kearney 2002; Pollock et al. 2002; Zwickl and Hillis 2002; Kearney and Clark 2003; Agnarsson and May-Collado 2008). We will discuss here the results of our comparative phylogenetic analyses of datasets with varied constraints on taxon density, number of characters and genomic regions, and amounts of missing data using the Caryophyllales. As a means of measuring phylogenetic robustness and accuracy, we compared the trees recovered here with the robust tree obtained in the

Table 2.2 Amount of missing data in each dataset attributed to missing characters (?) and ambiguous character states (N). Missing data are calculated for individual genomic regions and then combined to give the total amount of missing data for the dataset.

Dataset	Char Type	# of Taxa	*matK*	*trnK* intron	*rbcL*	*atpB*	*ndhF*	Total
MT-51	Both	51	0.13%	0.58%	N/A	N/A	N/A	0.27%
	?		0.00%	0.45%	N/A	N/A	N/A	0.14%
	N		0.12%	0.13%	N/A	N/A	N/A	0.13%
MT-652	Both	652	23.73%	53.78%	N/A	N/A	N/A	38.35%
	?		23.72%	53.78%	N/A	N/A	N/A	38.35%
	N		0.01%	0.01%	N/A	N/A	N/A	0.01%
MT-136	Both	136	14.10%	36.30%	N/A	N/A	N/A	21.19%
	?		14.00%	36.22%	N/A	N/A	N/A	21.10%
	N		0.10%	0.07%	N/A	N/A	N/A	0.09%
5GR-136	Both	136	14.10%	36.30%	32.16%	74.10%	64.38%	46.26%
	?		14.00%	36.22%	31.64%	74.06%	64.02%	46.03%
	N		0.10%	0.07%	0.52%	0.03%	0.35%	0.24%
3/5GR	Both	98	3.86%	11.90%	21.59%	64.05%	57.61%	35.56%

	?		3.76%	11.79%	21.04%	64.01%	57.12%	35.27%
	N		0.10%	0.10%	0.55%	0.05%	0.49%	0.29%
4/5GR	Both	48	1.32%	2.52%	1.00%	39.08%	29.59%	17.07%
	?		1.20%	2.40%	0.27%	39.00%	28.78%	16.64%
	N		0.11%	0.13%	0.73%	0.08%	0.82%	0.43%
5/5GR	Both	15	0.19%	0.56%	1.56%	1.93%	4.90%	2.22%
	?		0.00%	0.50%	0.01%	1.77%	4.77%	1.81%
	N		0.19%	0.06%	1.55%	0.16%	0.13%	0.41%

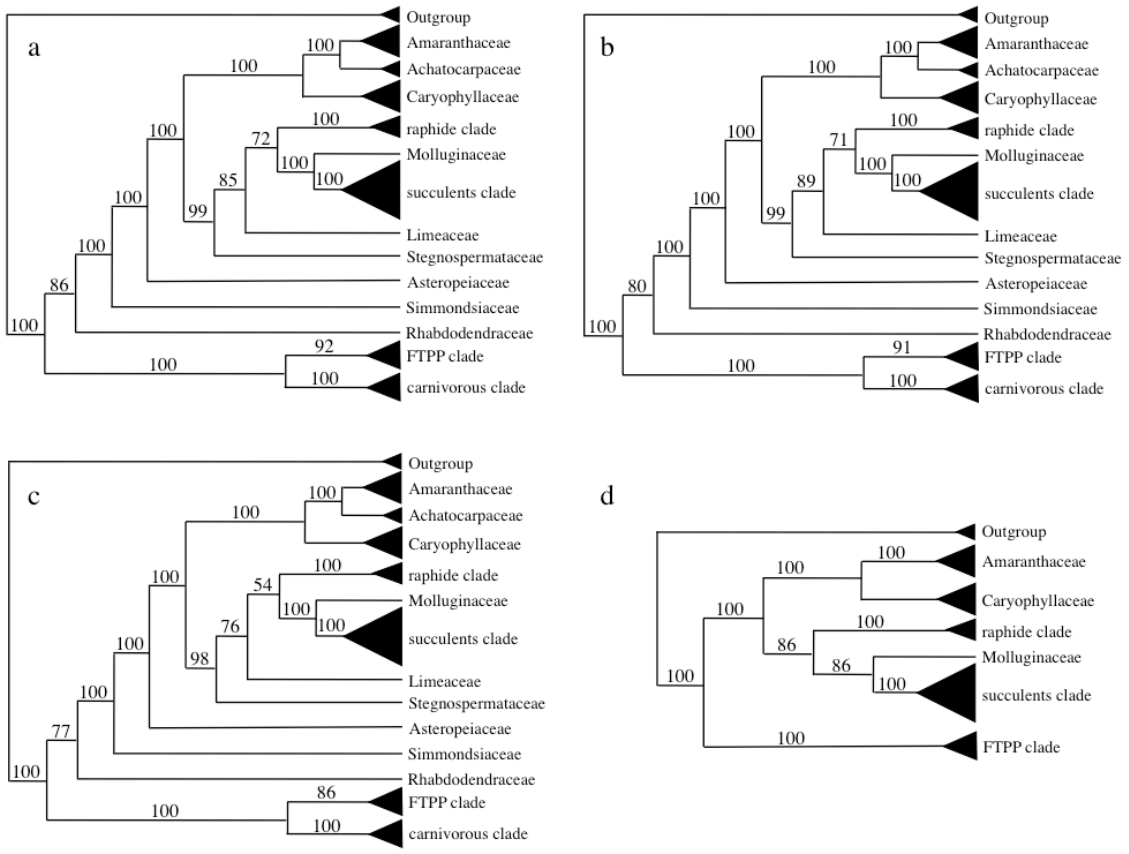

Figure 2.7 Summaries of ML trees displaying the phylogenetic impact of constraining the original five genomic region dataset (5GR-136) by retaining taxa based on number of genomic regions available. **a** Unconstrained tree based on the 5GR-136 dataset. **b** A tree constrained by allowing taxa that have at least a partial sequence for 3 of the 5 genomic regions (3/5GR; 98 taxa; 38% missing data). **c** A tree constrained by allowing taxa that have at least partial sequences for 4 of the 5 genomic regions represented (4/5GR; 48 taxa; 17% missing data). **d** A tree constrained by allowing taxa that have all 5 genomic regions represented by at least a partial sequence (5/5GR; 15 taxa; 2% missing data). Percent bootstrap values greater than 50% are noted on branches.

Brockington et al. (2009) study. We will focus on results of the ML analyses in accordance with the Brockington et al. (2009) paper; nevertheless we will make reference to MP results as it becomes relevant.

Taxon Density and Phylogenetic Accuracy

The first analyses were based on 51 taxa and a nearly complete *matK*/*trnK* intron dataset (only 0.3% missing data; MT-51) representing 33 of the 34 caryophyllid families (APG III 2009). Because this dataset is almost completely lacking missing data and the taxa provide balanced representation for the order, it will be used as a point of reference in our discussion.

The ML tree recovered from the MT-51 analysis was congruent in topology and overall support to that of Brockington et al. (2009) except for placement of Limeaceae and some relationships within the carnivorous clade (Figs. 2.1 and 2.2). Approximately 75%, of the nodes received >80% BS support (Figs. 2.2 and 2.3a). This moderate to strong support (Pirie et al. 2008; Rønsted et al. 2008; Kawahara et al. 2009; Wilson 2009) is quite comparable to the Brockington et al. (2009) tree (79% of nodes; Fig. 2.1). It is worth noting that most of the nodes that were not well supported in this ML tree also received very low support in the Brockington et al. (2009) ML tree. Examples of those nodes are the placement of Droseraceae within the carnivorous clade and the sister group relationship of the raphide clade to the succulents + *Mollugo* clade (Figs. 2.1 and 2.2). These nodes have been historically problematic as is evident from incongruent topologies and/or insufficient support (e.g. Albert et al. 1992; Williams et al. 1994; Savolainen et al. 2000; Cameron et al. 2002; Cuénoud et al. 2002; Hilu et al. 2003). Therefore, it seems that some lineages are difficult to place regardless of number of characters utilized, probably due to inherent rapid radiation and/or slow rates of gene evolution in these lineages. The placement of Droseraceae may be due to its long branch as demonstrated in our study (Fig. 2.6 and Figs. A.3c and A.5) and in previous ones (Cuénoud et al. 2002; Brockington et al. 2009). In contrast, the divergence of the raphide clade from the

succulents + *Mollugo* clade displayed relatively short branches in the studies cited above, implying potential rapid radiation.

Therefore, with the exception of the low BS support for the problematic Simmondsiaceae and Stegnospermataceae and the misplacement of the Limeaceae (<50%; Figs. 2.2, and 2.3a), phylogenetic information from 2,350 characters of rapidly evolving *matK*/*trnK* intron provided a robust Caryophyllales phylogeny that is highly congruent with the Brockington et al. (2009) tree. The latter tree was based on 42,006 characters (18 fold difference) from predominantly slowly evolving regions. This finding underscores the utility of rapidly evolving genomic regions at this historic level.

Increasing taxon density from 51 to 652 (MT-652) using GenBank sequences regardless of amount of missing characters recovered an ML tree with a backbone identical to the Brockington et al. (2009) tree, correcting the placement of Limeaceae (Figs. 2.1, 2.2, and 2.4). This was achieved despite 38% missing data, largely due to a relatively low representation of *trnK* intron and 5' *matK* sequences in GenBank. While major nodes experienced only slight decrease in BS support (Fig. 2.3a), pronounced overall decline in bootstrap values (51% of nodes with >80% BS) could be attributed to intrageneric relationships where denser sampling was used for some genera. This is not unexpected since phylogenetic resolution for *matK* ORF is low at this taxonomic level. To confirm this, the ML MT-652 tree was pruned to match the sampling of Brockington et al. (2009). In that case, BS support for nodes receiving >80% increased to 76%, which is quite comparable to 79% in the Brockington et al. (2009) analysis. However, the low support for the three problematic lineages (Simmondsiaceae, Stegnospermataceae, and Limeaceae) remained unchanged from the MT-51 analysis. Therefore, expanding taxon density by about 13 folds but maintaining the same genomic regions and despite 38% missing data, an accurate (compared with Brockington et al. 2009) and detailed picture of caryophyllid phylogeny was achieved. Our finding is in agreement with previous conclusions indicating that increased taxon sampling can enhance phylogenetic accuracy and robustness while allowing for broader sampling (Graybeal 1998; Pollock et al. 2002; Zwickl and Hillis 2002; Kearney and Clark 2003; Agnarsson and May-Collado 2008).

The pattern of the distribution of missing data in alignments was examined by Wiens (2003a, 2005) and found that it does not impact phylogenetic reconstruction particularly when large amounts of data are used. McMahon and Sanderson (2006) indicated that a minimum of two genomic regions per taxon is a necessary criterion when incomplete taxa are included. To assess the potential impact of unequal distribution of missing data across lineages in the expanded (MT-652) dataset, resolution and support for the five major lineages was compared with those obtained from the analysis of our MT-51 dataset. The percent missing data contributed by each clade ranged from 5.9% in the carnivorous to 45.0% in the AAC clade, with the FTPP, raphide and succulents clades contributing 25.4%, 6.0%, and 16.0%, respectively (Fig. 2.8). Despite this uneven distribution of missing data among lineages, resolution and support for the backbone of the Caryophyllales and within these major lineages was not impacted (Figs. 2.2, 2.4). The amount of missing data does not seem to be a factor with regard to the problematic lineages (Limeaceae, Stegnospermataceae, and Simmondsiaceae) that have been inconsistently resolved or received low BS support (also see Fig. 2.3), since it is negligible in our combined analyses (Fig. 2.8) as well as in the partitioned analyses. Conversely, the presence of the highest proportion of missing data in the Cactaceae (22.7%; Fig. 2.8a) did not impact the family position or phylogenetic structure. Similar conclusions can be drawn for families that have relatively high amounts of missing data such as Plumbaginaceae (9.6%), Amaranthaceae (7.4%), and Nyctaginaceae (6.2%) (Fig. 2.8a).

Increasing sample sizes affected the CI and RI values inversely; the RI increased substantially (0.560 to 0.862) whereas the CI decreased (0.478 to 0.284). The decrease in CI value is not unexpected as has been noted in other large-scale studies (Savolainen et al. 2000; Hilu et al. 2003; Soltis et al. 2003; Qiu et al. 2005). The usefulness of the CI value as a measure of homoplasy has been questioned (Farris 1989; Källersjö et al. 1999; Whittall et al. 2006). For example, Källersjö et al. (1999) stated that "the consistency index… gives exactly the wrong impression of the relative merits of positions" and that

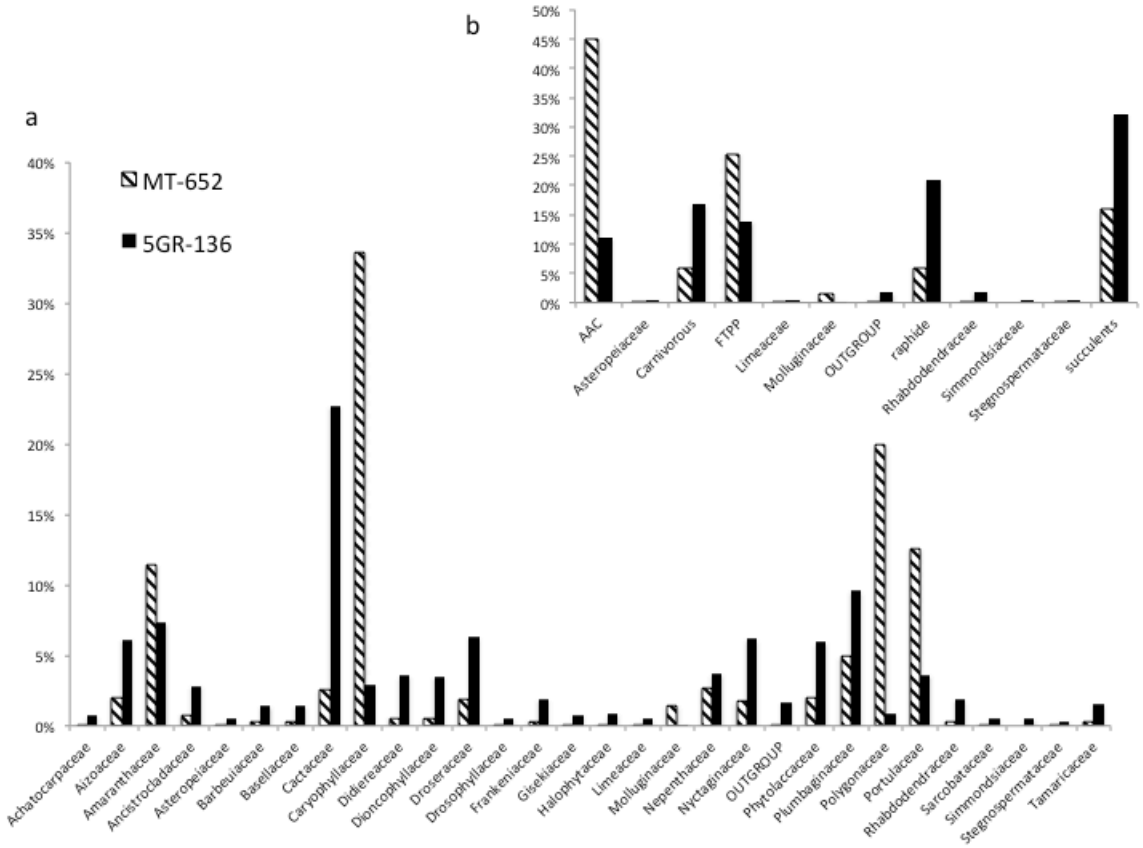

Figure 2.8 Proportion of missing data in the *matK*/*trnK* 652 taxon dataset (MT-652) and the five genomic region dataset (5GR-136). **a** Distribution of missing data among families. **b** Distribution of missing data among major clades.

"a much better evaluation is provided by the retention index." This study further supports the usefulness of RI as a measure of homoplasy.

Expanding Number of Characters

We will compare here the phylogenetic trees based on the 5-genomic regions, 136 taxon/7,234 character dataset (5GR-136) with those derived from a corresponding *matK*/*trnK* intron dataset with 2,349 characters (MT-136) and contrast the two with the Brockington et al. (2009) study (42,006 characters). The ML tree recovered in the MT-136 analysis was congruent in topology with that of the MT-51 ML tree (with the exception of the placement of Limeaceae), having similar levels of support and topological incongruence with the Brockington et al. (2009) tree (Figs. 2.2 and 2.5). The topology of the ML 5GR-136 tree, however, was completely congruent with that obtained by Brockington et al. (2009) and displayed very strong BS support (Figs. 2.1, 2.3b, and 2.6). In fact, support for the problematic lineages Simmondsiaceae, Stegnospermataceae, and Limeaceae approached that in Brockington et al. (2009) (Figs. 2.1 and 2.6). About 73% of the nodes received >80% BS support compared to 79% in the Brockington et al. (2009) tree. This robustness in the 5GR-136 tree was attained despite immense difference in the number of characters and genomic regions utilized (five genomic regions vs. entire plastid IR region plus 11 plastid and nuclear genes) and 46% missing data from across the five genomic regions (Table 2.2). Pruning the bootstrap tree to match the Brockington et al. (2009) sampling increased nodes receiving >80% BS support to 76%.

We conducted a series of partitioned analyses to determine the source of signal from the different genomic regions for the placement of the Limeaceae and Stegnospermataceae and the increased support for the Simmondsiaceae. Phylogenetic trees based on partitioned *matK*, *trnK* intron, and *atpB* data (*ndhF* sequences unavailable for these lineages) showed conflicting topologies for these three lineages, but with <50% BS support for the misplacement (data not shown). In contrast, *rbcL* recovered a topology similar to that obtained by Brockington et al. (2009) with strong support for their

placement (Stegnospermataceae 100%, Simmondsiaceae 87%, and Limeaceae 84% BS). It is evident that the problems of incongruence in the partitioned analyses are due to low signal rather than strong conflicting signal from the genomic regions. These results show that the inclusion of regions with different mode and tempo of evolution despite incomplete representation of genomic regions can positively impact phylogenetic reconstruction. It has been suggested that it is the proportion of signal to noise from different genomic regions that ultimately results in increased tree robustness (Qiu et al. 2005).

Influence of Missing Data

We approached the impact of missing data on phylogenetic reconstruction in three ways, one by increasing taxon sampling but maintaining genomic regions (MT-51 vs. MT-652), two by maintaining taxon sampling but expanding in genomic regions (MT-136 vs. 5GR-136), and three by applying constraints on the number of missing genomic regions in the 5GR-136 dataset (3/5GR, 4/5GR, and 5/5GR). The subsequent increase in missing data in the first and second approaches to 38% and 46%, respectively, did not negatively impact tree structure and accuracy. To the contrary, support for the major nodes increased and placements of the problematic lineages were corrected when compared with our MT-51 dataset (0.3% missing data; Figs. 2.2, 2.3a, 2.3b, 2.4, 2.5, 2.6). The increase in BS support is evidently due to increase in the number of overall characters in the dataset.

In the third approach, overall topology in all cases was identical to that obtained by Brockington et al. (2009) (Figs. 2.1 and 2.7b-d). Surprisingly, with the systematic elimination of taxa with high proportion of missing data, due to lack of whole genomic regions, support for some of the major nodes declined (Figs. 2.3c and 2.7). Most notably, support for the difficult to place lineages Stegnospermataceae and Limeaceae declined drastically when 3/5 genomic regions (17% missing data) were required (Figs. 2.3c and 2.7b). This decline in support might be due to the elimination of potential phylogenetic signal when taxa were trimmed. Although the BS support was high when applying

maximum constraint that required all genomic regions to be present (5/5GR, 2% missing data), the number of taxa was reduced drastically from 136 to 15 and only 5 of the 13 major nodes were represented (Figs. 2.3c and 2.7d). This drastic reduction in taxon density was accompanied by elimination of whole lineages including those that are difficult to place, which becomes an impediment to phylogenetic and taxonomic assessments.

In this sense, the inclusion in multi-loci phylogenetic studies of taxa with biological significance (e.g. carnivorous or C_4 photosynthesis) or suspected in prior studies to occupy crucial position in the tree (e.g. early diverging or potentially useful in breaking long branches) overweighs the issue of missing data. Our findings underscore the validity of sampling strategies that include incomplete taxa even if it leads to datasets punctuated with missing genomic regions. In the missing data-strict 5/5GR data set, a biologically important lineage like the carnivorous had to be excluded. This lineage was correctly placed in the alternative analyses based on data sets where more relaxed criteria were placed on missing data. Similar argument can also be made for effort to achieve broad taxon representation of species-rich biological lineages.

Conclusion

Our results have demonstrated that inclusion of partial genomic regions and/or incomplete taxa can improve phylogenetic structure in the Caryophyllales. This is in line with previous simulation and empirical studies which have suggested that the inclusion of additional taxa and genomic regions regardless of missing data can improve phylogenetic accuracy and rescue an analysis from long-branch attraction (Graybeal 1998; Wiens 1998, 2003a, b, 2005; Burleigh et al. 2009). Such flexibility definitely allows for denser taxon sampling at lower cost and effort using sequences from data repositories despite such shortcomings as missing data. Further, we have found here that the use of a few genomic regions with different mode and tempo of evolution can improve phylogenetic reconstruction. This is particularly important as concerted effort is being made to reconstruct the tree of life with greater taxonomic detail. We acknowledge however, that there is probably a reasonable threshold for the amount and type of missing data that an

analysis can accommodate, and our positive results are an indication that we have stayed within the bounds of that threshold here. It would be interesting to combine whole genome sequence data for some taxa with data for a few selected genomic regions of many more taxa to test the limits of this threshold for missing data in an extreme case.

Acknowledgments

The authors thank J. Gordon Burleigh for his contributions to this manuscript; D. and P. Soltis, S. Brockington, and M. Moore, as well as the Missouri Botanical Garden and the Royal Botanic Garden at Kew for providing DNA samples for several taxa. We thank M. Barthet for help in designing a primer, A. Hinckle for helping with specimen collection, S. Newman for assistance in lab work, and A. Ferraioli for assistance with figures. We also thank two anonymous reviewers for their comments and suggestions. This work is part of the AToL-Angiosperm project supported by grants from the National Science Foundation, USA - EF-043105 and REU-477683 3 to KWH.

References:

Agnarsson, I., May-Collado, L.J., 2008. The phylogeny of Cetartiodactyla: The importance of dense taxon sampling, missing data, and the remarkable promise of cytochrome b to provide reliable species-level phylogenies. Molecular Phylogenetics and Evolution 48, 964-985.

Albert, V.A., Williams, S.E., Chase, M.W., 1992. Carnivorous Plants: Phylogeny and Structural Evolution. Science 257, 1491-1495.

Alverson, W.S., Whitlock, B.A., Nyffeler, R., Bayer, C., Baum, D.A., 1999. Phylogeny of the core Malvales: evidence from *ndhF* sequence data. American Journal of Botany 86, 1474-1486.

APG II, 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Botanical Journal of the Linnean Society 141, 399-436.

APG III, 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Botanical Journal of the Linnean Society 161, 105-121.

Applequist, W.L., Wallace, R.S., 2001. Phylogeny of the Portulacaceous Cohort Based on *ndhF* Sequence Data. Systematic Botany 26, 406-419.

Barthet, M.M., Hilu, K.W., 2007. Expression of *matK*: Functional and Evolutionary Implications. American Journal of Botany 94, 1402-1412.

Behnke, H.-D., 1994. Sieve-Element Plastids: Their Significance for the Evolution and Systematics of the Order. In: Behnke, H.-D., Mabry, T.J. (Eds.), Caryophyllales: Evolution and Systematics. Springer Verlag, Berlin, Germany, pp. 87-121.

Bittrich, V., 1993. Introduction to Centrospermae. In: Kubitzki, K., Rohwer, J.G., Bittrich, V. (Eds.), The families and genera of vascular plants, vol. II, Magnoliid, hamamelid, and caryophyllid families. Springer Verlag, Berlin, Germany, pp. 13-19.

Borsch, T., Hilu, K. W., Quandt, D., Wilde, V., Neinhuis, C. Barthlott, W., 2003. Noncoding plastid *trnT-trnF* sequences reveal a well resolved phylogeny of basal angiosperms. Journal of Evolutionary Biology 16, 558-576.

Brockington, S.F., Alexandre, R., Ramdial, J., Moore, M.J., Crawley, S., Dhingra, A., Hilu, K., Soltis, D.E., Soltis, P.S., 2009. Phylogeny of the Caryophyllales Sensu Lato: Revisiting Hypotheses on Pollination Biology and Perianth Differentiation in the Core Caryophyllales. International Journal of Plant Sciences 170, 627-643.

Burleigh, J.G., Hilu, K.W., Soltis, D.E., 2009. Inferring Phylogenies with Incomplete Data Sets: A 5-Gene, 567-Taxon Analysis of Angiosperms. BMC Evolutionary Biology 9, 61.

Cameron, K.M., Wurdack, K.J., Jobson, R.W., 2002. Molecular Evidence for the Common Origin of Snap-Traps Among Carnivorous Plants. American Journal of Botany 89, 1503-1509.

Chase, M.W., Soltis, D.E., Olmstead, R.G., Morgan, D., Les, D.H., Mishler, B.D., Duvall, M.R., Price, R.A., Hills, H.G., Qiu, Y.-L., Kron, K.A., Rettig, J.H., Conti, E., Palmer, J.D., Manhart, J.R., Systma, K.J., Michaels, H.J., Kress, W.J., Karol, K.G., Clark, W.D., Hedren, M., Gaut, B.S., Jansen, R.K., Kim, K.-J., Wimpee,

C.F., Smith, J.F., Furnier, G.R., Strauss, S.H., Xiang, Q.-Y., Plunkett, G.M., Soltis, P.S., Swensen, S.M., Williams, S.E., Gadek, P.A., Quinn, C.J., Eguiarte, L.E., Golenberg, E., Learn Jr., G.H., Graham, S.W., Barrett, S.C.H., Dayanandan, S., Albert, V.A., 1993. Phylogenetics of Seed Plants: An Analysis of Nucleotide Sequences from the Plasitd Gene *rbcL*. Annals of the Missouri Botanical Garden 80, 528-580.

Clark, L.G., Zhang, W., Wendel, J.F., 1995. A Phylogeny of the Grass Family (Poaceae) Based on *ndhF* Sequence Data. Systematic Botany 20, 436-460.

Cuénoud, P., Savolainen, V., Chatrou, L.W., Powell, M.P., Grayer, R.J., Chase, M.W., 2002. Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL*, *atpB*, and *matK* DNA sequences. American Journal of Botany 89, 132-144.

Donoghue, M.J., Doyle, J.A., Gauthier, J., Kluge, A.G., Rowe, T., 1989. The Importance of Fossils in Phylogeny Reconstruction. Annual Review of Ecology and Systematics 20, 431-460.

Downie, S.R., Katz-Downie, D.S., Cho, K.-J., 1997. Relationships in the Caryophyllales as suggested by phylogenetic analyses of partial chloroplast DNA ORF2280 homolog sequences. American Journal of Botany 84, 253-273.

Downie, S.R., Palmer, J.D., 1994. Phylogenetic Relationships Using Restriction Site Variation of the Chloroplast DNA Inverted Repeat. In: Behnke, H.-D., Mabry, T.J. (Eds.), Caryophyllales: Evolution and Systematics. Springer Verlag, Berlin, Germany, pp. 223-233.

Doyle, J.J., Doyle, J.L., 1990. Isolation of plant DNA from fresh tissue. Focus 12, 13-25.

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

Edwards, E.J., Nyffeler, R., Donoghue, M.J., 2005. Basal Cactus Phylogeny: Implications of Pereskia (Cactaceae) Paraphyly for the Transition to the Cactus Life Form. American Journal of Botany 92, 1177-1188.

Farris, J.S., 1989. The Retention Index and the Rescaled Consistency Index. Cladistics 5, 417-419.

Felsenstein, J., 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution 39, 783-791.

Felsenstein, J., 2004. Inferring phylogenies. Sinauer Associates pg 344.

Fior, S., Karis, P.O., Casazza, G., Minuto, L., Sala, F., 2006. Molecular phylogeny of the Caryophyllaceae (Caryophyllales) inferred from chloroplast *matK* and nuclear rDNA ITS sequences. American Journal of Botany 93, 399-411.

Fior, S., Karis, P.O., 2007. Phylogeny, evolution and systematics of Moehringia (Caryophyllaceae) as inferred from molecular and morphological data: a case of homology reassessment. Cladistics 23, 362-372.

Freudenstein, J.V., Davis, J.I., 2010. Branch support via resampling; an empirical study. Cladistics 26, 643-656.

Gao, K., Norell, M.A., 1998. Taxonomic Revision of Carusia (Reptilia: Squamata) from the Late Cretaceous of the Gobi Desert and Phylogenetic Relationships of Anguimorphan Lizards. American Museum Novitates 3230, 1-52.

Gauthier, J., 1986. Saurischian Monophyly and the Origin of Birds. Memoirs of the California Academy of Sciences Number 8, 1-56.

GenBank (2009). (http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html).

Giannasi, D.E., Zurawski, G., Learn, G., Clegg, M.T., 1992. Evolutionary Relationships of the Caryophyllidae Based on Comparative *rbcL* Sequences. Systematic Botany 17, 1-15.

Graybeal, A., 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Systematic Biology 47, 9-17.

Hillis, D.M., 1996. Inferring complex phylogenies. Nature 383, 130-131.

Hillis, D.M., Pollock, D.D., McGuire, J.A., Zwickl, D.J., 2003. Is Sparse Taxon Sampling a Problem for Phylogenetic Inference? Systematic Biology 52, 124-126.

Hilu, K.W., Alice, L.A., 1999. Evolutionary Implications of *matK* Indels in Poaceae. American Journal of Botany 86, 1735-1741.

Hilu, K.W., Borsch, T., Müller, K., Soltis, D.E., Soltis, P.S., Savolainen, V., Chase, M.W., Powell, M.P., Alice, L.A., Evans, R., Sauquet, H., Neinhuis, C., Slotta, T.A.B., Jens, G.R., Campbell, C.S., Chatrou, L.W., 2003. Angiosperm phylogeny

based on *matK* sequence information. American Journal of Botany 90, 1758-1776.

Hoot, S.B., Culham, A., Crane, P.R., 1995. The Utility of *atpB* Gene Sequences in Resolving Phylogenetic Relationships: Comparison with *rbcL* and 18S Ribosomal DNA Sequences in the Lardizabalaceae. Annals of the Missouri Botanical Garden 82, 194-207.

Huelsenbeck, J.P., 1991. When are Fossils better than Extant Taxa in Phylogenetic Analysis? Systematic Zoology 40, 458-469.

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., dePamphilis, C.W., Leebens-Mack, J., Müller, K.F., Guisinger-Bellian, M., Haberle, R.C., Hansen, A.K., Chumley, T.W., Lee, S.-B., Peery, R., McNeal, J.R., Kuehl, J.V., Boore, J.L., 2007. Analysis of 81 Genes from 64 Plastid Genomes Resolves Relationships in Angiosperms and Identifies Genome-Scale Evolutionary Patterns. PNAS 104, 19369-19374.

Jansen, R.K., Saski, C., Lee, S.-B., Hansen, A.K., Daniell, H., 2011. Complete Plastid Genome Sequences of Three Rosids (*Castanea*, *Prunus*, *Theobroma*): Evidence for At Least Two Independent Transfers of rpl22 to the Nucleus. Mol Biol Evol 28, 835-847.

Johnson, L.A., Soltis, D.E., 1995. Phylogenetic inference in Saxifragaceae sensu stricto and Gilia (Polemoniaceae) using *matK* sequences. Annals of the Missouri Botanical Gardens 82, 149-175.

Judd, W.S., Campbell, C.S., Kellogg, E.A., Stevens, P.F., Donoghue, M.J., 2008. Plant Systematics: A Phylogenetic Approach. Sinauer Associates, Inc., Sunderland, MA 01375 USA.

Kadereit, G., Borsch, T., Weising, K., Freitag, H., 2003. Phylogeny of Amaranthaceae and Chenopodiaceae and the evolution of C4 photosynthesis. International Journal of Plant Sciences 164, 959-986.

Källersjö, M., Farris, J.S., Chase, M.W., Bremer, B., Fay, M.F., Humphries, C.J., Petersen, G., Seberg, O., Bremer, K., 1998. Simultaneous parsimony jackknife analysis of 2538 *rbcL* DNA sequences reveals support for major clades of green

plants, land plants, seed plants, and flowering plants. Plant Systematics and Evolution 213, 259-287.

Källersjö, M., Albert, V.A., Farris, J.S., 1999. Homoplasy Increases Phylogenetic Structure. Cladistics 15, 91-93.

Kawahara, A.Y., Mignault, A.A., Regier, J.C., Kitching, I.J., Mitter, C., 2009. Phylogeny and Biogeography of Hawkmoths (Lepidoptera: Sphingiae): Evidence from Five Nuclear Genes. PLoS One 4, 1-11.

Kearney, M., 2002. Fragmentary Taxa, Missing Data, and Ambiguity: Mistaken Assumptions and Conclusions. Systematic Biology 51, 369-381.

Kearney, M., Clark, J.M., 2003. Problems Due to Missing Data in Phylogenetic Analyses Including Fossils: A Critical Review. Journal of Vertebrate Paleontology 23, 263-274.

Kelchner, S. A., 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. Annals of the Missouri Botanical Garden 87, 482-498.

Kubitzki, K., Rohwer, J. G., & Bittrich, V. 1993 (eds), The Families and Genera of Vascular Plants. II. Flowering Plants: Dicotyledons, Magnoliid, Hamamelid and Caryophyllid Families. Springer, Berlin.

Leebens-Mack, J., Raubeson, L.A., Cui, L., Kuehl, J.V., Fourcade, M.H., Chumley, T.W., Boore, J.L., Jansen, R.K., dePamphilis, C.W., 2005. Identifying the Basal Angiosperm Node in Chloroplast Genome Phylogenies: Sampling One's Way Out of the Felsenstein Zone. Molecular Biology and Evolution 22, 1948-1963.

Li, J., 2008. Phylogeny of Catalpa (Bignoniaceae) inferred from sequences of chloroplast *ndhF* and nuclear ribosomal DNA. Journal of Systematics and Evolution 46, 341-348.

Liang, H., Hilu, K.W., 1996. Application of the *matK* gene sequences to grass systematics. Canadian Journal of Botany 74, 125-134.

McMahon, M.M., Sanderson, M.J., 2006. Phylogenetic Supermatrix Analysis of GenBank Sequences from 2228 Papilionoid Legumes. Systematic Biology 55, 818-836.

Meimberg, H., Wistuba, A., Dittrich, P., Heubl, G., 2001. Molecular Phylogeny of Nepenthaceae Based on Cladistic Analysis of Plastid *trnK* Intron Sequence Data. Plant Biology 3, 164-175.

Moore, M.J., Bell, C.D., Soltis, P.S., Soltis, D.E., 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci U S A 104, 19363-19368.

Müller, J., Müller, K., 2003. QuickAlign: A New Alignment Editor. Plant Molecular Biology Reporter 21, 5.

Müller, K., 2004. PRAP-computation of Bremer support for large data sets. Molecular Phylogenetics and Evolution 31, 780-782.

Müller, K.F., Borsch, T., 2005. Phylogenetics of Amaranthaceae Based on *matK*/*trnK* Sequence Data-Evidence from Parsimony, Likelihood, and Bayesian Analyses. Annals of the Missouri Botanical Gardens 92, 66-102.

Müller, K.F., Borsch, T., Hilu, K.W., 2006. Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: contrasting *matK*, *trnT-F* and *rbcL* in basal angiosperms. Molecular Phylogenetics and Evolution 41, 99-117.

Nixon, K.C., 1999. The Parsimony Ratchet, a New Method for Rapid Parsimony Analysis. Cladistics 15, 407-414.

Novacek, M.J., 1992. Fossils, Topologies, Missing Data, and the Higher Level Phylogeny of Eutherian Mammals. Systematic Biology 41, 58-73.

Nyffeler, R., 2002. Phylogenetic relationships in the cactus family (Cactaceae) based on evidence from *trnK*/*matK* and *trnL-trnF* sequences. American Journal of Botany 89, 312-326.

Nyffeler, R., 2007. The Closest Relatives of Cacti: Insights from Phylogenetic Analyses of Chloroplast and Mitochondrial Sequences with Special Emphasis on Relationships in the Tribe Anacampseroteae. American Journal of Botany 94, 89-101.

Olmstead, R.G., Michaels, H.J., Scott, K.M., Palmer, J.D., 1992. Monophyly of the Asteridae and Identification of Their Major Lineages Inferred From DNA Sequences of *rbcL*. Annals of the Missouri Botanical Garden 79, 249-265.

Olmstead, R.G., Zjhra, M.L., Lohmann, L.G., Grose, S.O., Eckert, A.J., 2009. A Molecular Phylogeny and Classification of Bignoniaceae. American Journal of Botany 96, 1731-1743.

O'Quinn, R., Hufford, L., 2005. Molecular Systematics of Montieae (Portulacaceae): Implications for Taxonomy, Biogeography and Ecology. Systematic Botany 30, 314-331.

Philippe, H., Snell, E.A., Bapteste, E., Lopez, P., Holland, P.W.H., Casane, D., 2004. Phylogenomics of Eukaryotes: Impact of Missing Data on Large Alignments. Molecular Biology and Evolution 21, 1740-1752.

Pirie, M.D., Humphreys, A.M., Galley, C., Barker, N.P., Verboom, G.A., Orlovich, D., Draffin, S.J., Lloyd, K., Baeza, C.M., Negritto, M., Ruiz, E., Sanchez, J.H.C., Reimer, E., Linder, H.P., 2008. A novel supermatrix approach improves resolution of phylogenetic relationships in a comprehensive sample of danthonioid grasses. Molecular Phylogenetics and Evolution 48, 1106-1119.

Pollock, D.D., Zwickl, D.J., McGuire, J.A., Hillis, D.M., 2002. Increased Taxon Sampling Is Advantageous for Phylogenetic Inference. Systematic Biology 51, 664-671.

Pryer, K.M., Schuettpelz, E., Wolf, P.G., Schneider, H., Smith, A.R., Cranfill, R., 2004. Phylogeny and Evolution of Ferns (Monilophytes) with a Focus on the Early Leptosporangiate Divergences. American Journal of Botany 91, 1582-1598.

Qiu, Y.-L., Dombrovska, O., Lee, J., Li, L., Whitlock, B.A., Bernasconi-Quadroni, F., Rest, J.S., Davis, C.C., Borsch, T., Hilu, K.W., Renner, S.S., Soltis, D.E., Soltis, P.S., Zanis, M.J., Cannone, J.J., Gutell, R.R., Powell, M., Savolainen, V., Chatrou, L.W., Chase, M.W., 2005. Phylogenetic Analyses of Basal Angiosperms Based on Nine Plastid, Mitochondrial, and Nuclear Genes. International Journal of Plant Sciences 166, 815-842.

Qiu, Y.-L., Li, L., Wang, B., Chen, Z., Knopp, V., Groth-Malonek, M., Dombrovska, O., Lee, J., Kent, L., Rest, J., Estabrook, G.F., Hendry, T.A., Taylor, D.W., Testa, C.M., Ambros, M., Crandall-Stotler, B., Duff, R.J., Stech, M., Frey, W., Quandt, D., Davis, C.C., 2006. The deepest divergences in land plants inferred from phylogenomic evidence. PNAS 103, 15511-15516.

Rannala, B., Huelsenbeck, J.P., Yang, Z., Nielsen, R., 1998. Taxon Sampling and the Accuracy of Large Phylogneies. Systematic Biology 47, 702-710.

Rettig, J.H., Wilson, H.D., Manhart, J.R., 1992. Phylogeny of the Caryophyllales - gene sequence data. Taxon 41, 201-209.

Rokas, A., Carroll, S.B., 2005. More Genes or More Taxa? The Relative Contribution of Gene Number and Taxon Number to Phylogenetic Accuracy. Mol Biol Evol 22, 1337-1344.

Rønsted, N., Weiblen, G.D., Clement, W.L., Zerega, N.J.C., Savolainen, V., 2008. Reconstructing the phylogeny of figs (Ficus, Moraceae) to reveal the history of the fig pollination mutualism. Symbiosis 45, 1-12.

Rosenberg, M.S., Kumar, S., 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. PNAS 98, 10751-10756.

Sanchez, A., Kron, K.A., 2008. Phylogenetics of Polygonaceae with an Emphasis on the Evolution of Eriogonoideae. Systematic Botany 33, 87-96.

Savolainen, V., Chase, M.W., Hoot, S.B., Morton, C.M., Soltis, D.E., Bayer, C., Fay, M.F., DeBruijn, A.Y., Sullivan, S., Qiu, Y.-L., 2000. Phylogenetics of Flowering Plants Based on Combined Analysis of Plastid *atpB* and *rbcL* Gene Sequences. Systematic Biology 49, 306-362.

Smissen, R.D., Clement, J.C., Garnock-Jones, P.J., Chambers, G.K., 2002. Subfamilial relationships within Caryophyllaceae as inferred from 5' *ndhF* sequences. American Journal of Botany 89, 1336-1341.

Smith, J.F., Wolfram, J.C., Brown, K.D., Carroll, C.L., Denton, D.S., 1997. Tribal Relationships in the Gesneriaceae: Evidence from DNA Sequences of the Chloroplast Gene *ndhF*. Annals of the Missouri Botanical Garden 84, 50-66.

Soltis, D.E., Soltis, P.S., Chase, M.W., Mort, M.E., Albach, D.C., Zanis, M., Savolainen, V., Hahn, W.H., Hoot, S.B., Fay, M.F., Axtell, M., Swensen, S.M., Prince, L.M., Kress, W.J., Nixon, K.C., Farris, J.S., 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. Botanical Journal of the Linnean Society 133, 381-461.

Soltis, D.E., Senters, A.E., Zanis, M.J., Kim, S., Thompson, J.D., Soltis, P.S., Ronse De Craene, L.P., Endress, P.K., Farris, J.S., 2003. Gunnerales are sister to other core

eudicots: implications for the evolution of pentamery. American Journal of Botany 90, 461-470.

Soltis, D.E., Albert, V.A., Savolainen, V., Hilu, K., Qiu, Y.-L., Chase, M.W., Farris, J.S., Stefanovic, S., 2004. Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. TRENDS in Plant Science 9, 477-483.

Stamatakis, A., Hoover, P., Rougemont, J., 2008. A Fast Bootstrapping Algorithm for the RAxML Web Servers. Systematic Biology 57, 758-771.

Stevens, P. F. (2010). Angiosperm Phylogeny Website. Version 9, June 2008. http://www.mobot.org/MOBOT/research/APweb/.

Swofford, D.L., 2003. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4, Sinauer Associates, Sunderland, Massachusetts, USA.

Turmel, M., Gagnon, M.-C., O'Kelly, C.J., Otis, C., Lemieux, C., 2009. The Chloroplast Genomes of the Green Algae *Pyramimonas*, *Monomastix*, and *Pycnococcus* shed new light on the evolutionary history of Prasinophytes and the origin of the secondary chloroplasts of Euglenids. Mol Biol Evol 26, 632-648.

Wang, H., Moore, M.J., Soltis, P.S., Bell, C., Brockington, S.F., Alexandre, R., Davis, C.C., Latvis, M., Manchester, S.R., Soltis, D.E., 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc Natl Acad Sci U S A 106, 3853-3858.

Whittall, J.B., Carlosn, M.L., Beardsley, P.M., Meinke, R.J., Liston, A., 2006. The Mimulus moschatus Alliance (Phrymaceae): Molecular and Morphological Phylogenetics and their Conservation Implications. Systematic Botany 31, 380-397.

Wiens, J.J., 1998. Does Adding Characters with Missing Data Increase or Decrease Phylogenetic Accuracy? Systematic Biology 47, 625-640.

Wiens, J.J., 2003a. Incomplete Taxa, Incomplete Characters, and Phylogenetic Accuracy: Is There a Missing Data Problem? Journal of Vertebrate Paleontology 23, 297-310.

Wiens, J.J., 2003b. Missing Data, Incomplete Taxa, and Phylogenetic Accuracy. Systematic Biology 52, 528-538.

Wiens, J.J., 2005. Can Incomplete Taxa Rescue Phylogenetic Analyses from Long-Branch Attraction? Systematic Biology 54, 731-742.

Wiens, J.J., 2006. Missing data and the design of phylogenetic analyses. Journal of Biomedical Informatics 39, 34-42.

Wiens, J.J., Reeder, T.W., 1995. Combining Data Sets with Different Numbers of Taxa for Phylogenetic Analysis. Systematic Biology 44, 548-558.

Wilkinson, M., 1995. Coping with Abundant Missing Entries in Phylogenetic Inference Using Parsimony. Systematic Biology 44, 501-514.

Williams, S.E., Albert, V.A., Chase, M.W., 1994. Relationships of Droseraceae: A Cladistic Analysis of *rbcL* Sequence and Morphological Data. American Journal of Botany 81, 1027-1037.

Wilson, C.A., 2009. Phylogenetic Relationships Among the Recognized Series in Iris Section Limniris. Systematic Botany 34, 277-284.

Wolf, P.G., 1997. Evaluation of *atpB* nucleotide sequences for phylogenetic studies of ferns and other pteridophytes. American Journal of Botany 84, 1429-1440.

Chapter 3: Caryophyllales: Evaluating phylogenetic signal in *matK* and *trnK* intron

Abstract

We compared phylogenetic signal from the plastid *trnK* intron with rapidly evolving *matK* in reconstructing phylogeny at the deep historic level of the order Caryophyllales. The *matK* gene, a putative group II intron maturase, is nested within an intron of *trnK*, the tRNA gene encoding Lysine$^{(UUU)}$. The two genomic regions are often co-amplified and co-sequenced making *trnK* intron an appealing source of genetic information. They appeared comparable in terms of proportion of variable sites, displayed a similar pattern of substitution rates per site, and overall phylogentic informativeness. Maximum Parsimony, Maxmimum Likelihood, and Bayesian analyses showed strong congruence between the phylogenetic trees based on *matK* and *trnK* intron sequences from 45 genera representing 30 of the 34 recognized Caryophyllales families. The *trnK* intron alone provided a relatively well-resolved topology for the order. Combining the *trnK* intron with the *matK* sequence data resulted in 6 most parsimonious trees, differing only in the placement of *Claytonia* (Montiaceae) within the raphid clade. A major basal split in the order into core and non-core Caryophyllales was evident with very strong support. Both partitioned and combined data analyses provided the highest support yet for the single origin of carnivory in the Caryophyllales. This study promotes the use of the *trnK* intron (both substitutions and insertions/deletions) at deeper level phylogenies, particularly with its convenient sequencing since the intron is concurrently amplified in most *matK*-based studies.

Introduction

Although slowly evolving genes have traditionally been used for inferring phylogenies at higher taxonomic level, recent studies have demonstrated the effectiveness of rapidly evolving genomic regions such as *matK* at deep historic levels (see Hilu et al. 2003). For example, the use of partial sequences from *matK* provided a well-resolved phylogeny for angiosperms that was comparable in robustness to 3-11 genes + plastid IR combined (Graham and Olmstead 2000; Zanis et al. 2002; Hilu et al. 2003; Brockington, et al., 2009). The *matK* gene is nested within the *trnK* intron encoding the tRNA for Lysine$^{(UUU)}$, splitting the intron into 5' and 3' regions. We will collectively refer to these 5' and 3' regions as the *trnK* intron to distinguish them from the *matK* open reading frame (ORF). Most studies have used *trnK* intron sequences to address systematic questions below the family level (Hu et al. 2000; Young and dePamphilis 2000; Edwards and Gadek 2001; Lavin et al. 2001; Wilson 2004; Ronsted et al. 2005). The utility of *trnK* intron at deep historic levels has not been well explored despite its concurrent amplification in most *matK*-based studies. A recent assessment of phylogenetic signal in the *trnK* intron shows that it evolves at the same rate as *matK* (Hilu et al., 2008). Phylogenetic signal from the *trnK* intron as well as *matK* should not be confined to substitutions alone, since both regions are rich in insertions and deletions (indels) that could potentially contribute to phylogenetic informativeness.

The Caryophyllales are one of the major lineages in flowering plants that include various life forms, such as carnivorous and desert plants. Traditionally, Caryophyllales was circumscribed by synapomorphies such as P3 type of sieve element plastid, curved embryo, free central placentation (sometimes basal), and perisperm (Bittrich 1993; Behnke 1994). In addition, betalain pigments represent another synapomorphy for the group, except for Caryophyllaceae and Molluginaceae (Clement and Mabry 1996). The Caryophyllales as envisioned in Cronquist and Thorne (1994), was comprised of 15 families according to Thorne, but reduced to 12 in Cronquist's view. However, studies based on molecular information over the past 14 years have considerably expanded our

knowledge of the phylogeny of the Caryophyllales and its composition. Several new families have been added including the Droseraceae, Nepenthaceae, Plumbaginaceae, Polygonaceae, Asteropeiaceae, Frankeniaceae, Rhabdodendraceae, Simmondsiaceae, Limeaceae, Montiaceae, and Tamaricaceae (Albert, Williams, and Chase 1992; Williams, Albert, and Chase 1994; Fay et al. 1997; Morton, Karol, and Chase 1997; Källersjö et al. 1998; APG II 2003; APG III 2009). Currently the APG III (2009) recognizes 34 families in the order. Two major groups have been defined, core and non-core Caryophyllales with several well defined subclades (Albert, Williams, and Chase 1992; Nandi, Chase, and Endress 1998; Brockington et al., 2009).

Although the monophyly of the Caryophyllales in its current broad circumscription is strongly supported in most studies, the relationship among some of the families within the core and non-core Caryophyllales has been historically disputed, due to incongruence and/or low resolution and support among phylogenies. The phylogenetic position of families such as Caryophyllaceae, Stegnospermataceae, and Molluginaceae has been inconsistently resolved with different gene combinations (Savolainen et al. 2000; Cuénoud et al. 2002; Hilu et al. 2003; Soltis et al. 2003; Brockington et al., 2009). Further, the uncertain position of families such as Rhabdodendraceae and Simmondsiaceae is problematic, affecting the backbone of the Caryophyllales phylogeny as they appear sister to remaining Caryophyllales or deeper in the tree. In one of two molecular systematic studies focusing on the phylogenetic structure of the Caryophyllales, Cuénoud et al. (2002) used sequence information from 18S rDNA, *rbcL*, *atpB*, and part of *matK*, but concluded that we are far from understanding of the phylogeny of the group. However, phylogenetic signal in *matK* is distributed along the entire ORF and, thus, the exclusion of sectors of *matK* can result in reducing the amount of phylogenetic information. Brockington et al. (2009) however, used sequences from nine plastid genes (*atpB*, *matK*, *ndhF*, *psbB*, *psbT*, *pbsN*, *rbcL*, *rpoC2*, and *rps4*), two nuclear genes (18S and 26S), and the entire plastid inverted repeat (IR) from 36 taxa, and the resulting phylogeny provided a strongly supported backbone for most of the major clades of Caryophyllales.

The existing well-established phylogenies for the Caryophyllales provides a suitable platform to contrast phylogenetic signal and tree robustness from the rapidly evolving *trnK* intron and *matK* ORF at the ordinal level (Cuénoud et al., 2002; Brockington et al., 2009). We constructed two completely overlapping datasets for *trnK* intron and *matK* from 45 species of the Caryophyllales to assess the phylogenetic utility of substitutions and indels from these two rapidly evolving genomic regions at the ordinal evolutionary history. We chose the Caryophyllales because it 1) has lineages that are historically difficult to place and others with well established phylogenetic position, 2) the focus on resolving its phylogenetic structure has primarily been by using varied number and combinations of mainly slowly evolving genes, and 3) a robust phylogeny for the order has recently been obtained in the Brockington et al. (2009) study.

Materials and Methods

Taxon Sampling and Sources of Material

This study examined 45 caryophyllid species representing 30 of the 34 families recognized by APG III (2009). Six core eudicot species from the families Vitaceae, Dilleniaceae, Berberidopsidaceae, and Santalaceae were used as outgroup (Appendix B). Our choice of outgroup was based on the phylogenetic position of Caryophyllales in core eudicots as depicted in Hilu et al. (2003) and Soltis et al. (2011). Fifty complete or partial sequences were generated for this study; the remaining sequences were obtained from GenBank (Appendix B). Efforts were made to use the same species/accession for completing partial sequences, but in a few cases a placeholder from the same genus was substituted. DNA samples were either extracted from plant material collected in the field and stored at –80°C or obtained from various sources (Appendix B).

DNA Isolation, PCR Amplification, and Sequencing

DNA extraction followed the CTAB method of Doyle and Doyle (1990). To generate completely new sequences for the entire *matK* ORF and the *trnK* intron, the whole region

was amplified in three overlapping sections using a combination of external primers located in the *trnK* exons and internal *matK* primers. Partial sequences from GenBank were used as template to search for and design *matK* internal primers that were used for completing the missing regions and providing a considerable overlap with existing sequences. The previously published primers used in this study were: trnK3914Fdi (Johnson and Soltis 1995), corematK1 (Barthet and Hilu 2007), TOmatK480F (Hilu et al. 2003), MG1 (Liang and Hilu 1996), and CaryomatK291F, CaryomatK467R, Caulo1100F, PImatK1326R, and AsteromatK500R (Crawley and Hilu 2011). The protocols for Polymerase chain reaction (PCR) amplification and sequencing of the genomic regions are detailed in Crawley and Hilu (2011).

Sequence Alignment and Phylogenetic Analysis

All sequences were manually aligned using QuickAlign (Müller and Müller 2003). The insertion of gaps in both *matK* and *trnK* intron sequence alignments followed the rules noted in Kelchner (2000) and Borsch et al. (2003). The *matK* sequences were translated into amino acids to provide further guideline for the position of gaps while maintaining the open reading frame. Frame shift mutations are primarily confined to the very end of the 3' region *(*Hilu and Alice, 1999; Hilu et al., 2003). We further tested our alignment by submitting the original datasets to the CIPRES server (http://www.phylo.org/) to generate an alignment with MUSCLE (Edgar 2004). The *matK* ORF alignment was unambiguous and, thus, the whole alignment was used in the analyses. In the case of the *trnK* intron, a large number of mutational changes across the dataset were evident. Certain sectors, however, represented mutational hotspots, making reliable homology assessments difficult to attain. Consequently, these sectors along with poly-A/T strings (371 characters total) were excluded from the analyses. Three datasets were generated representing *matK* ORF alone, *trnK* intron alone, and combined *matK* ORF and *trnK* intron. Indels in all datasets were coded as binary characters using the "Simmons & Ochoterena (2000) – simple coding" option in SeqState (Müller 2005). The datasets were analyzed phylogenetically with and without indels to test the effect of the inclusion of indels on phylogenetic reconstruction for each genomic region.

The distribution of rates of substitution across sites were assessed for *matk* and *trnK* intron regions by assigning a rate class to each site based on the general reversible model (see model choice below) using the HyPhy program (Kosakovsky Pond, Frost, and Muse 2005). Four rate classes were established as well as a non-variable class following a discrete gamma distribution, and each site was assigned to one of these classes followed by calculation of percentage of sites in each class. Additionally, assessment of rates of substitutions per site across the *matK* and *trnK* intron genomic regions was also examined using HyPhy (Kosakovsky Pond, Frost, and Muse 2005).

Maximum parsimony, maximum likelihood and Bayesian approaches were used to analyze the partitioned and combined datasets with and without indels. Prior to combing the two datasets, a partition homogeneity test (Farris et al., 1995) was performed using PAUP* version 4.0b10 (Swofford, 2003) to test for congruence of the *trnK* intron and *matK* regions. The test was executed with 100 replicates, and $P = 0.07$ indicated that these regions were combinable. Maximum parsimony (MP) was performed using PAUP* version 4.0b10 (Swofford, 2003). Heuristic tree searches used TBR branch swapping with 1000 random addition sequence replicates and a strict consensus tree was generated in each case. Confidence estimates were obtained by performing 1000 bootstrap (BS) (Felsenstein, 1985) replicates each with 10 random sequence addition replicates and TBR branch swapping.

Prior to executing likelihood and Bayesian analyses each dataset was tested for appropriate model choice using the ModelTest3.7 program (Posada and Crandall, 1998). The GTR+I+G model was suggested for all three datasets. Maximum likelihood (ML) analyses were executed using RAxML (Stamatakis et al., 2008). The data were submitted to the CIPRES portal (www.phylo.org) using the default settings for GTR+I+G model. Bootstrapping was also performed in RAxML on the CIPRES portal with 1000 bootstrap replicates and a 50% majority rule consensus tree was generated in PAUP* to obtain BS values. Bayesian analyses were carried out using Mr. Bayes version 3.2 (Ronquist and Huelsenbeck, 2003). For each analysis, four Markov chains were run

simultaneously starting with a random tree, and sampling trees and parameters every 100 generations for 1,000,000 generations at which point the stationary state had been reached. The first 25% of trees were discarded as 'burn in' trees, and a 50% majority rule consensus tree was computed from those remaining.

PhyDesign server (http://phydesign.townsend.yale.edu; Townsend 2007; López-Giráldez and Townsend, 2011) was used to estimate phylogenetic informativeness of *matK* and *trnK* intron regions across the caryophyllids evolutionary history. We generated ultrametric trees using PATHd8, executed the file in MEGA 5.0 (Tamura et al. 2011) and used the trees to overlay the historic changes in substitution rates for the genomic regions. The root of the tree was set at an evolutionary time of 1.0 to obtain relative ages of the clades. The datasets and their corresponding trees were used as input data to generate profiles of phylogenetic informativeness. Both net phylogenetic informativeness and per site informativeness were calculated for the two genomic regions. The latter assessment avoids bias caused by differences in sequence length between the two regions.

Results

Phylogenetic signal in matK/trnK intron

The average rate of substitution per site across *matK* was 1.29 compared with 1.43 and 1.66 for 5' and 3' *trnK* introns, respectively. The two regions of the *trnK* intron combined had an average rate of 1.49. Comparing the rate of substitutions per site across the *matK* ORF to the *trnK* intron using the *t*-test showed that there is not a significant difference between the two regions ($p = 0.19$). The 5' *trnK* intron, however, displayed some sites with elevated rates of substitution (Fig. 3.1). Similarly, the upstream region of *matK* and to a certain degree Domain X (the putative functional domain), appears to experience higher rates of substitution in the Caryophyllales (Fig. 3.1), which is in line with previous findings (Barthet and Hilu, 2008). However, extremely high spikes in the rates are only evident in vary few sites of both regions. Homogeneity in substitution rates

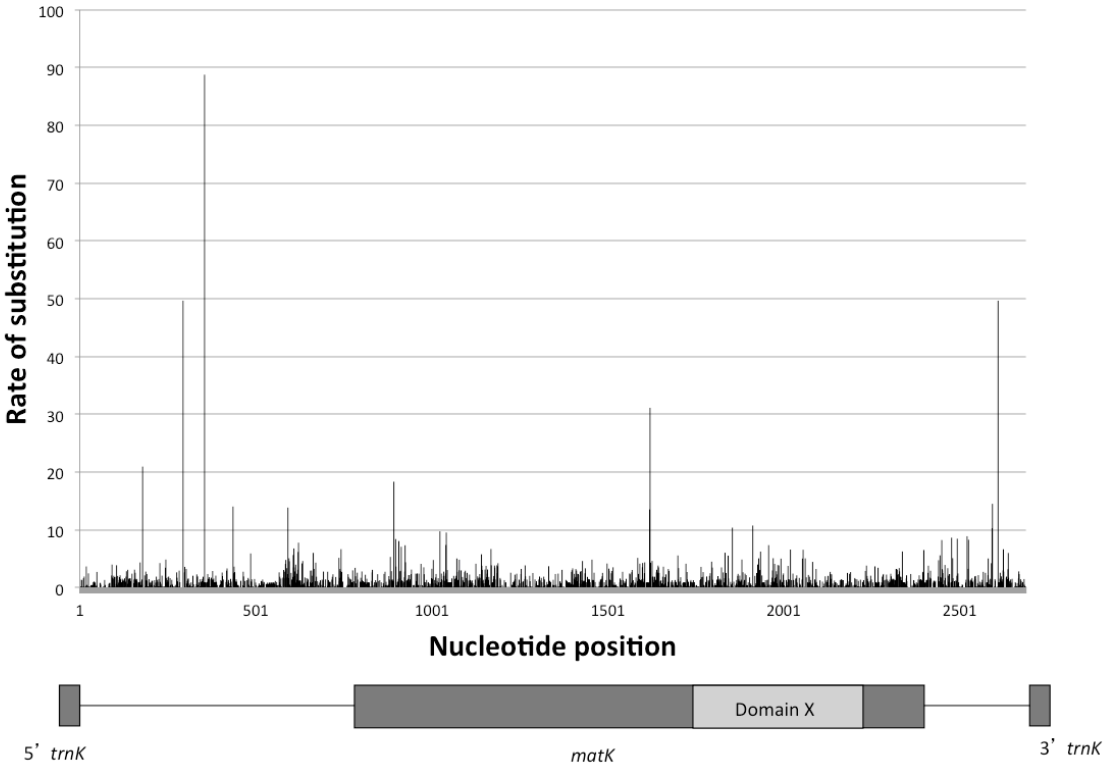

Figure 3.1 Distribution of substitution rates across 5' and 3' *trnK* introns and the *matK* gene as calculated in HyPhy using the GTR model of evolution. A diagram of *matK* and *trnK* is included to indicate the approximate position of the sites along these genomic regions.

across genomic regions is a desirable feature in molecular systematics (Lió and Goldman, 1998).

The rate classes analyses showed that about one third of the sites were invariable (RC0) in both regions (Table 3.1). The RC 1 class, which represents the slowest rate class, comprised the least proportion of variable sites in *matK* and *trnK* intron. The other faster evolving classes (RC 2 – RC 4), although displaying overall similar proportions of sites, show higher numbers of sites for the RC 2 class in *matK* and RC 4 for 3' *trnK* intron (Table 3.1).

The net informativeness profiles of *matK* and *trnK* intron demonstrated that *matK* is more informative than *trnK* intron, although both profiles depict a decline in informativeness deeper in caryophyllids history (Fig. 3.2). This decline was less pronounced in the *trnK* intron profile. Combining sequence information from the two genomic regions enhanced the informativeness considerably (Fig. 3.2). However, when per site informativenes was calculated, the *matK* profile approached that of the *trnK* intron, but the profile based on the combined data appeared intermediate between the two (Fig. 3.2). The mean rate of substitution for *matK* was 0.666 (SD= 11.470) compared with 0.164 (SD= 0.210) for *trnK* intron.

Phylogenetic analyses

The statistical information comparing analyses of data with and without the inclusion of indels as characters are noted in Table 3.2. Since the inclusion of indel characters did not cause topological differences but resulted in an overall increase in tree robustness, we will present the results obtained from datasets that included the indel characters. Nevertheless, statistical results obtained from the comparative analyses of the indel characters are included in the tables

matK ORF partition

Table 3.1 Percentage of sites in each rate class arranged from invariant (RC 0) to fastest (RC 4).

	RC 0	RC 1	RC 2	RC 3	RC 4
matK ORF	32.1	3.9	26.5	17.9	19.5
5' *trnK* Intron	35.8	5.0	21.0	19.2	19.0
3' *trnK* Intron	36.5	3.2	15.6	18.1	26.6
trnK Intron Combined	36.0	4.5	19.5	18.9	21.1
MatK ORF/*trnK* Intron Combined	33.6	4.1	23.8	18.3	20.1

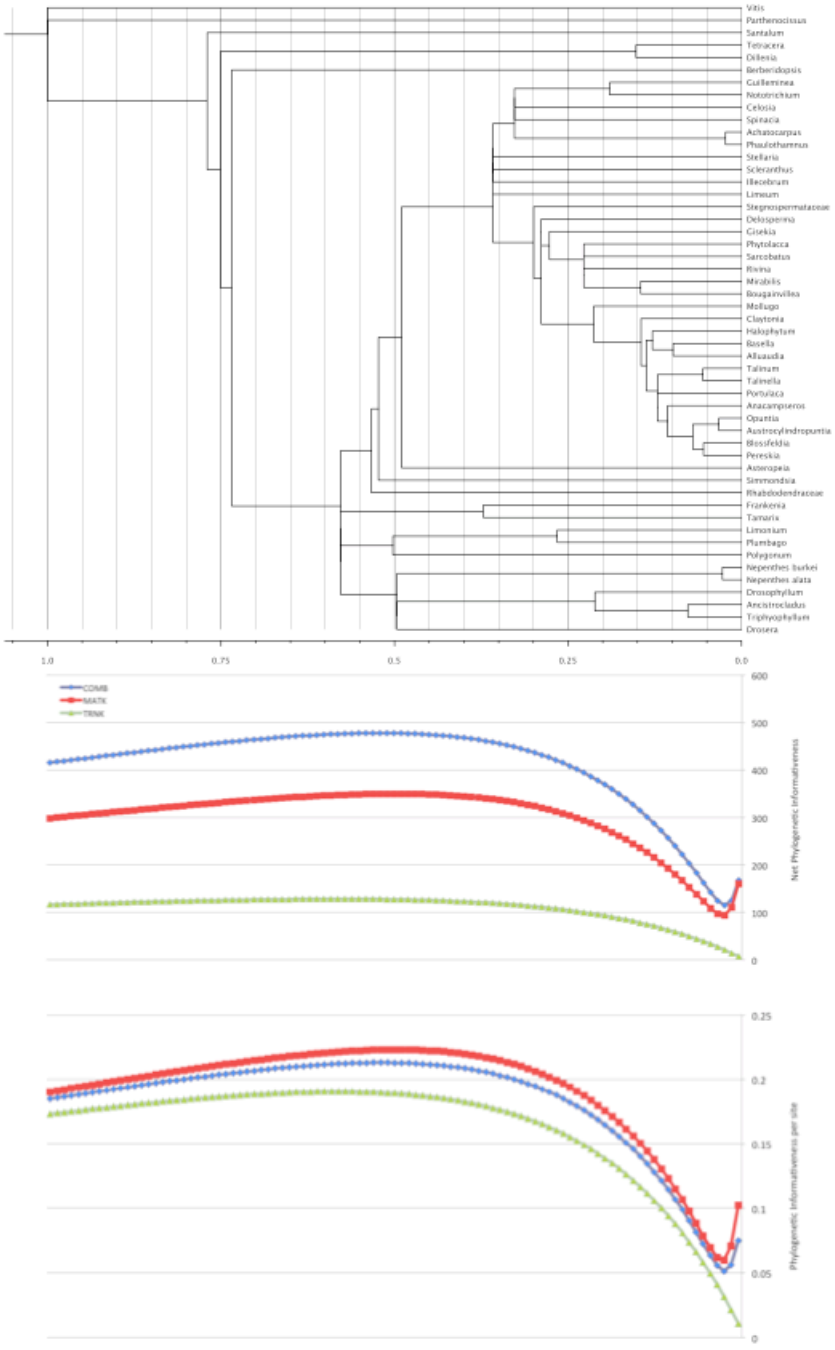

Figure 3.2 Phylogenetic informativeness profiles for *matK* ORF (red), *trnK* intron (green), and combined *matK*/*trnK* intron (blue). An ultrametric tree based on the combined *matK*/*trnK* intron dataset (obtained with RAxML and ultrametrisized with PATHd8) with relative divergence times is shown at top and the profiles of net and per-site phylogenetic informativeness are shown below.

Table 3.2 Contribution of indels to each of the three datasets

	matK ORF	*trnK* Intron	*matK*/*trnK* Combined
Total Number of Indels	61	223	284
Variable Indel Characters	60 (98%)	223 (100%)	283 (99.6%)
Parsimony Informative Indel Characters	21 (35%)	90 (40.4%)	111 (39.2%)
Change in CI[a] due to Indels	+ 0.014	+ 0.024	+ 0.012
Change in Number of trees due to Indels	No Change	- 3567	- 30

[a] CI = Ensemble Consistency Index.

+ indicates an increase and – a decrease in CI and tree number with inclusion of indel characters in the analyses compared to the respective dataset without indel characters.

The length of the *matK* ORF ranged from 1500 bp (*Stellaria media*, Caryophyllaceae) to 1572 bp (*Mollugo verticellata*, Molluginaceae), with the aligned length comprising 1650 characters. Additionally, 61 indel characters were coded for this dataset. Among the total characters, 1179 (69%) were variable and 843 (72%) of the variable characters were potentially parsimony informative (PI). The MP analysis resulted in four most parsimonious trees of 4102 steps in length, with ensemble consistency index (CI) and ensemble retention index (RI) values of 0.480 and 0.564, respectively (Table 3.3). The topological differences among these four trees were confined either to the relationships between *Anacampseros* (Anacampserotaceae) and *Portulaca* (Portulacaceae) or among members of the Cactaceae. The strict consensus tree is shown in figure 3.3b. The monophyly of the Caryophyllales was supported by 100% BS_{MP} (Fig. 3.3b). A basal split into non-core and core Caryophyllales (including Rhabdodendraceae, Asteropieaceae, and Simmondsiaceae) was evident. The non-core Caryophyllales (55% BS_{MP}) was divided into two subclades, one (FTPP clade; 80% BS_{MP}) containing Frankeniaceae + Tamaricaceae (100% BS_{MP}) sister to Plumbaginaceae + Polygonaceae (100% BS_{MP}). The other subclade (carnivorous clade; 95% BS_{MP}) was comprised of Nepenthaceae sister to the two clades of Droseraceae + Drosophyllaceae (<50% BS_{MP}) and Ancistrocladaceae + Dioncophyllaceae (100% BS_{MP}).

The monophyly of core Caryophyllales received 63% BS_{MP} support. Rhabdodendraceae and Simmondsiaceae form a clade (<50% BS_{MP}) sister to remaining core Caryophyllales, excluded by 84% BS_{MP} (Fig. 3.3b). Asteropeiaceae diverged next (excluded by 100% BS_{MP}) followed by Caryophyllaceae + *Limeum* (Limeaceae) (<50% BS_{MP}), and then Stegnospermataceae as sister to the remaining core Caryophyllales. The latter group split into Amaranthaceae + Achatocarpaceae (92% BS_{MP}) and a clade comprised of the remaining families (56% BS_{MP}). This clade consisted of two well-supported lineages, corresponding to the raphide clade (100% BS_{MP}) and the succulents + *Mollugo* clade (84% BS_{MP}) of Brockington et al., (2009); a nomenclature we will adopt here (Fig. 3.3b). Among the families in these clades, the Phytolaccaceae did not appear monophyletic (Fig. 3.3b).

Table 3.3 Maximum Parsimony statistics for the *matK* ORF, *trnK* intron, and combined datasets analyzed with and without indels

	Alignment Length	Characters included in analyses[a]	Variable Characters	Parsimony Informative Characters	Number of Trees	Tree Length	Nodes Resolved	CI[b]	RI[c]
matK ORF	1650	1650	1119 (67.8 %)	882 (73.5 %)	4	4015	42	0.476	0.558
matK ORF plus indels	N/A	1711	1179 (68.9 %)	843 (71.5 %)	4	4102	42	0.480	0.564
trnK Intron	1410	1039	666 (64.1 %)	440 (66.1 %)	3690	2226	30	0.497	0.554
trnK Intron plus indels	N/A	1262	889 (70.4 %)	530 (59.6 %)	123	2553	36	0.521	0.575
matK/*trnK* Combined	3060	2689	1785 (66.4 %)	1262 (70.7 %)	36	6259	37	0.482	0.554
matK/*trnK* Combined plus indels	N/A	2973	2068 (69.6 %)	1373 (66.4 %)	6	6677	42	0.494	0.565

[a] Number of characters used in the analyses after the exclusion of regions that were difficult to align.

[b] CI = Ensemble Consistency Index.

[c] RI = Ensemble Retention Index.

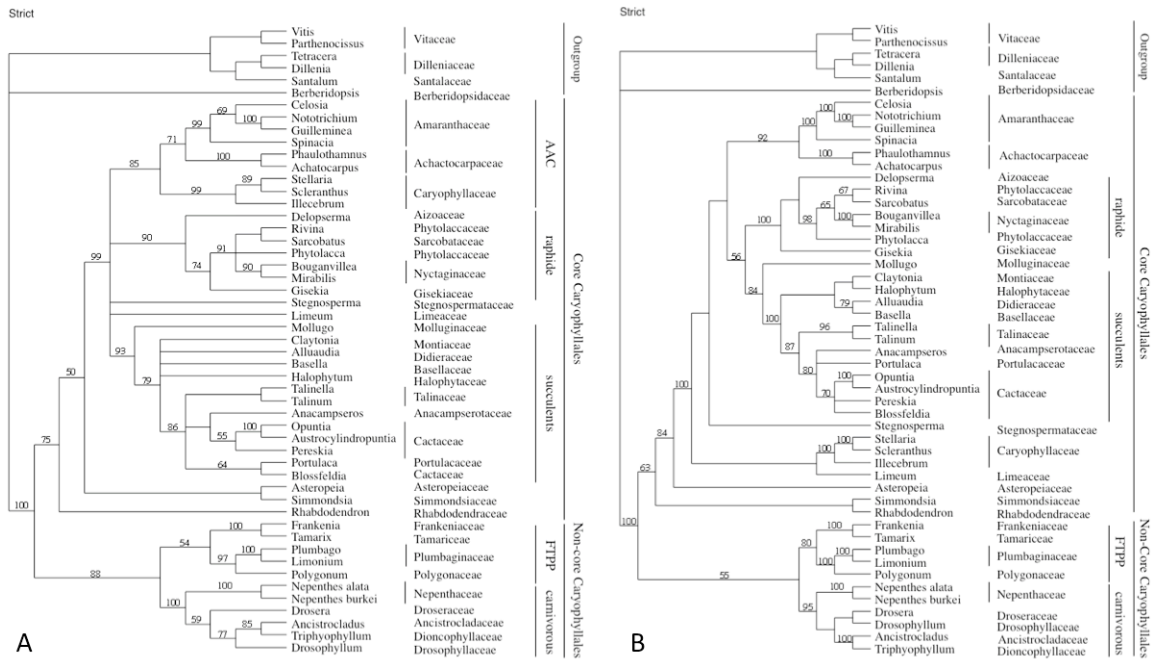

Figure 3.3 Strict consensus tree for the Caryophyllales derived from Maximum Parsimony analysis. Bootstrap values are noted on the nodes. A) Phylogeny based on substitutions and indels from *trnK* intron. B) Phylogeny based on substitutions and indels from *matK* ORF.

The topology of the Bayesian tree was very similar to that of the MP tree (Fig. E.2). Major differences include the position of the Caryophyllaceae, which appeared in the Bayesian analysis as sister to the Achatocarpaceae + Amaranthaceae (0.97 PP) forming the AAC clade, the emergence of *Limeum* (Limeaceae) as sister to the latter clade (0.73 PP), and the appearance of *Rhabdodendron* and *Simmondsia* in a grade instead of a clade (Fig. E.2). However, support for these relationships was very low in both Bayesian and MP analyses. The ML based tree was similar in topology and support to the Bayesian tree but with improved resolution (Fig. 3.4b).

trnK Intron partition

The length of the *trnK* intron ranged from 830 bp (*Stegnosperma halmifolium*, Stegnospermataceae) to 1084 bp (*Scleranthus perennis*, Caryophyllaceae), The insertion of gaps and deletion of mutational hotspots/polyA-T regions resulted in an alignment of 1039 characters. Additionally, 223 indel characters were added to the alignment in the indel coding process. Among the total characters, 889 (70%) were variable and 530 (60%) were potentially parsimony informative. The MP analysis resulted in 123 most parsimonious trees of 2553 steps in length, with CI and RI values of 0.521 and 0.575, respectively (Table 3.3). The strict consensus tree is shown in figure 3.3a. After examining these trees we noted that the differences in topologies were mostly due to the position of *Stegnosperma* (Stegnospermataceae) and *Limeum* (Limeaceae). The exclusion of these two species from the MP analysis reduced the number of trees to only 12, which mainly differed in the placement of *Phytolacca* (Phytolaccaceae), and the relationships among members of the succulents clade.

As in the *matK* analysis, the monophyly of the Caryophyllales was supported by 100% BS_{MP}. A basal split into non-core Caryophyllales and core Caryophyllales (including Rhabdodendraceae, Simmondsiaceae, and Asteropeiaceae) was also evident. Non-core Caryophyllales received higher support (88% BS_{MP}) compared with *matK*. Relationships

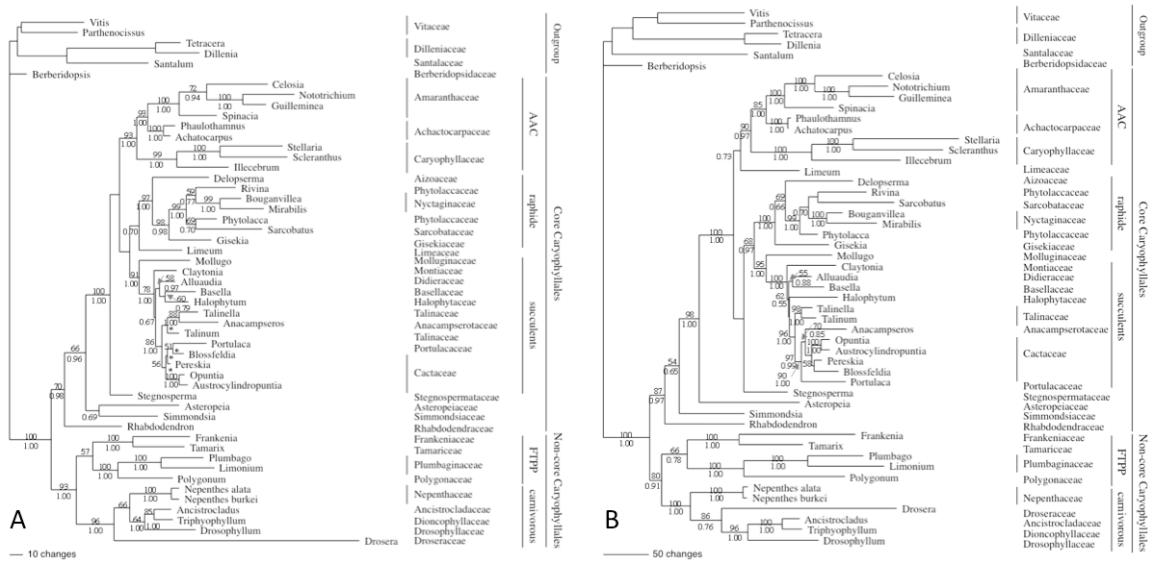

Figure 3.4 The 50% majority rule trees derived from Maximum Likelihood analyses. Bootstrap values are noted on the nodes with Posterior Probability values noted below the nodes. The * denotes nodes that differ between the ML and BI topologies (See Figs. E1 and E2). A) Phylogeny based on substitutions and indels from *trnK* intron. B) Phylogeny based on substitutions and indels from *matK* ORF.

among families were similar to those obtained with *matK*, except for the position of Drosophyllaceae in relation to the Droseraceae (Fig. 3.3).
Monophyly of core Caryophyllales received 75% BS_{MP} support. Rhabdodendraceae emerged first, followed by Simmondsiaceae + Asteropeiaceae (<50% BS_{MP}) and a clade containing the remaining Caryophyllales. The latter clade was supported by 99% BS_{MP} and showed a polytomy of five lineages. These lineages represent: the raphide clade (90% BS_{MP}); the succulents + *Mollugo* clade (93% BS_{MP}); Amaranthaceae + Achatocarpaceae (71% BS_{MP}) sister to Caryophyllaceae (AAC clade; 85% BS_{MP}); Stegnospermataceae; and Limecaceae (Fig. 3.3a). The Cactaceae did not appear monophyletic in the *trnK* intron analysis, and as in the *matK* analysis, Phytolaccaceae did not appear monophyletic either.

The topologies of the Bayesian and ML trees were very similar to each other, and were for the most part congruent with that of the MP tree (Fig. 3.4a and Fig. E.1). Major exceptions are the placement of the Limeaceae and the monophyly of the Cactaceae. Both the PP and BS_{ML} support increased for some relationships, but remained comparable to the BS_{MP} support for most of the nodes.

matK ORF and trnK Intron Combined

The *matK* ORF and the *trnK* intron combined ranged in length from 2342 bp (*Stegnosperma halmifolium*, Stegnospermataceae) to 2639 bp (*Drosera capensis*, Droseraceae), with an alignment length of 2689 characters following the insertion of gaps and the deletion of mutational hotspots/polyA-T regions in the *trnK* intron. Additionally, 284 indel characters were included in the dataset. Variable characters were 2068 (70%), of which 1373 (66%) were potentially parsimony informative. The MP analysis resulted in 6 trees, 6677 steps in length, with CI and RI values of 0.494 and 0.565, respectively (Fig. 3.5a, Table 3.3). The strict consensus tree is show in figure 3.5a. Examination of these trees revealed, as in the *trnK* intron analysis, that the differences in topologies were mostly due to the positions of *Stegnosperma* (Stegnospermataceae) and *Limeum* (Limeaceae). The exclusion of these two genera from the MP analysis reduced the

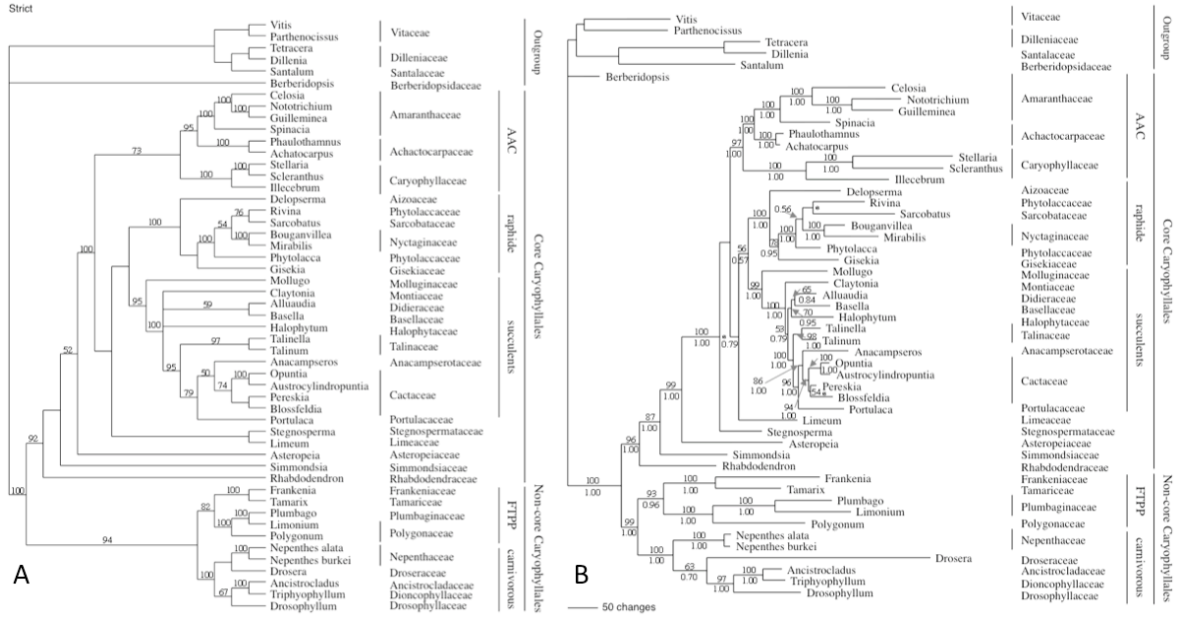

Figure 3.5 Phylogeny of Caryophyllales based on substitutions and indels from *matK* ORF/*trnK* intron combined. A) Strict consensus tree derived from Maximum Parsimony analysis. Bootstrap values are noted on the nodes. B) The 50% majority rule tree derived from Maximum Likelihood analysis. Bootstrap values are noted on the nodes with Posterior Probability values noted below the nodes. The * denotes nodes that differ between the ML and BI topologies.

number of trees to three, where the differences were due to the problematic placement of *Claytonia* (Montiaceae).

The Caryophyllales monophyly was again supported by 100% BS_{MP}. The major division into non-core Caryophyllales and core Caryophyllales (including Rhabdodendraceae, Simmondsiaceae, and Asteropeiaceae) was also evident. Support for these two major clades was 94% BS_{MP} and 92% BS_{MP}, respectively. The topology and support for non-core Caryophyllales was similar to that derived from *trnK* intron sequence information, except for the position of Droseraceae (Figs. 3.3a and 3.5). In core Caryophyllales, a grade of Rhabdodendraceae/Simmondsiaceae/Asteropeiaceae emerged first. This was followed by the AAC clade (73% BS_{MP}) with Amaranthaceae + Achatocarpaceae (95% BS_{MP}), and another clade of Stegnospermataceae + *Limeum* (Limeaceae) (<50% BS_{MP}) sister to the raphide clade (100% BS_{MP}) and the succulents + *Mollugo* clade (95% BS_{MP}).

The structure of the Bayesian and ML trees (Fig. 3.5b and Fig. E.3) was again very similar to that of the combined MP analysis (Fig. 3.5a). Unlike the MP analysis, *Limeum* and *Stegnosperma* did not form a clade in the ML and Bayesian trees, and the position of Droseraceae was shifted within its clade. There is, however, increased resolution and generally increased support in the Bayesian and ML trees where it was lacking in the MP analysis.

Discussion

Substitution patterns in matK ORF vs. trnK Intron

Despite the exclusion of some sectors from the *trnK* intron where homology assessment was unreliable, the remaining sectors provided 666 variable and 440 PI substitution characters (Table 3.3). This corresponds to about half of the substitution characters derived from *matK* (1119 variable and 882 PI). Although the percentages of variable and parsimony informative characters were comparable between the two regions (Table 3.3), there is a general trend of a higher proportion of variable and parsimony informative

characters in the *matK* ORF compared with *trnK* intron (Müller and Borsch, 2005; Wanke et al., 2007). Information from substitutions in *trnK* intron alone (indels not coded) resulted in a consensus tree with a well-resolved backbone (tree is not shown) where the major lineages received good support (core 64% BS_{MP}; non-core 74% BS_{MP}; AAC 73% BS_{MP}; raphide 82% BS_{MP}; succulents + *Mollugo* 66% BS_{MP}; FTTP <50% BS_{MP}; carnivorous 97% BS_{MP}). The CI and RI values were very comparable to those obtained from substitutions alone with the *matK* ORF (Table 3.3). To compare phylogenetic structure from *trnK* intron vs. *matK* ORF, we determined the average overall support for nodes and the number of nodes resolved following Källersjö, Albert, and Farris (1999). The two regions provided comparable amounts of signal in MP analyses with average BS_{MP} support values being 89% for *matK* and 85% for *trnK* intron for the 33 and 32 nodes, respectively, that received ≥50% BS_{MP} (Fig. 3.3). The number of nodes resolved, however, using the *matK* sequence information exceeded those of the *trnK* intron, being 42 for *matK* both with and without indels and 30 or 36 for *trnK* intron with and without indels, respectively. The extra resolution provided by the *matK* data compared with the *trnK* intron include the resolution *matK* provided for the five major lineages and the relationships among members of the succulents clade. Despite the slight difference in resolution, a considerable amount of phylogenetic signal is found in the *trnK* intron in the Caryophyllales.

Notable also was a general trend in branch lengths obtained with *matK* or *trnK* intron in ML analyses (Fig. 3.4). In general non-core Caryophyllales displayed relatively long branches, particularly Droseraceae; the exception being the branches leading to the families Dioncophyllaceae and Ancistrocladaceae in the carnivorous clade. The Ancistrocladaceae is known to have lost carnivory and in the Dioncophyllaceae only one of its three genera (*Triphyophyllum*) is carnivorous, though only during one stage of its life cycle (Meimberg et al., 2000; Bringmann et al., 2002). Within core Caryophyllales, sequence information from both genomic regions showed Rhabdodendraceae, Simmondsiaceae, and Asteropeiaceae to have considerably long branches. The taxonomic placement of these families has been problematic (Fay et al., 1997; Lledó et al., 1998; Nandi et al., 1998; Savolainen et al., 2000; Soltis et al., 2000; Cuénoud et al.,

2002; Hilu et al., 2003; Brockington et al., 2009). Similar long branches were displayed by members of the Caryophyllaceae sampled here, which may have contributed to its unusual placement in the *matK* MP analysis. In contrast, difficult to resolve families of the succulents clade displayed relatively short branches in both *matK* and *trnK* intron analyses (Fig. 3.4).

Considering the substitution rates for individual sites, it appears that both the 5' *trnK* intron and 5' region of *matK* possess some sites that have undergone extensive amounts of substitution (Fig. 3.1). In addition, domain X of *matK* also appears to have some sites with high rates of substitution (Fig. 3.1), which is in agreement with a recent study on rate of substitution in *matK* across land plants (Barthet and Hilu 2008). Considering the five rate classes, it appears that about one third of the sites were invariable in both regions (Table 3.1). However, the slowest rate class (RC 1) comprised the least proportion of variable sites in *matK* and *trnK* intron, reflecting the rapidly evolving nature of the two genomic regions. The other faster evolving classes (RC 2 – RC 4) although displaying overall similar proportions of sites, there are higher numbers of sites for the RC 2 class in *matK* and RC 4 for 3' *trnK* intron (Table 3.1). These differences are a reflection of the evolutionary constraints on *matK* being a protein-coding gene compared with the *trnK* intron; the 5' *trnK* intron may contain some gene regulatory elements for the downstream *matK* open reading frame, and thus may be under evolutionary constraints (Learn et al., 1992; Young and dePamphilis, 2000; Kelchner, 2002).

It is important to note that the trees resulting from the *trnK* intron partitioned analyses provided well-supported relationships that were congruent with those obtained by *matK* alone, such as the division into core and non-core Caryophyllales and the recognition of the raphide and succulents clades as well (detailed relationships are discussed below). In fact the *trnK* intron data provided higher support for non-core Caryophyllales than *matK* alone (88% BS_{MP} vs. 55% BS_{MP}, 93% BS_{ML}vs. 80% BS_{ML}, and 1.00 PP vs. 0.91 PP, Figs. 3.3, 3.4 and Figs. E.1, E.2). In addition, combining the *trnK* intron and *matK* datasets resulted in increased support for relationships across the order (Figs. 3.5 and 3.6).

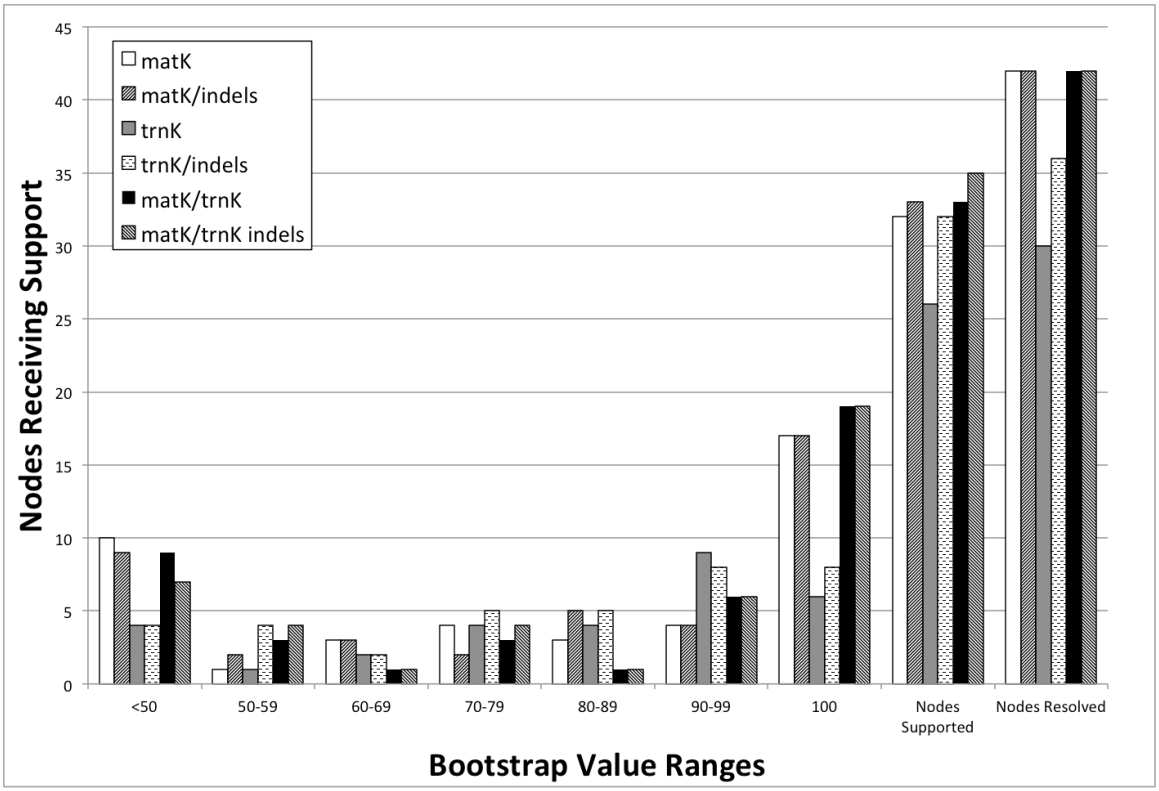

Figure 3.6 Comparing *matK* ORF and *trnK* intron separately and combined (substitutions vs. indels) for the degree of bootstrap support and number of nodes resolved from the Maximum Parsimony analyses.

Therefore, overall *trnK* intron provided useful phylogenetic information in the Caryophyllales, although *matK* provided more phylogenetic structure by resolving polytomies that existed in the *trnK* intron analysis alone (Figs. 3.3 and 3.4).

Phylogenetic informativeness profiles

Net informativeness profiles of the two genomic regions calculated in PhyDesign (Townsend 2007; Lopez-Giraldez and Townsend 2011) show the *matK* ORF to provide more phylogenetic signal than *trnK* intron across the history of the Caryophyllales with parallel decline deep in the history of the order (Fig. 3.2). Following the exclusion of certain regions of *trnK* intron due to difficulties in achieving accurate alignments, the number of characters used in the analyses was reduced to 889 compared with the 1179 characters used for *matK* (i.e. 33% fewer characters). Such a difference in number of characters may bias the net informativeness profiles in favor of *matK*. This was evident when the per-site informativeness profiles were examined, which the excludes sequence length variable; the per-site informativeness profile for *trnK* intron appears closer to that of *matK* (Fig. 3.2). Consequently, it appears that strength of phylogenetic signal of substitutions and indels in the trnK intron are comparable to that of the matK ORF. Therefore, in studies where homology assessments are achievable for higher proportions of *trnK* intron sequences, additional characters can be included, potentially resulting in increased phylogenetic information from this region.

Homoplasy, however, is a definite consideration since PhyDesign does not account for it in calculating informativeness. Degree of homoplasy measured by ensemble consistency index (CI) show that *trnK* intron has lower amount of homoplasy compared with *matK* (0.521 vs. 0.480), adding further support for the cost-effectiveness of utilizing the *trnK* intron in phylogenetic reconstruction at this phylogenetic depth. In both net and per-site analyses, combining *matK* and *trnK* intron sequences boosted the phylogenetic informativeness above either partition (Fig. 3.2), supporting the combined utility of these rapidly evolving genomic regions. The complementary information from *matK* and *trnK* intron combined improved phylogenetic accuracy when compared with the tree topology

of Brockington et al. (2009) by correcting the placement of the Caryophyllaceae in the *matK* partition, enhancing resolution within the raphide and succulents + *Mollugo* clades in the *trnK* intron partition (Figs. 3.3, 3.4, and 3.5), and increased total number of nodes resolved as well as increase in the number of nodes receiving 100% BS support throughout the tree (Fig. 3.6). Similar studies assessing cost-effectiveness of genomic regions using PhyDesign have demonstrated the differential contribution of genomic regions across different epochs for Ascomycota (Schoch et al., 2009) and muroid rodents by Townsend (2007) using the data set from Steppan et al. (2004).

Phylogenetic Utility of Indels in trnK intron and matK

Unlike most other genes commonly used in molecular phylogenetics, *matK* accommodates loss or gain of various stretches of nucleotides along its ORF. This mode of evolution takes place over both shallow and deep evolutionary histories, providing various numbers of indels in alignments that are potentially phylogenetically informative (Barthet and Hilu, 2008). Being a non-coding region, the *trnK* intron is very rich in indels particularly at this taxonomic depth. The *trnK* intron accommodated by far more indels than the *matK* ORF (*trnK* 223, *matK* 61). However, although the proportions of variable characters were very similar (*trnK* 70.4%, *matK* 68.9%), the proportion of PI indel characters was lower in the *trnK* intron analyses (*trnK* 59.6%, *matK* 71.5%; Table 3.3). This decline may indicate higher proportion of homoplasic indels in the *trnK* intron, as is also evident in the CI and RI values (Table 3.3). Nevertheless, the differential positive impact of the indels was apparent in the phylogenetic analyses of the partitioned *matK* and *trnK* intron datasets. In the MP analysis, the addition of indels to the *matK* dataset did not affect the number of most parsimonious trees, the number of nodes resolved nor tree topologies (Table 3.3). In contrast, the addition of indels to the *trnK* intron dataset greatly reduced the number of most parsimonious trees (3690 to 123) and number of nodes resolved (30 to 36; Table 3.3). In the combined data analysis, the number of trees was reduced from 36 to 6 with the inclusion of indels, the number of nodes resolved from 37 to 42 (Table 3.3). In terms of BS support for nodes, the values either increased or remained the same upon addition of indels to the datasets (Fig. 3.6).

A notable exception was in the partitioned *matK* dataset where there was a decrease in the number of nodes receiving 70%-79% BS_{MP} support (Fig. 3.6). In the Bayesian analyses, the inclusion of indels did not influence the number of nodes resolved in partitioned and combined analyses, however, there was a slight increase in the number of nodes supported by PP ≥ 0.95 (Fig. 3.6).

Phylogenetic Assessment: matK/trnK intron vs. other genomic regions

Global assessment of Caryophyllales phylogenetics was the focus of some studies (Cuénoud et al. 2002; Brockington et al. 2009; Crawley and Hilu 2011) but this subject has also received attention in several broad-scale investigations of angiosperm systematics (e.g. Savolainen et al. 2000; Hilu et al. 2003; Soltis et al. 2003; Soltis et al. 2011). These studies were based on one to several genomic regions with some incorporating *trnK* intron and *matK*. Although these phylogenetic studies differed in taxon sampling (both number and kind of placeholders), they provide a useful platform for comparison of phylogenetic signal in the *trnK* intron alone and when combined with *matK*. Since indels are part of the molecular features of *trnK* intron and *matK* sequence information (signal and noise), we will emphasize here results obtained from using indels as phylogenetic characters.

Caryophyllales backbone – The monophyly of the Caryophyllales was resolved with 100% $BS_{MP/ML}$ and 1.00 PP in both partitioned and combined data analyses. Maximum BS support for the caryophyllids monophyly was also obtained in MP analyses with varied representations by Fay et al. (1997; *rbcL*), Lledó et al. (1998; *rbcL*), Cuénoud et al. (2002; 18S rDNA, *rbcL*, *atpB*, and *matK* combined), and Brockington et al., (2009; nuclear: 18S and 26S, plastid: *atpB*, *matK*, *ndhF*, *psbB*, *psbT*, *psbN*, *rbcL*, *rpoC2*, *rps4*, and the entire plastid IR). In a global phylogenetic study of angiosperms by Savolainen et al. (2000), monophyly of the order received 74% JK_{MP} with *atpB*, 84% JK_{MP} with *rbcL*, and 97% JK_{MP} with *atpB*/*rbcL*. This support increased to 100% JK_{MP} with the addition of 18S rDNA to those two gene datasets (Soltis et al., 2000), but declined to 83% JK_{MP} with the inclusion of 26S rDNA sequences (Soltis et al., 2003). Support for the order was

99% JK_{MP} in the partial *matK* sequence study of Hilu et al. (2003). Therefore, the maximum support for the order with partial sequences of *trnK* intron reflects inherently strong phylogenetic signal in this intron (Figs. 3.3 and 3.4).

A basal split in the order into non-core and core Caryophyllales has not been consistently recovered and the placement of Rhabdodendraceae, Simmondsiaceae and Asteropeiaceae has been inconsistently resolved in previous molecular and non-molecular studies (Fay et al., 1997; Lledó et al., 1998; Nandi et al., 1998; Savolainen et al., 2000; Soltis et al., 2000; Cuénoud et al., 2002; Hilu et al., 2003; Brockington et al., 2009). Such a split was evident in partitioned *trnK* intron, *matK* and combined analyses (Figs. 3.3 and 3.4). Both *trnK* intron and *matK* analyses recovered the split into non-core and core Caryophyllales with the Rhabdodendraceae, Simmondsiaceae and Asteropeiaceae being sister to remaining core taxa. Non-core Caryophyllales received 94% BS_{MP}/99% BS_{ML}/1.00 PP and core Caryophyllales received 92% BS_{MP} /96% BS_{ML} /1.00 PP support (Figs. 3.3, 3.4, Figs. E.1, and E.2) in combined analyses. Support was lower with the partitioned data, but support with *trnK* intron superseded *matK* particularly for the non-core lineage (Figs. 3.3 and 3.4). The placement of these three families in the backbone of the order was unequivocal. However, the support for the relationships among the three was not strong in the partitioned analyses, but the ML and Bayesian analyses of the combined data provide convincing support (1.00 PP and 96%, 87%, 99% BS_{ML}) for a grade of *Rhabdodendron*/*Simmondsia*/*Asteropeia* (Fig. 3.5 and Fig. E.3).

A similar major split into two lineages was also recovered in the analyses of *matK* sequences of Cuénoud et al. (2002) and Hilu et al. (2003) and in the multi-gene study of Brockington et al., (2009). The non-core and core Caryophyllales lineages in the *matK*-based tree of Cuénoud et al. (2002) received 63% BS_{MP} and 53% BS_{MP}, and in Hilu et al. (2003) received 74% JK_{MP} and 68% JK_{MP}, respectively while in Brockington et al., (2009) they both received 100% BS_{ML}. Using *atpB* sequences alone, Savolainen et al. (2000) recovered a basal polytomy among non-core Caryophyllales, *Rhabdodendron* + *Simmondsia*, and a clade of remaining Caryophyllales. In contrast, their *rbcL* data analysis depicted the basal split but included *Rhabdodendron* and *Simmondsia* with non-

core Caryophyllales, albeit with weak support. Using *rbcL*, Fay et al. (1997) and Lledó et al. (1998) showed *Rhabdodendron* sister to remaining Caryophyllales (with 100% BS_{MP} in the latter study) while *Simmondsia* was nested within non-core Caryophyllales or sister to remaining non-core taxa. In Cuénoud et al. (2002), a *Rhabdodendron* + *Simmondsia* clade was depicted basal to core Caryophyllales in their MP analysis of *matK* sequences, but with <50% support. Therefore, *trnK* intron alone or in combination with *matK* contributed significant signal for resolving these relationships.

Non-core Caryophyllales – This lineage was recovered in all single-gene and multigene analyses of the Caryophyllales (Fay et al., 1997; Lledó et al., 1998; Savolainen et al., 2000; Soltis et al., 2000; Cuénoud et al., 2002; Hilu et al., 2003; Brockington et al., 2009). However, the clade either lacked support or received moderate support in most analyses (Fay et al., 1997; Lledó et al., 1998; Savolainen et al., 2000; Soltis et al., 2000; Cuénoud et al., 2002; Hilu et al., 2003). The highest support (95% BS_{MP}) was achieved in Cuénoud et al. (2002) by combining sequences from 18S rDNA, *rbcL*, *atpB*, and *matK* and in Brockington et al., (2009; 100% BS_{ML}) using 9 plastid genes, 2 nuclear genes and the entire IR. The *trnK* intron alone provided 88% BS_{MP}/93% BS_{ML}/1.00 PP support and the inclusion of *matK* further increased support to 94% BS_{MP}/99% BS_{ML}/1.00 PP (Figs. 3.3-3.5 and Figs. E1 and E3).

Within non-core Caryophyllales, the carnivorous clade, one of the two lineages of non-core Caryophyllales, received 95% to 100% $BS_{MP/ML}$ and 1.00 PP support in the partitioned and combined data analyses (Figs. 3.3 and 3.4). This is the highest support for the single origin of carnivorous plants in the Caryophyllales except for the Hilu et al. (2003) *matK* study (96% JK) and the Brockington et al. (2009; 100% BS_{ML}) study. The group either appeared in a major polytomy with a three gene analysis (Soltis et al., 2000) or was resolved but with moderate (Meimberg et al., 2000; Cameron et al., 2002; Cuénoud et al., 2002) to weak support (Lledó et al., 1998; Savolainen et al., 2000). The relationships among families of this group were consistent in our partitioned and combined data analyses except for the placement of the Droseraceae (Figs. 3.3-3.5). The placement of the Droseraceae in non-core Caryophyllales appears to be problematic in

previous studies as well due to topological differences and insufficient support (Fay et al., 1997; Lledó et al., 1998; Meimberg et al., 2000; Savolainen et al., 2000; Soltis et al., 2000; Cameron et al., 2002; Cuénoud et al., 2002; Hilu et al., 2003; Meimberg and Heubl, 2006; Brockington et al., 2009). The most supported relationship was between, Ancistrocladaceae and Dioncophyllaceae, the two non-carnivorous families in the lineage (except *Triphyophyllum*), which is in agreement with previous studies (Fay et al., 1997; Lledó et al., 1998; Meimberg et al., 2000; Soltis et al., 2000; Cameron et al., 2002; Cuénoud et al., 2002; Hilu et al., 2003; Meimberg and Heubl, 2006; Brockington et al., 2009).

Support of the monophyly of the FTPP clade was lower for *trnK* intron (54% BS_{MP}/57% BS_{ML}) compared with *matK* (80% BS_{MP}/66% BS_{ML}/0.78 PP) but increased when the signal from the two regions were combined (82% BS_{MP}/93% BS_{ML}/0.96 PP) (Figs. 3.3-3.5). Support for the monophyly of the FTPP clade has been low in most studies, except for the 90% JK support of Hilu et al. (2003) and the 97% BS_{ML} in Brockington et al. (2009). The internal structure of the clade in this study mirrors all previous studies in terms of topology and high support (Fay et al., 1997; Lledó et al., 1998; Meimberg et al., 2000; Soltis et al., 2000; Cameron et al., 2002; Cuénoud et al., 2002; Hilu et al., 2003; Brockington et al., 2009).

Core Caryophyllales – Since the placement of the Rhabdodendraceae, Simmondsiaceae, and Asteropeiaceae have been discussed previously, we will address here the phylogenetic relationship among remaining core Caryophyllales. $BS_{MP/ML}$ and PP support was maximum with *trnK* intron, *matK* and *trnK*/*matK* for the monophyly of remaining core Caryophyllales. The *trnK* intron recovered the AAC clade with similar internal topology and comparable support (85% BS_{MP}/93% BS_{ML}/1.00 PP) to that obtained with combined *trnK*/*matK* (73% BS_{MP}/97% BS_{ML}/1.00 PP; Figs. 3.3-3.5). However, *matK* alone placed Caryophyllaceae after *Asteropeia* (Asteropeiaceae) in a clade with *Limeum* (Limeaceae) with <50% BS_{MP} support whereas *matK* ML and Bayesian analyses reproduced the topology obtained with *trnK* intron and *trnK*/*matK* with 90% BS_{ML} /0.97 PP support (Figs. 3.3-3.5 and Fig E.2). This latter topology was similar to that obtained

in previous single to multi-genomic regions analyses (Soltis et al. 2000; Cuénoud et al. 2002; Kadereit et al. 2003; Müller & Borsch 2005; Brockington et al. 2009; Crawley and Hilu 2011).

The second clade to emerge in the MP combined data analysis contains *Limeum* (Limeaceae) and *Stegnosperma* (Stegnospermataceae), albeit with <50% support. The position of these two families was inconsistently resolved in the remaining analyses of our datasets. It is to be noted that the exclusion of these families from the analyses reduced the number of most parsimonious trees from 6 to 3 in the combined analysis and from 123 to 12 in the *trnK* intron analysis.

Remaining core Caryophyllales form two highly supported lineages (Figs. 3.3-3.5), the raphide clade and the succulents + *Mollugo* clade. These lineages posses the shared derived character of a globular crystal in sieve-element plastids (Behnke, 1994). Members of the raphide clade (Aizoaceae, Phytolaccaceae, and Nyctaginaceae) are recognized by the presence of raphide crystal as a synapomorphy (Judd et al., 1999; Dequan and Hartmann, 2003; Stevens, 2008). Sequence information from *trnK* intron recovered the raphide clade with 90% BS_{MP}/93% BS_{ML}/1.00 PP, which approaches closely the maximum support of *matK* and combined *matK*/*trnK* (Figs. 3.3-3.5). Support for this clade ranged from 51% BS_{MP} with *rbcL* (Savolainen et al., 2000) to 100% BS_{ML} in Brockington et al. (2009). The lack of monophyly of the Phytolaccaceae is clear, although the exclusion of *Phytolacca* from *Rivina* received low support (Figs. 3.3-3.5). Support for monophyly of the Phytolaccaceae has not been shown in previous studies (Savolainen et al., 2000; Soltis et al., 2000; Cuénoud et al., 2002; Hilu et al., 2003; Brockington et al., 2009) and APG III (2009) acknowledges that there is still much to be learned about this family, which is "almost certainly not monophyletic."

The succulents clade was recovered in the partitioned and combined *matK*/*trnK* intron analyses with high to very strong support (84%-95% BS_{MP}, 91%-99% BS_{ML}, 1.00 PP) with *Mollugo* (Molluginaceae) sister to remaining members (Figs. 3.3-3.5). Historically, the relationships within this clade are not well-resolved and lacked support (Carolin,

1987; Hunziker et al., 2000; Cuénoud et al., 2002; Hilu et al., 2003; Brockington et al., 2009). *Mollugo* appeared in this position in 2-4 combined gene analyses (Soltis et al., 2000; Cuénoud et al., 2002) but with very low support while receiving 100% BS_{ML} in the Brockington et al., (2009).

Phylogenetic signal from the combined sequences of *matK* and *trnK* intron resulted in an overall robust phylogeny for the Caryophyllales (Fig. 3.5). The two regions in some cases displayed differential resolving capabilities such as the case of non-core Caryophyllales where BS_{MP} support from *matK* was 55% whereas for *trnK* intron it was 88% (Figs. 3.3 and 3.4). Problems in Caryophyllales remain in terms of potential polyphyly of Phytolaccaceae and placement of some problematic genera such as *Stegnosperma*, *Claytonia*, and *Limeum*.

Conclusion

The *trnK* intron was comparable to *matK* in terms of proportion of variable sites, PI sites, and the distribution of those sites among rate classes. The phylogenetic trees obtained from partitioned *trnK* intron and *matK* datasets were fairly congruent in terms of topology and support. However, a more robust phylogeny was obtained from the combined datasets. Phylogenetic reconstruction based on the partial sequences of *trnK* intron at the ordinal level in the Caryophyllales is in line with many recent studies and tree robustness approaches those derived from multi-gene studies. Therefore, the co-amplified and often co-sequenced *trnK* intron should be included in phylogenetic studies at deeper historic level as indicated in this study and our previous work on early diverging eudicots (Hilu et al. 2008). The outcome of this investigation provides further support for the utility of rapidly evolving genomic regions for deeper phylogenetic reconstructions.

Acknowledgements

We would like to thank Douglas and Pamela Soltis, Samuel Brockington, and Michael Moore, as well as the Missouri Botanical Garden and the Royal Botanic Garden at Kew for their generosity in providing DNA samples for several taxa. We thank Michelle Barthet for help in designing one of the primers used here, Anya Hinckle for helping with specimen collection, Shelli Newman for assistance in lab work, and Gordon Burleigh for advice on the data analyses. This work was supported by a grant from the National Science Foundation, USA - EF-043105 to KWH.

References:

Albert, V. A., S. E. Williams, and M. W. Chase. 1992. Carnivorous Plants: Phylogeny and Structural Evolution. Science **257**:1491-1495.

APG II. 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Botanical Journal of the Linnean Society **141**:399-436.

APG III. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Botanical Journal of the Linnean Society **161**:105-121.

Barthet, M. M., and K. W. Hilu. 2007. Expression of *matK*: Functional and Evolutionary Implications. American Journal of Botany **94**:1402-1412.

Barthet, M. M., and K. W. Hilu. 2008. Evaluating evolutionary constraint on the rapidly evolving gene *matK* using protein composition. J Mol Evol **66**:85-97.

Behnke, H.-D. 1994. Sieve-Element Plastids: Their Significance for the Evolution and Systematics of the Order. Pp. 87-121 *in* H.-D. Behnke, and T. J. Mabry, eds. Caryophyllales: Evolution and Systematics. Springer Verlag, Berlin, Germany.

Bittrich, V. 1993. Introduction to Centrospermae. Pp. 13-19 *in* K. Kubitzki, J. G. Rohwer, and V. Bittrich, eds. The families and genera of vascular plants, vol. II, Magnoliid, hamamelid, and caryophyllid families. Springer Verlag, Berlin, Germany.

Borsch, T., K. W. Hilu, D. Quandt, V. Wilde, C. Neinhuis, and W. Barthlott. 2003. Noncoding plastid trnT-trnF sequences reveal a well resolved phylogeny of basal angiosperms. Journal of Evolutionary Biology **16**:558-576.

Bringmann, G., H. Rischer, J. Schlauer, and K. Wolf. 2002. The tropical liana *Triphyophyllum peltatum* (Dioncophyllaceae): Formation of carnivorous organs is only a facultative prerequisite for shoot elongation. Carnivorous Plant Newsletter **31**:44-52.

Brockington, S. F., R. Alexandre, J. Ramdial, M. J. Moore, S. Crawley, A. Dhingra, K. Hilu, D. E. Soltis, and P. S. Soltis. 2009. Phylogeny of the Caryophyllales Sensu Lato: Revisiting Hypotheses on Pollination Biology and Perianth Differentiation in the Core Caryophyllales. International Journal of Plant Sciences **170**:627-643.

Cameron, K. M., K. J. Wurdack, and R. W. Jobson. 2002. Molecular Evidence for the Common Origin of Snap-Traps Among Carnivorous Plants. American Journal of Botany **89**:1503-1509.

Carolin, R. 1987. A Review of the Family Portulacaceae. Australian Journal of Botany **35**:383-412.

Clement, J. S., and T. J. Mabry. 1996. Pigment Evolution in the Caryophyllales: a Systematic Overview. Botanica Acta **109**:360-367.

Crawley, S. S., and K. W. Hilu. 2011. Impact of missing data, gene choice, and taxon sampling on phylogenetic reconstruction: the Caryophyllales (angiopserms). Plant Systematics and Evolution.

Cronquist, A., and R. F. Thorne. 1994. Nomenlatural and Taxonomic History. Pp. 5-25 *in* H.-D. Behnke, and T. J. Mabry, eds. Caryophyllales: Evolution and Systematics. Springer Verlag, Berlin, Germany.

Cuénoud, P., V. Savolainen, L. W. Chatrou, M. P. Powell, R. J. Grayer, and M. W. Chase. 2002. Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL, atpB,* and *matK* DNA sequences. American Journal of Botany **89**:132-144.

Dequan, L., and H. E. K. Hartmann. 2003. Molluginaceae. Flora of China **5**:437-439.

Doyle, J. J., and J. L. Doyle. 1990. Isolation of plant DNA from fresh tissue. Focus **12**:13-25.

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

Edwards, K. J., and P. A. Gadek. 2001. Evolution and biogeography of *Alectryon* (Sapindaceae). Molecular Phylogenetics and Evolution **20**:14-26.

Farris, J. S., M. Källersjö, A. G. Kluge, and C. Bult. 1995. Testing significance of incongruence. Cladistics **10**:315-319.

Fay, M. F., K. M. Cameron, G. Prance, T., M. D. Lledo, and M. W. Chase. 1997. Familial relationships of *Rhabdodendron* (*Rhabdodendraceae*): plastid *rbc*L sequences indicate a caryophyllid placement. Kew Bulletin **54**:923-932.

Felsenstein, J. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution **39**:783-791.

Graham, S. W., and R. G. Olmstead. 2000. Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. American Journal of Botany **87**:1712-1730.

Hassan, N. M. S., U. Meve, and S. Liede-Schumann. 2005. Seed coat morphology of Aizoaceae-Sesuvioideae, Gisekiaceae and Molluginaceae and its systematic significance. Botanical Journal of the Linnean Society **148**:189-206.

Hilu, K. W., and L.A. Alice. 1999. Evolutionary Implications of *matK* Indels in Poaceae. American Journal of Botany **86**:1735 - 1741.

Hilu, K. W., T. Borsch, K. Müller, D. E. Soltis, P. S. Soltis, V. Savolainen, M. W. Chase, M. P. Powell, L. A. Alice, R. Evans, H. Sauquet, C. Neinhuis, T. A. B. Slotta, G. R. Jens, C. S. Campbell, and L. W. Chatrou. 2003. Angiosperm phylogeny based on *matK* sequence information. American Journal of Botany **90**:1758-1776.

Hilu, K. W., C. Black, D. Diouf, and J.G. Burleigh. 2008. Phylogenetic signal in *matK* vs. *trnK*: a case study in early diverging eudicots (angiosperms). Molecular Phylogenetics and Evolution **48**:1120-1130.

Hu, J.-M., M. Lavin, M. F. Wojciechowski, and M. J. Sanderson. 2000. Phylogenetic Systematics of the Tribe Millettieae (Leguminosae) Based on Chloroplast *trnK/matK* Sequences and its Implications for Evolutionary Patterns in Papilionoideae. American Journal of Botany **87**:418-430.

Hunziker, J. H., R. Pozner, and A. Escobar. 2000. Chromosome number in *Halophytum ameghinoi* (Halophytaceae). Plant Systematics and Evolution **221**:125-127.

Johnson, L. A., and D. E. Soltis. 1995. Phylogenetic inference in Saxifragaceae sensu stricto and *Gilia* (Polemoniaceae) using *matK* sequences. Annals of the Missouri Botanical Gardens **82**:149-175.

Judd, W. S., C. S. Campbell, E. A. Kellogg, and P. F. Stevens. 1999. Plant Systematics: A Phylogenetic approach. Sinauer, Sunderland, Massachusetts, USA.

Kadereit, G., T. Borsch, K. Weising, H. Freitag. 2003. Phylogeny of Amaranthaceae and Chenopodiaceae and the evolution of C_4 photosynthesis. International Journal of Plant Sciences **164**:959-986.

Källersjö, M., V. A. Albert, and J. S. Farris. 1999. Homoplasy *Increases* Phylogenetic Structure. Cladistics **15**:91-93.

Källersjö, M., J. S. Farris, M. W. Chase, B. Bremer, M. F. Fay, C. J. Humphries, G. Petersen, O. Seberg, and K. Bremer. 1998. Simultaneous Parsimony Jackknife Analysis of 2538 *rbcL* DNA Sequences Reveals Support For Major Clades of Green Plants, Land Plants, Seed Plants, and Flowering Plants. Plant Systematics and Evolution **213**:259-287.

Kelchner, S.A. 2000. The Evolution of Non-Coding Chloroplast DNA and Its Application in Plant Systematics. Annals of the Missouri Botanical Garden **87**:482-498.

Kelchner, S. A. 2002. Group II introns as phylogenetic tools: structure, function, and evolutionary constraints. American Journal of Botany **89**:1651 - 1669.

Kosakovsky Pond, S. L., S. D. W. Frost, and S. V. Muse. 2005. HyPhy: Hypothesis Testing Using Phylogenies. Bioinformatics **21**:676-679.

Lavin, M., R. T. Pennington, B. B. Klitgaard, J. I. Sprent, H. Cavalcante de Lima, and P. E. Gasson. 2001. The Dalberdioid Legumes (Fabaceae): Delimitation of a Pantropical Monophyletic Clade. American Journal of Botany **88**:503-533.

Learn Jr., G. H., J. S. Shore, G. R. Furnier, G. Zurawski, and M. T. Clegg. 1992. Constraints on the evolution of plastid introns: the group II intron in the gene encoding tRNA-Val(UAC). Mol Biol Evol **9**:856-871.

Liang, H., and K. W. Hilu. 1996. Application of the *mat*K gene sequences to grass systematics. Canadian Journal of Botany **74**:125-134.

Lió, P., and N. Goldman. 1998. Models of molecular evolution and phylogeny. Genome Research **8**:1233-1244.

Lledo, M. D., M. B. Crespo, K. M. Cameron, M. F. Fay, and M. W. Chase. 1998. Systematics of Plumbaginaceae Based upon Cladistic Analysis of rbcL Sequence Data. Systematic Botany **23**:21-29.

López-Giráldez, F., and J. P. Townsend. 2011. PhyDesgin: an online application for profiling phylogenetic informativeness. BMC Evolutionary Biology **11**:152.

Meimberg, H., P. Dittrich, G. Bringmann, J. Schlauer, and G. Heubl. 2000. Molecular Phylogeny of Caryophyllidae s.l. Based on *MatK* Sequences with Special Emphasis on Carnivorous Taxa. Plant Biology **2**:218-228.

Meimberg, H., and G. Heubl. 2006. Introduction of a Nuclear Marker for Phylogenetic Analysis of Nepenthaceae. Plant Biology **8**:831 - 840.

Morton, C. M., K. G. Karol, and M. W. Chase. 1997. Taxonomic Affinities of *Physena* (Physenaceae) and *Asteropeia* (Theaceae). The Botanical Review **63**:231-239.

Müller, J., and K. Müller. 2003. QuickAlign: A New Alignment Editor. Plant Molecular Biology Reporter **21**:5.

Müller, K. F. 2005. SeqState-primer design and sequence statistics for phylogenetic DNA data sets. Applied Bioinformatics **4**:65-69.

Müller, K., and T. Borsch. 2005. Multiple origins of a unique pollen feature: stellate pore ornamentation in Amaranthaceae. Grana **44**:266-281.

Nandi, O. I., M. W. Chase, and P. K. Endress. 1998. A Combined Cladistic Analysis of Angiopserms Using rbcL and Non-Molecular Data Sets. Annals of the Missouri Botanical Garden **85**:137-214.

Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14**:817-818.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572-1574.

Ronsted, N., S. Law, H. Thronton, M. F. Fay, and M. W. Chase. 2005. Molecular Phylogenetic evidence for monophly of *Fritillaria* and *Lilium* (Liliaceae; Liliales)

and the infrageneric classification of Fritillaria. Molecular Phylogenetics and Evolution **35**:509-527.

Savolainen, V., M. W. Chase, S. B. Hoot, C. M. Morton, D. E. Soltis, C. Bayer, M. F. Fay, A. Y. DeBruijn, S. Sullivan, and Y.-L. Qiu. 2000. Phylogenetics of Flowering Plants Based on Combined Analysis of Plastid *atpB* and *rbcL* Gene Sequences. Systematic Biology **49**:306-362.

Schoch, C. L., G.-H. Sung, and F. López-Giráldez, et al. 2009. The ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. Systematic Biology **58**:224-239.

Simmons, M. P., and H. Ochoterena. 2000. Gaps as Characters in Sequence-Based Phylogenetic Analyses. Systematic Biology **49**:369-381.

Soltis, D. E., A. E. Senters, M. J. Zanis, S. Kim, J. D. Thompson, P. S. Soltis, L. P. Ronse De Craene, P. K. Endress, and J. S. Farris. 2003. Gunnnerales Are Sister to Other Core Eudicots: Implications for the Evolution of Pentamery. American Journal of Botany **90**:461-470.

Soltis, D. E., P. S. Soltis, M. W. Chase, M. E. Mort, D. C. Albach, M. Zanis, V. Savolainen, W. H. Hahn, S. B. Hoot, M. F. Fay, M. Axtell, S. M. Swensen, L. M. Prince, W. J. Kress, K. C. Nixon, and J. S. Farris. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. Botanical Journal of the Linnean Society **133**:381-461.

Soltis, D. E., S. A. Smith, N. Cellinese, et al. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. American Journal of Botany **98**:704-730.

Stamatakis, A., P. Hoover, and J. Rougemont. 2008. A Fast Bootstrapping Algorithm for the RAxML Web Servers. Systematic Biology **57**:758-771.

Steppan, S. J., R. M. Adkins, and J. Anderson. 2004. Phylogeny and divergence-date estimates of rapid radiations in muroid rodents based on multiple nuclear genes. Systematic Biology **53**:533-553.

Stevens, P. F. 2008. Angiosperm Phylogeny Website. Version 9, June 2008. http://www.mobot.org/MOBOT/research/APweb/.

Swofford, D. L. 2003. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4, Sinauer Associates, Sunderland, Massachusetts, USA.

Townsend, J. P. 2007. Profiling Phylogenetic Informativeness. Systematic Biology **56**:222-231.

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

Wanke, S., M. A. Jaramillo, T. Borsch, M. S. Samain, D. Quandt, C. Neinhuis. 2007. Evolution of Piperales - *matK* gene and *trnK* intron sequence data reveal lineage specific resolution contrast. Molecular Phylogenetics and Evolution **42**:477 - 497.

Williams, S. E., V. A. Albert, and M. W. Chase. 1994. Relationships of Droseraceae: A Cladistic Analysis of rbcL Sequence and Morphological Data. American Journal of Botany **81**:1027-1037.

Wilson, C. 2004. Phylogeny of *Iris* based on chloroplast *matK* gene and *trnK* intron sequence data. Molecular Phylogenetics and Evolution **33**:402-412.

Young, N. D., and C. W. dePamphilis. 2000. Purifying selection detected in the plastid gene *matK* and flanking ribozyme regions within a group II intron of nonphotosynthetic plants. Molecular Biology and Evolution **17**:1933-1941.

Zanis, M. D., D. E. Soltis, P. S. Soltis, S. Mathews, and M. J. Donoghue. 2002. The root of angiosperms revisited. Proceedings of the National Academy of Science USA **99**:6848-6853.

Chapter 4: RNA editing of *matK* in the Gnetales?

Abstract

The gymnosperm order Gnetales is comprised of three distinct genera; *Gnetum* (Gnetaceae), *Ephedra* (Ephedraceae), and *Welwitschia* (Welwitschiaceae). These genera have proven difficult to place phylogenetically despite persistent efforts to resolve their position in the tree of life. Additionally, this group has been shown to exhibit unique characteristics with respect to the chloroplast gene *matK* in that it has a higher proportion of non-polar amino acids in Gnetales than other land plants that corresponds to a transmembrane domain predicted solely for members of this order. These features make it an intriguing sample group for the investigation of RNA editing. The inclusion of known RNA editing sites in phylogeny reconstruction is not widely practiced, and yet there is little evidence to discourage the practice. We show here that there is no RNA editing of *matK* in the Gnetales. We did, however, find a total of 20 RNA edited sites in two Zamiaceae and 3 Pinaceae species. Additionally, based on the presence of a 38 base pair (bp) insertion/deletion in *matK* 62 bp downstream from the consensus start codon for some members of *Ephedra*, we propose an alternate start codon for this genus. Our phylogenetic analyses of DNA alone or DNA plus cDNA, for those species shown to undergo RNA editing, revealed no impact of including RNA editing sites on tree topology, support, or resolution.

Introduction

The plant order Gnetales (gymnosperms) consist of 96 species grouped into 3 very distinct genera; *Gnetum* (Gnetaceae), *Ephedra* (Ephedraceae), and *Welwitschia* (Welwitschiaceae). This order has been difficult to place phylogenetically in the tree of life for seed plants for a number of reasons. They possess features that are shared with flowering plants, such as net-veined leaves in Gnetum, flower-like structures, double fertilization (Friedman and Carmichael, 1996), and the presence of vessels in the wood

(Carlquist 1996; Donoghue and Doyle, 2000). They also have features similar to conifers including tracheids with circular bordered pits, a lack of scalariform pitting in primary xylem, and scale-like and strap-shaped leaves (Donoghue and Doyle, 2000). These morphological features along with molecular data have made the phylogenetic position of Gnetales controversial. The group has been regarded as sister to angiosperms, sister to gymnosperms, sister to seed plants, or sister to the pine group within gymnosperms in the "gnepine" hypothesis (Bowe et al., 2000; Chaw et al., 2000; Donoghue and Doyle, 2000; Graham and Olmstead 2000; Gugerli et al., 2001; Magallon and Sanderson 2002; Soltis et al., 2002; Burleigh and Mathews 2004; Hajibabaei et al., 2006; Rai et al., 2008; Braukmann et al., 2009; Magallón and Hilu, 2009; Mathews 2009).

In addition to their controversial phylogenetic placement, this group has exhibited unique characteristics with respect to the chloroplast gene *matK*. The *matK* gene is the only putative group II intron maturase in the chloroplast genome (Barthet and Hilu, 2008). It is currently one of the most widely used genomic regions in reconstructing plant phylogenies (Hilu et al., 2003; Harrington et al., 2005; Müller et al., 2006; Hilu et al., 2008; Mort et al., 2008; Cadotte et al., 2009). It is nested within the intron of tRNA Lys (UUU), *trnK*, in the large single copy region of the plastid genome. The *matK* gene undergoes nucleotide substitution at a very high rate (3 times that of *rbcL*) and results in amino acid mutations at an even higher rate (6 times faster than *rbcL*) (Müller et al., 2006; Johnson and Soltis, 1995). It has been shown that this rapidly evolving gene in the Gnetales has a higher proportion of non-polar amino acids when compared to other land plants (Barthet and Hilu, 2008). Additionally, TMHMM and TMAP transmembrane prediction programs identified a transmembrane domain corresponding to this increased hydrophobicity unique to the Gnetales (Barthet 2006; Barthet and Hilu 2008).

The combination of the Gnetales difficult phylogenetic placement and their uniquely predicted transmembrane domain within *matK* led us to explore the possibility of RNA editing within the *matK* transcript of this group since the presence of RNA editing sites in phylogenetic studies has the potential to impact the phylogenetic placement of the order. RNA editing is a mechanism that cells employ to alter an RNA message post

transcription but before translation can occur (Tillich et al., 2006). Generally this editing occurs in the form of a C to U edit, and has been documented in almost all land plant chloroplasts studied to date with the exception of the Marchantiid liverworts (Freyer et al., 1997; Tillich et al., 2006). It has been noted that there is a general preference for sites to be edited in the second codon position with a bias toward certain amino acid transitions with proline (P) to leucine (L), serine (S) to leucine (L), and serine (S) to phenylalenine (F) being among the most common (Freyer et al., 1997). RNA editing of the *matK* transcript has been shown in several species including barely (Vogel et al., 1997), the maidenhair fern *Adiantum* (Wolf et al., 2003), and rice (Inada et al., 2004). RNA editing has also been examined in gymnosperms. Wakasgui et al., (1996) noted RNA editing of 26 sites in 12 plastid genes of black pine (*Pinus thunbergii*) but *matK* was not one of the genes analyzed in this study. Chen et al. (2011) conducted a study of RNA editing in the cycad *Cycas taitungensis* and detected 85 editing sites in 25 transcripts, of which one corresponds to partial editing of a single site in *matK*. We have sequenced DNA and cDNA transcripts of *matK* for nine Gnetales species (16 individuals) along with several representatives of Cycadaceae, Ginkgoaceae, Pinaceae, and Zamiaceae to determine the pattern of RNA editing within this enigmatic group.

Materials and Methods

Plant material

Representatives from each of the three Gnetales genera were sampled (16 accessions total) as were three Pinaceae, two Zamiaceae, one Cycadaceae, and *Ginkgo bioloba* (Table 4.1). All plant material was collected from plants grown in the Virginia Tech Greenhouse or in the Hahn Horticulture Garden on Virginia Tech Campus in Blacksburg, VA. The only exceptions being one *Gnetum gnemon* species (designated *G. gnemon* DUKE) and *Gnetum leyboldii,* which were obtained from the Duke University Greenhouse (Table 4.1). Leaf material was cut from the living plant, immediately submerged in liquid nitrogen, and then transported on dry ice for storage at -80°C until DNA/RNA isolation was completed.

Table 4.1 Species included in the study. For several species, multiple individuals were sampled in which case the individual identifiers are listed in parenthesis following the species name.

Family	Genus species
Gnetales	
Ephedraceae	*Ephedra equisetina* (1, 2, and 3)
	Ephedra intermedia (1, 2, and 3)
	Ephedra nevadensis
	Ephedra sinica (1, 2, and 3)
	Ephedra viridis
Gnetaceae	*Gnetum gnemon* (DUKE, VT)
	Gnetum leyboldii
	Gnetum montanum
Welwitschiaceae	*Welwitschia mirabilis*
Other	
Cycadaceae	*Cycas revoluta*
Ginkgoaceae	*Ginkgo biloba*
Pinaceae	*Picea omorkia*
	Pinus sylvestris
	Pinus thunbergii
Zamiaceae	*Dioon edule*
	Zamia fisheri

Generation of new matK DNA/RNA sequences

Genomic DNA extraction followed the CTAB method of (Doyle and Doyle 1990). RNA isolation was done either using TRIzol Reagent (Invitrogen, Carlsbad CA) or the Spectrum Plant Total RNA Kit (Sigma, Saint Louis MO). Total RNA was digested with DNase I (Sigma, Saint Louis MO) to ensure the samples were free of DNA contamination followed by first strand synthesis reactions using and Superscript II reverse transcriptase (RT) (Invitrogen, Carlsbad CA). The product of this reaction was then PCR amplified as described below. To further ensure that RNA samples were free of DNA contamination, each sample was prepared for first strand syntheses, with and without the RT enzyme and subjected to PCR. Any samples that showed a PCR product where no RT enzyme was used were discarded and a fresh sample was treated with DNase I followed by repeating the first strand synthesis and PCR reactions.

To generate new sequences for the entire *matK* ORF, the region was amplified in two overlapping sections using a combination of external primers located immediately upstream of the *matK* start codon and downstream of the *matK* stop codon (within the *trnK* intron) and internal *matK* primers (Table 4.2). Primers were designed specific to each family and based on existing GenBank sequences for species of the same genus/genera where possible (Table 4.2). The DNA and cDNA were amplified on either a PTC-100 or a PTC-200 Peltier Thermal Cycler (MJ Research, Waltham, Massachusetts). The 20 μL PCR reactions contained 0.5 μL of 20 mM primer, 0.2 μL of 2,000 U/mL Phusion polymerase (New England BioLabs, Ipswich MA), 4.0 μL 5X Phusion HF buffer (New England BioLabs, Ipswich MA), and 0.4 μL of 10 mM dNTPs (Promega, Madison, Wisconsin). The thermocycling profile for amplification was: 98°C for 30 s, then 34 cycles of 98°C for 10 s, 48°C for 30 s, and 72°C for 1 min, and a final elongation step of 10 mins at 72°C. PCR products were cleaned using Promega's Wizard SV gel and PCR clean-up system (Promega, Madison WI). Sequences were generated at the Duke University sequencing lab (Applied Biosystems 3730 XL DNA Analyzer) using

Table 4.2 Primers used in amplification of DNA/cDNA templates.

Family	Primer name	Primer sequence
Ephedraceae		
	matK f-1	CAG GAG AAC GCC TGG TTG C
	matK f-2	CTT CGA TTC ATT CAG AGC TG
	matK r-1	GCA CAC GGC TTT CTC TCT G
	matK r-2	CTA GAA TAG TAG TTC CCA GC
Gnetaceae		
	matK f-1	CGA CCA AAC TAG ATT GCA C
	matK f-2	GTT TCC AGG TGG GAG TTA C
	matK r-1b	GGG CTT GCA ATT TTC ATC GC
	matK r-2	GGT TTT CCC GTA ATG TCG C
Welwitschiaece		
	matK f-1	GCA CCG TGT GTC TGT GTG C
	matK f-2	GTT TCC AAA GGG GAG TCT AG
	matK r-1	GCA AAC AAG CCT TTC GTT CG
	matK r-2	CAA CTT ACT AAT GGG TCT TCC C
Pinaceae		
	matK f-1	GGT TAT TCT CAT GAA CGA GGG
	matK f-2	GGT TCG AAC CTT TCG TCG C
	matK r-1	GAT TCT GAA TCG AGG CAA TTA C
	matK r-2	CCG AAC CCT CAG AAA ATA ACC
Cycadaceae/Zamiaceae		
	matK f-1	CGT AAC TCA AGA GGT CAC CC
	matK f-2	GCT GGA TCC AAG ATG CTC C
	Cycas matK r-1	CGT AAT CAT CCT AAT TCC CGG
	matK r-1	CGA TCG TGG ATC GAT TCC GG
	matK r-2	CGC TCA ATA AAT AGC CCA GG
Ginkgoaceae		

matK f-1	GTT ACT CTC ACA AGG GCC
matK f-2	CTC ATC GCT CCG GAA GG
matK r-1	CGA TCG TGA ATT GAT ACT TTC
matK r-2	CCA CGG AAG TAT TCA TTC G

a Big Dye Terminator Cycle Sequencing Ready Reaction Kit (ABI, Foster City, California).

Identification of the matK *start codon*

To ensure that all sequences being compared were for the coding region of *matK* only and did not include *trnK* intron sequences, sequences were examined in each of the three possible reading frames to determine the correct frame. The correct reading frame was assumed to be the one that provided the longest stretch of amino acids before encountering any stop codons. This frame was then used to search for ATG start codons that corresponded to consensus start codons as found in sequences for species of the same genera on GenBank and in previous studies (Magallón and Hilu, 2009).

Determination of edited sites

Based on overlapping sequence fragments, consensus sequences were determined for both DNA and cDNA samples which were then aligned in QuickAlign (Müller and Müller, 2003). Any instances in which the DNA and cDNA sequences did not match were noted and the corresponding chromatograms were examined to ensure the correct nucleotide had been recorded. A sequence was considered to have been edited if at least 3 sequences (for both DNA and cDNA) clearly and consistently showed the nucleotide in question to be different.

Dataset alignment and phylogenetic analyses

Two *matK* datasets were generated, 1) DNA sequences only and 2) DNA sequences for un-edited species and cDNA sequences for those species that have undergone RNA editing. Both datasets were uploaded to the CIPRES portal (www.phylo.org) for alignment in MUSCLE (Edgar, 2004). Datasets were not adjusted manually but followed an alignment in MUSCLE to ensure that any differences observed in the phylogenetic trees were due to differences in the datasets themselves, not potential subjectivity from

the manual adjustment of alignments. Maximum-likelihood (ML) analyses using RAxML (Stamitakis et al., 2008) under the GTR+CAT model were executed through the CIPRES portal (www.phylo.org) with the maximum-likelihood search option selected and non-parametric bootstrapping of 1,000 replicates. A 50% majority rule consensus tree was subsequently generated in PAUP* version 4.0b10 (Swofford, 2003) to obtain bootstrap support (BS) values.

Results

Evidence of RNA editing

Based on the sampling used in this study, we found no evidence of RNA editing among members of the Gnetales for the *matK* gene. Neither was there editing detected in *Ginkgo biloba* nor *Cycas revoluta*. RNA editing was apparent, however, in those species sampled from the Zamiaceae and Pinaceae (Fig 4.1). In all cases, editing followed the most common type reported for the chloroplast genome, namely C to U changes. Three of the species exhibited five edited sites, *Dioon edule*, *Pinus sylvestris*, and *Pinus thunbergii*; *Picea omorika* had three edited sites, while *Zamia fiseheri* showed the least number of edits with only two (Fig 4.1). The two edited nucleotides of *Zamia* (codon 176 and 435) matched two of the five nucleotides that were edited in *Dioon*. The first three edited sites of *Dioon* correspond to the second codon position, while the last two are for first codon positions (Fig 4.1). In each case, a nonsynonymous mutation resulted in amino acid changes from proline to leucine (P to L; codon 105), threonine to isoleucine (T to I; codon 155), serine to leucine (S to L; codon 176), histamine to tyrosine (H to Y; codon 416), and proline to serine (P to S; codon 435), respectively. Among members of the Pinaceae, all three were shown to have one of the edited sites in common, at codon number 175 (Fig 4.1). Again, there is a mix of 1st and 2nd codon position site edits with changes from serine to leucine (S to L), proline to leucine (P to L), arginine to tryptophan (R to W), and serine to phenylalanine (S to F; Fig 4.1).

Phylogenetic analyses

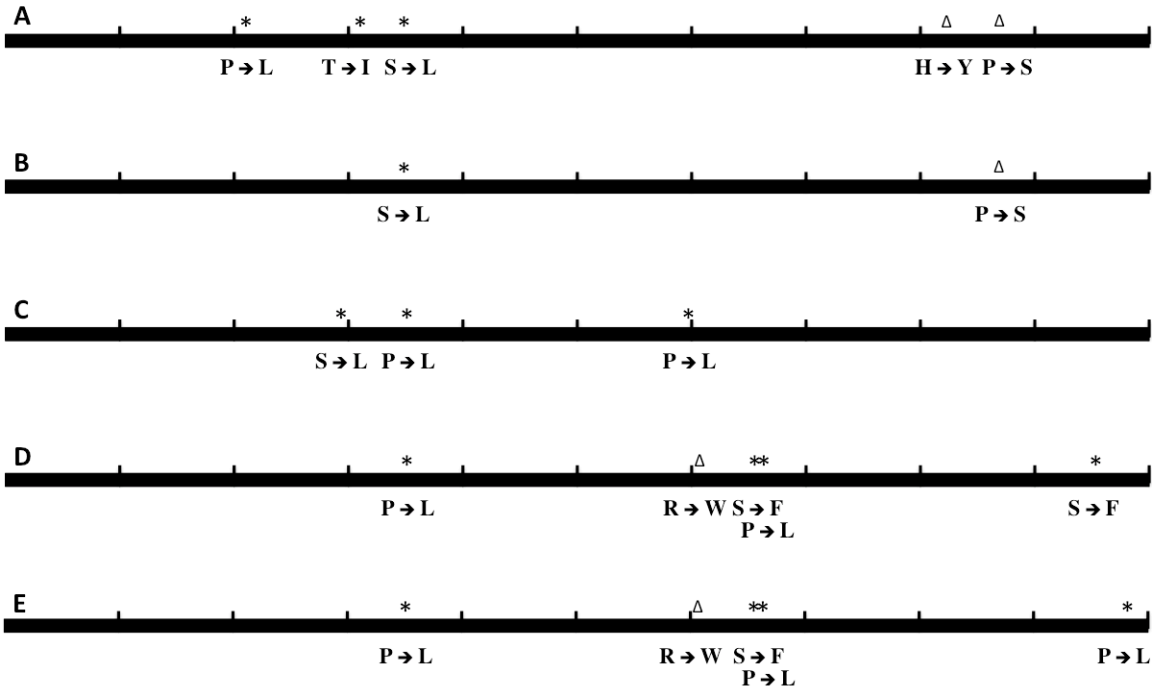

Figure 4.1 Schematic of the *matK* gene indicating where RNA editing events are located in each of the species. The gene is approximately 500 amino acids long, each marked segement above corresponds to 50 amino acids. A * indicates that the editing event takes place in the second codon position while the Δ denotes a chage in the first codon position. A. *Dioon edule* B. *Zamia fisheri* C. *Picea omorkia* D. *Pinus sylvestris* E. *Pinus thunbergii*.

The analysis of the dataset corresponding to DNA sequences alone recovered each family as monophyletic units with 100% BS support (Fig 4.2a). The two accessions of *Gnetum gnemon* were sister to one another with 100% BS and the other two *Gnetum* species were recovered as sisters with 85% BS support (Fig. 4.2a). However, within Ephedraceae, the species with multiple accessions did not group together, and BS support was ≤50% for all relationships within the family with the exception of the 53% BS support excluding *E. nevadensis* and *E. viridis* from remaining *Ephedra* species (Fig 4.2a).

When the cDNA sequences were used in place of the DNA sequences for the five species with editing, again the monophyly all families received 100% BS support. Relationships in *Gnetum* showed *G. gnemon* (VT) + *G. gnemon* (DUKE) with 100% BS support and G. leyboldii + G. montanum had a slight increase in support to 86% BS (Fig. 4.2b). Similar pattern of divergence was shown for basal *Ephedra* species with *E. nevadensis* emerging first followed by *E. viridis* and then remaining *Ephedra* (excluded by 54% BS; Fig. 4.2b). Other relationships within *Ephedra* differed slightly but support was ≤51% BS for these changes (4.2b).

Discussion

RNA editing of Gnetales

The Gnetales are an enigmatic group with regards to their phylogenetic placement, biological features and their distinct characteristics of the *matK* gene. The relatively higher proportion of non-polar amino acids in *matK* for this group corresponding to an additional predicted transmembrane domain (Barthet 2006; Barthet and Hilu 2008), suggest that this gene might behave differently within Gnetales. We thus undertook the investigation of RNA editing of *matK* for the three families that make up this order; Gnetaceae, Ephedraceae, and Welwitschiaceae. While we did not detect any RNA editing within the order, evidence of editing was found for members of Zamiaceae and

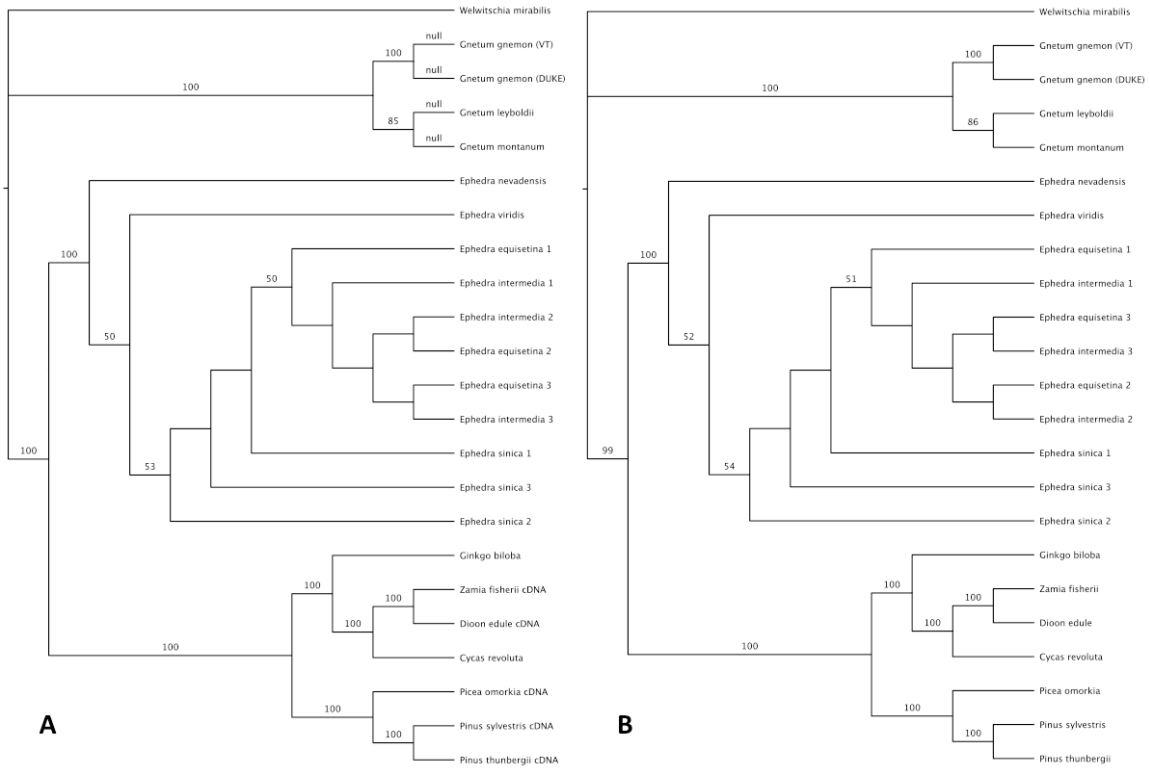

Figure 4.2 Maximum likelihood trees of the Gnetales and other gymnosperm species. BS support values noted above the branches. **A.** Dataset comprised of DNA sequences only. **B.** Dataset comprised of both DNA and cDNA sequences for those species shown to undergo RNA editing.

Pinaceae suggesting that editing of the *matK* transcript has been lost within the Gnetales. The lack of RNA editing of *matK* for these species while being present in near relatives supports the unique nature of this gene with the Gnetales.

Although there was no evidence of RNA editing detected among the Gnetales studied here, we did note an interesting feature of *matK* in Ephedraceae. Our sequencing of three individuals of *Ephedra equisetina* revealed that two of the three appeared to have a 38 base pair (bp), frame-shifting, insertion or deletion (indel) 62 nucleotides downstream from the consensus ATG that was not present in other members of *Ephedra* sampled here (Fig. 4.3). This indel is an exact repeat of 38 bp found in all *Ephedra* sampled. A search of *matK*/*trnK* sequences for Ephedraceae on GenBank revealed two other sequences with the same 38 bp insertion. A single repeat was found in *Ephedra intermedia* (AB453797) and a double repeat in a second accession of *Ephedra intermedia* (AB453798). Since this indel is 38 bp in length (76 bp for the double repeat), when the consensus ATG is used, these sequences are riddled with premature stop codons with the first occurring in the 46th amino acid position (Fig 4.3). If, however, the ATG occurring at the very start of the 38 bp section (in species without the repeated indel), 62 nucleotides downstream from the consensus ATG is used, the sequence returns to the proper frame and a full-length protein is encoded (Fig 4.3). In those species that have either a single or double repeat of the indel however, this newly described ATG corresponds to a TTG (Leu) codon (Fig 4.3). It has been shown that TTG can be recognized as a start codon in yeast (Yoon and Donahue, 1992; Touriol et al., 2003) and therefore has the potential to be a viable start codon in these species of *Ephedra*. Aside from the unique indel in these two species of *Ephedra* (*E. equisetina* and *E. intermedia*), the remaining nucleotide sequence is highly similar to other *Ephedra* species and should therefore not be considered a pseudogene, despite the TTG start codon.

RNA editing of other species

We did not detect any RNA editing of *matK* in either *Ginkgo bioloba* or *Cycas revoluta*. Partial editing of a single codon with the *matK* transcript has been shown in *Cycas*

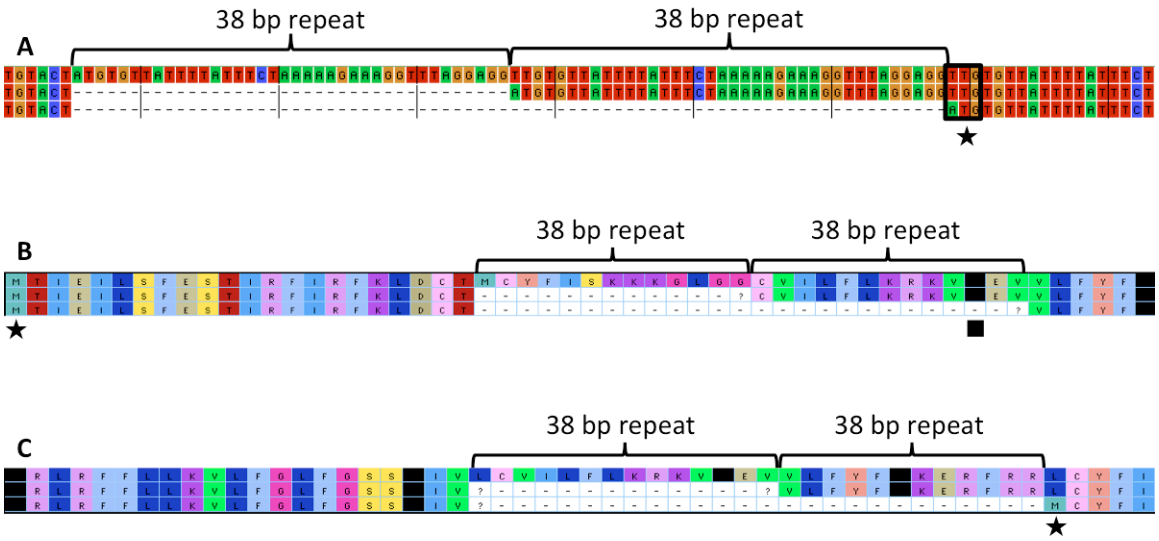

Figure 4.3 Diagram of the 38 base pair (bp) indel in some species of *Ephedra*. **A.** The consensus ATG is 62 bp upstream of the proposed alternate ATG (★). **B.** Amino acid translation of the DNA sequence for *Ephedra* when the consensus ATG (★) is used. Note the presense of stop codons (black boxes) 46 and 54 amino acids downstream of the start. **C.** Amino acid translation of the DNA sequence for *Ephedra* when the proposed alternate ATG (★) is used. Note that in species with the indel, this codes for Leu, not Met.

taitungensis (Chen et al., 2011). Close re-examination of our sequences of *Cycas revoluta* confirmed the lack of editing at this same codon or any other codon within the sequence. We did note editing of *matK* in both species of Zamiaceae and all three Pinaceae species studied here.

Within the Zamiaceae, *Zamia fisherii* was shown to have two sites undergo RNA editing, at codons 176 (S to L) and 435 (P to S), both as a result of a C to T change, at the second codon position of codon 176 and at the first codon position for codon 435 (Fig 4.1). In both cases the change results in an amino acid that is unique to *Zamia fisherii*, not matching other members of the Zamiaceae. *Dioon edule* was shown to have five sites that are edited; codons 105 (P to L), 155 (T to I), 176 (S to L), 416 (H to Y), and 435 (P to S; Fig. 4.1). Each amino acid change is the result of a C to T nucleotide change, and each results in an amino acid that is unique amongst Zamiaceae, but the changes at codons 105 and 416 result in amino acids that are homologous with those of the Cycadaceae and *Ginkgo*. This might indicate that there is RNA editing happening in other members of Zamiaceae to restore these codons to the more ancient state, or that editing of these sites evolved after the divergence of Zamiaceae from other cycads.

Slightly higher levels of editing were noted in the Pinaceae, with *Picea omorkia* having 3 sites edited and both *Pinus sylvestris* and *Pinus thunbergii* having 5 sites edited (Fig 4.1). Of the 13 sites edited within Pineaceae, one of them is common to all three species studied, located at codon 175 where a C changed to a T in the second codon position resulted in a P to L amino acid change (Fig. 4.1). The resulting leucine is unique to these species (unless other members of Pinaceae also undergo editing at this site). The other two sites that are edited within *Picea omorika* correspond to C to T nucleotide changes at the second codon position of codons 147 and 299 resulting in S to L and P to L amino acid changes, respectively (Fig. 4.1). This two edits result in leucines that appear within other members of *Pinus* but not other members of *Picea*. Of the four sites unique to the two *Pinus* species studied here, three of them are shared by both *P. sylvestris* and *P. thunbergii* (codons 302 R to W, 327 S to P, and 328 P to L; Fig. 4.1). The change at codon 302 is due to a first codon position change while the other two take place in the

second codon position. The first two changes (codons 302 and 327) appear to result in an amino acid that is not shared with other members of the Pinaceae. The change at the third (codon 328) however, results in a leucine that is homologous with members of *Picea*, although it is as the result of a different codon (CTA in *Pinus* and TTA in *Picea*). The only change that is unique to *P. sylvestris* is in codon 473 (S to F), resulting from a C to T change in the second codon position that restores the consensus phenylalanine. The fifth edited site of *P. thunbergii* is in codon 492 (P è L), resulting from a C to T change in the second codon position, and producing a leucine that is not shared with other members of Pinaceae. Over half of the total changes noted result in an amino acid change to leucine and a majority of total changes (85%) result in hydrophobic amino acids. The remaining 15% of changes are to amino acids that are neutral. One half of the amino acids that are changed originally code for proline residues which are hydrophobic, so there is very little change in the behavior of the amino acid chain as a result of the RNA editing.

The exclusive observance of C to T nucleotide edits is very consistent with the pattern of RNA editing observed in plant chloroplast genes (Maier et al., 1996; Freyer et al., 1997; Bock 2000; Tillich 2006) as is the tendency to see edits in the first and second codon positions (Freyer et al., 1997). It is interesting to note that a majority of the amino acid changes noted here did not result in an amino acid matching the consensus sequence, but instead departed from the consensus to produce a novel amino acid. This result goes against the expectation that RNA editing should restore the amino acid to the corresponding consensus sequence, and might indicate that the widely used method of predicting RNA editing sites based on this fact is missing a large proportion of RNA edited sites (Freyer et al., 1997; Chaw et al., 2008; Cuenca et al., 2008; Diekmann et al., 2009; Chen et al., 2011).

RNA editing and phylogeny reconstruction

Bowe and dePamphilis (1996) indicated that RNA editing itself doesn't pose a problem in phylogenetic reconstruction, but that contend that cDNA and DNA sequences should

not be combined due to the potential for signal from edited sites to outweigh signal across the dataset. Duffy et al. (2009) indicate that they changed several nucleotides in their alignment to match known RNA editing sites, but they do not explain the reason for their choice. The results of our analyses based on *matK* sequences for 23 species, five of which contain sites subject to a total of 20 RNA editing events, indicate that these changes do not impact phylogenetic reconstruction. The ML analyses based on either DNA data alone or combined DNA/cDNA data resulted in identical phylogenies for nodes with >50% BS support and with very little impact on BS support values >50% (Fig. 4.2). Based on our analyses, the presence of RNA editing did not impact phylogenetic reconstruction in terms of topology or support.

Based on the evidence presented above for an alternate start codon for species of *Ephedra*, the two datasets were re-analyzed using the proposed ATG/TTG start for *matK* instead of the consensus ATG for all members of *Ephedra*. The resulting phylogenies in all cases were congruent with those obtained when the consensus ATG was used (Figs. 4.2 and 4.4). However, there is notable increase in BS support for relationships within Ephedraceae and slight decrease in support for the sister relationship of *Gnetum leyboldii* to *Gnetum montanum* (85% BS to 79% BS for DNA dataset, 86% BS to 84% BS for combined DNA/cDNA dataset; Figs. 4.2 and 4.4). In cases where the support within *Ephedra* increased, two of the three accessions for both *E. equisetina* and *E. intermedia* were shown to be sisters with support ranging from 71% - 87% BS, a drastic improvement on the <50% support received using the consensus start codon (Fig. 4.4). This increase provides additional support for the use of the proposed alternate start codon in these species.

Conclusions

Despite the unique nature of *matK* within Gnetlaes, their distinctly higher proportion of non-polar amino acids in *matK* corresponding to the prediction of a unique transmembrane domain, no evidence of RNA editing was detected for members of this group. We did, however, note an interesting pattern of *matK* evolution for members of

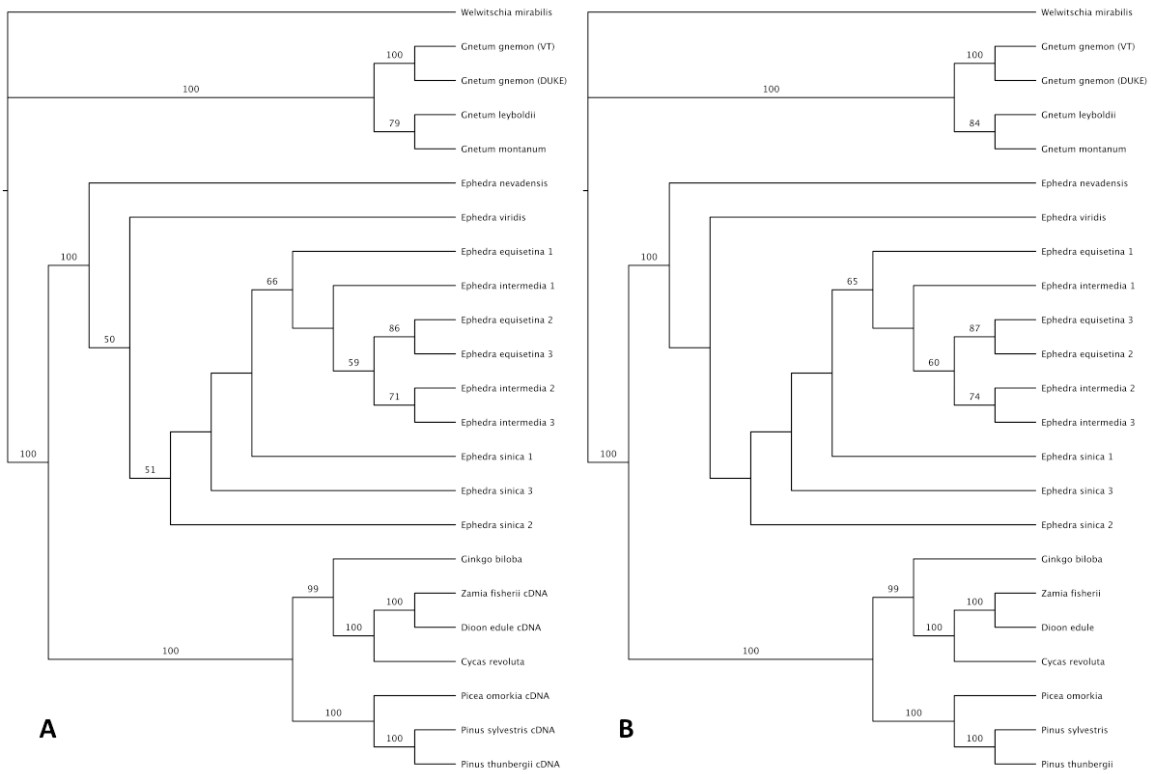

Figure 4.4 Maximum likelihood trees of the Gnetales and other gymnosperm species. In the datasets used to generate these trees, the alternate ATG for *Ephedra* has been used in place of the consensus ATG. BS support values noted above the branches. **A.** Dataset comprised of DNA sequences only. **B.** Dataset comprised of both DNA and cDNA sequences for those species shown to undergo RNA editing.

Ephedra. Members of this genus have been shown to contain a 38 bp repeat, sometimes with two repeats in tandem that appear to alter the open reading frame of the gene resulting in pre-mature stop codons. We propose an alternate ATG start codon for species lacking the repeat, and a TTG start codon for those shown to contain the repeat. Additionally, we provide evidence here for 20 RNA editing sites of *matK* for 5 species within the Zamiaceae and Pinaceae, most of which appear to result in an amino acid different from the consensus sequence suggesting that searches for RNA editing sites based on consensus sequence restoration might be missing a large proportion of actual editing. Our phylogenetic analysis of datasets based on DNA alone or combined DNA/cDNA sequences show that there is no impact of RNA editing sites on topology or support for the species studied here.

Acknowledgements

We would like to thank Adrianna Feriolli and Emily Steele for their assistance with specimen collection and lab work. This work was supported by grants from the Virginia Tech Graduate Research Development Program to SSC.

References:

Barthet, M. M., 2006: Expression and function of the chloroplast-encoded gene *matK*, Biological Sciences, Virginia Polytechnic Institute and State University, Blacksburg. Pages pp.

Barthet, M. M. and Hilu, K. W., 2008: Evaluating evolutionary constraint on the rapidly evolving gene *matK* using protein composition. *J Mol Evol,* 66: 85-97.

Bock, R., 2000: Sense from nonsense: how the genetic information of chloroplasts is altered by RNA editing. *Biochimie,* 82: 549-557.

Bowe, L. M. and dePamphilis, C. W., 1996: Effects of RNA editing on gene processing on phylogenetic reconstrcution. *Mol Biol Evol,* 13: 1159-1166.

Bowe, L. M., Coat, G., and dePamphilis, C. W., 2000: Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and Gnetales closest relatives are conifers. *Proc Natl Acad Sci U S A,* 97: 4092-4097.

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

Burleigh, J. G. and Mathews, S., 2004: Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. *American Journal of Botany,* 91: 1599-1613.

Cadotte, M. W., Hamilton, M. A., and Murray, B. R., 2009: Phylogenetic relatedness and plant invader success arcoss two spatial scales. *Diversity and Distributions,* 15: 481-488.

Carlquist, S., 1996: Wood, bark, and stem anatomy of Gnetales: a summary. *International Journal of Plant Sciences,* 157: S58-S76.

Chaw, S.-M., Parkinson, C. L., Cheng, Y., Vincent, T. M., and Palmer, J. D., 2000: Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. *Proc Natl Acad Sci U S A,* 97: 4086-4091.

Chaw, S.-M., Shih, A. C.-C., Wang, D., Wu, Y.-W., Liu, S.-M., and Chou, T.-Y., 2008: The mitochondrial genome of the gymnosperm *Cycas taitungensis* contains a novel family of short interspersed elements, Bpu sequences, and abundata RNA editing sites. *Mol Biol Evol,* 25: 603-615.

Chen, H., Deng, L., Jiang, Y., Lu, P., and Yu, J., 2011: RNA Editing Sites Exist in Protein-Coding Genes in the Chloroplast Genome of Cycas Taitungensis. *Journal of Integrative Plant Biology*: no-no.

Cuenca, A., Petersen, G., Seberg, O., Davis, J. I., and Stevenson, D. W., 2008: Are substitution rates and RNA editing correlated? *BMC Evolutionary Biology,* 10: 349.

Diekmann, K., Hodkinson, T. R., Wolfe, K. H., van den Bekerom, R., Dix, P. J., and Barth, S., 2009: Complete chloroplast genome sequence of a major allogamous

forage species, perennial ryegress (*Lolium perenne* L.). *DNA Research,* 16: 165-176.

Donoghue, M. J. and Doyle, J. A., 2000: Seed plant phylogeny: Demise of the anthophyte hypothesis? *Current Biology,* 10: R106-R109.

Doyle, J. J. and Doyle, J. L., 1990: Isolation of plant DNA from fresh tissue. *Focus,* 12: 13 - 25.

Duffy, A. M., Kelchner, S. A., and Wolf, P. G., 2009: Conservation of selection on *matK* follwoing an ancient loss of its flanking intron. *Gene,* 438: 17-25.

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

Freyer, R., Kiefer-Meyer, M.-C., and Kossel, H., 1997: Occurrence of plastid RNA editing in all major lineages of land plants. *Proc Natl Acad Sci U S A,* 94: 6285-6290.

Friedman, W. E. and Carmichael, J. S., 1996: Double Fertilization in Gnetales: Implications for Understanding Reproductive Diversification among Seed Plants. *International Journal of Plant Sciences,* 157: S77-S94.

Graham, S. W. and Olmstead, R. G., 2000: Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *American Journal of Botany,* 87: 1712-1730.

Gugerli, F., Sperisen, C., Büchler, U., Brunner, I., Brodbeck, S., Palmer, J. D., and Qiu, Y.-L., 2001: The evolutionary split of Pinaceae from other conifers: evidence from an intron loss and multigene phylogeny. *Molecular Phylogenetics and Evolution,* 21: 167-175.

Hajibabaei, M., Xia, J., and Drouin, G., 2006: Seed plant phylogeny: Gnetophytes are derived and a sister group to Pinaceae. *Molecular Phylogenetics and Evolution,* 40: 208-217.

Harrington, M. G., Edwards, K. J., Johnson, S. A., Chase, M. W., and Gadek, P. A., 2005: Phylogenetic inference in Sapindaceae sensu lato using plastid *matK* and *rbcL* DNA sequences. *Systematic Botany,* 30: 366-382.

Hilu, K. W., Borsch, T., Müller, K., Soltis, D. E., Soltis, P. S., Savolainen, V., Chase, M. W., Powell, M. P., Alice, L. A., Evans, R., Sauquet, H., Neinhuis, C., Slotta, T. A.

B., Jens, G. R., Campbell, C. S., and Chatrou, L. W., 2003: Angiosperm phylogeny based on *matK* sequence information. *American Journal of Botany,* 90: 1758 - 1776.

Hilu, K. W., Black, C., Diouf, D., and Burleigh, J. G., 2008: Phylogenetic signal in *matK* vs. *trnK*: a case study in early diverging eudicots (angiosperms). *Molecular Phylogenetics and Evolution,* 48: 1120-1130.

Inada, M., Sasaki, T., Yukawa, M., Tsudzuki, T., and Sugiura, M., 2004: A systematic search for RNA editing sites in pea chloroplasts: an editing event causes diversification from the evolutionarily conserved amino acid sequence. *Plant Cell Physiol,* 45: 1615-1622.

Johnson, L. A. and Soltis, D. E., 1995: Phylogenetic inference in Saxifragaceae sensu stricto and *Gilia* (Polemoniaceae) using *matK* sequences. *Annals of the Missouri Botanical Gardens,* 82: 149 - 175.

Magallón, S. and Sanderson, M. J., 2002: Relationships among seed plants inferred from highly conserved genes: sorting conflicting phylogenetic signal among ancient lineages. *American Journal of Botany,* 89: 1991-2006.

Magallón, S. and Hilu, K. W., 2009: Land plants (Embryophyta). *In* Hedges, S. B. and Kumar, S. (eds.), *The Timetree of Life*: Oxford University Press, 133-137.

Maier, R. M., Zeltz, P., Kössel, H., Bonnard, G., Gualberto, J. M., and Grienenberger, J. M., 1996: RNA editing in plant mitochondria and chloroplasts. *Plant Molecular Biology,* 32: 343-365.

Mathews, S., 2009: Phylogenetic relationships among seed plants: persistent questions and the limits of molecular data. *American Journal of Botany,* 96: 228-236.

Mort, M. E., Randle, C. P., Kimball, R. T., Tadesse, M., and Crawford, D. J., 2008: Phylogeny of Coreopsideae (Asteraceae) inferred from nuclear and plastid DNA sequences. *Taxon,* 57: 109-120.

Müller, J. and Müller, K., 2003: QuickAlign: a new alignment editor. *Plant Molecular Biology Reporter,* 21: 5.

Müller, K. F., Borsch, T., and Hilu, K. W., 2006: Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: contrasting *matK*, *trnT-F* and *rbcL* in basal angiosperms. *Molecular Phylogenetics and Evolution,* 41: 99 - 117.

Rai, H. S., Reeves, P. A., Peakall, R., Olmstead, R. G., and Graham, S. W., 2008: Inference of higher-order conifer relationships from a multi-locus plastid data set. *Botany,* 86: 658-669.

Soltis, P. S., Soltis, D. E., Savolainen, V., Crane, P. R., and Barraclough, T. G., 2002: Rate heterogeneity among lineages of tracheophytes: integration of molecular and fossil data and evidence for molecular living fossils. *PNAS,* 99: 4430-4435.

Stamatakis, A., Hoover, P., and Rougemont, J., 2008: A Fast Bootstrapping Algorithm for the RAxML Web Servers. *Systematic Biology,* 57: 758-771.

Swofford, D. L., 2003: *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*. Sinauer Associates, Sunderland, Massachusetts, USA.

Tillich, M., Lehwark, P., Morton, B. R., and Maier, U. G., 2006: The evolution of chloroplast RNA editing. *Mol Biol Evol,* 23: 1912-1921.

Touriol, C., Bornes, S., Bonnal, S., Audigier, S., Prats, H., Prats, A.-C., and Vagner, S., 2003: Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. *Biology of the Cell,* 95: 169-178.

Vogel, J., Hubschmann, T., Borner, T., and Hess, W. R., 1997: Splicing and intron-internal RNA editing of trnK-matK transcripts in barley plastids: support for MatK as an essential splicing factor. *Journal of Molecular Biology,* 270: 179 - 187.

Wakasugi, T., Hirose, T., Horihata, M., Tsudzuki, T., Kössel, H., and Sugiura, M., 1996: Creation of a novel protein-coding region at the RNA level in black pine chloroplasts: the patter of RNA editing in the gymnosperm chloroplast is different from that in angiosperms. *Proc Natl Acad Sci U S A,* 93: 8766-8770.

Wolf, P. G., Rowe, C. A., Sinclair, R. B., and Hasebe, M., 2003: Complete nucleotide sequence of the chloroplast genome from a leptosporangiate fern, *Adiantum capillus-veneris* L. *DNA Research,* 10.

Yoon, H. and Donahue, T. F., 1992: The *sui1* suppressor locus in *Saccharomyces cerevisiae* encodes a translation factor that functions during tRNAiMet recognition of the start codon. *Molecular and Cellular Biology,* 12: 248-260.

Chapter 5: Conclusion

The ongoing effort to discern the evolutionary origins, and phylogenetic relationships and classification of all living organisms has lead to varied degrees of large-scale phylogeny reconstruction. These efforts require a great deal of investment in terms of time and resources to piece together a detailed picture of the tree of life, and molecular phylogeneticists must consider the cost-effectiveness of sequence data. Using whole genomes (chloroplast and/or mitochondrial) have been used for a few species per group and thus will only provide an overall relationships (backbone) and is error prone due to relatively small number of species used. We have shown that expanding datasets by increasing the number/type of genomic regions utilized and/or increasing the taxon sampling using GenBank data for either partial or complete sequences can improve phylogenetic structure in the Caryophyllales despite the inclusion of missing data. We have shown that this approach provides a detailed picture of the order by expanding the number of species up to 652 species in this study. Such detailed phylogenetic trees can be used by a broader sector of biological scientists since it contains detailed information out to the species level. Such an approach should be tested in other biological groups in an effort to increase the details of the tree of life and enhance its accuracy.

Additionally, we have shown that the phylognetic information from the *trnK* intron that is quite often excluded in deep-level molecular phylogenetics, is comparable to *matK* in terms of proportion of variable sites, parsimony informative sites, and the distribution of those sites among rate classes. Phylogenetic analyses of the Caryophyllales based on the combined *matK*/*trnK* intron data produced a robust phylogeny comparable to those derived from multi-gene studies. Therefore, the *trnK* intron sequence data, often obtained concurrently with *matK*, should be included in phylogenetic studies even at deep historic levels. This almost doubles the amount of sequence information for slightly more cost and time than using *matK* alone. This is particularly important since *matK* alone provides phylogenetic information from sequences of as many as 3-11 genes plus the chloroplast inverted repeat combined.

We have shown that the inclusion of RNA edited sites for the plastid group II intron maturase *matK* in phylogeny reconstruction of the Gnetales and other gymnosperms had no impact on tree robustness when compared to a phylogeny resulting from analysis of DNA data only. We detected a 38 bp insertion in the *matK* ORF of some members of Ephedraceae that appear to result in a frame-shift. When the consensus ATG start codon is used, we found a number of premature stop codons that should render matK a pseudogene. We have suggested an alternate start codon for members of Ephedraceae in species containing this 38 bp indel that results in a full reading frame for the gene. The potential of the evolution of this alternate start codon is supported by the similarity in the frequency of nucleotide mutations in the alignment of the Ephedraceae across all species including those that lack the 38 bp indel.. Pseudogenes lack selectional constraints and thus its sequences accommodate large amounts of stochastic mutations. Using this alternate start codon in sequence alignment resulted in increased support for relationships with the Ephedraceae. Although no RNA editing of the *matK* transcript was detected in members of the Gnetales, a total of 20 editing sites were detected in two members of the Zamiaceae and three in Pinaceae. Therefore, it appears that RNA editing may not impact molecular phylogenetic reconstruction but may increase tree robustness by improving support for some relationships.

Appendix A

Supplemental Figures for Chapter 2

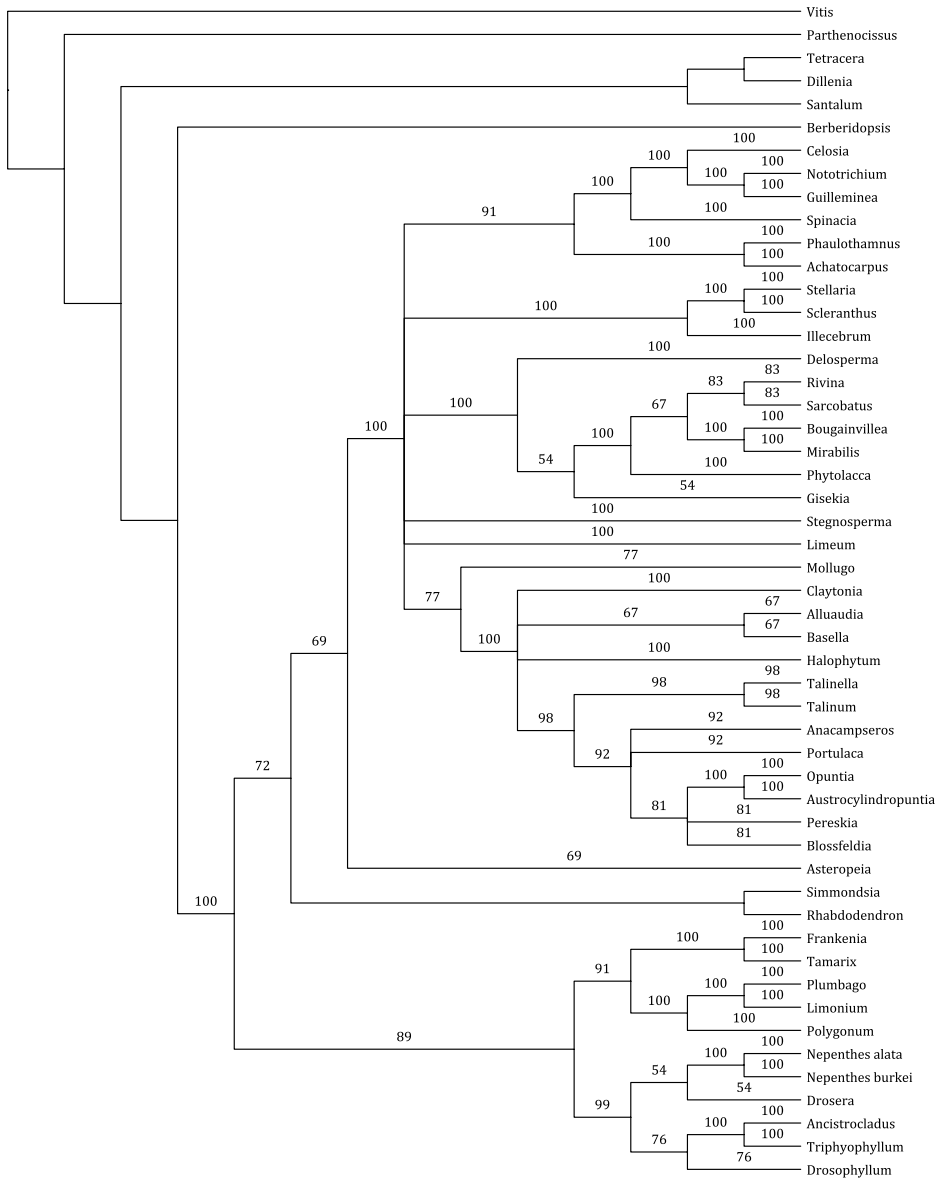

Figure A.1 MP strict consensus tree based on the *matK*/*trnK* intron dataset for 51 Caryophyllales taxa (0.3% missing data). Percent bootstrap values greater than 50% are noted on branches.

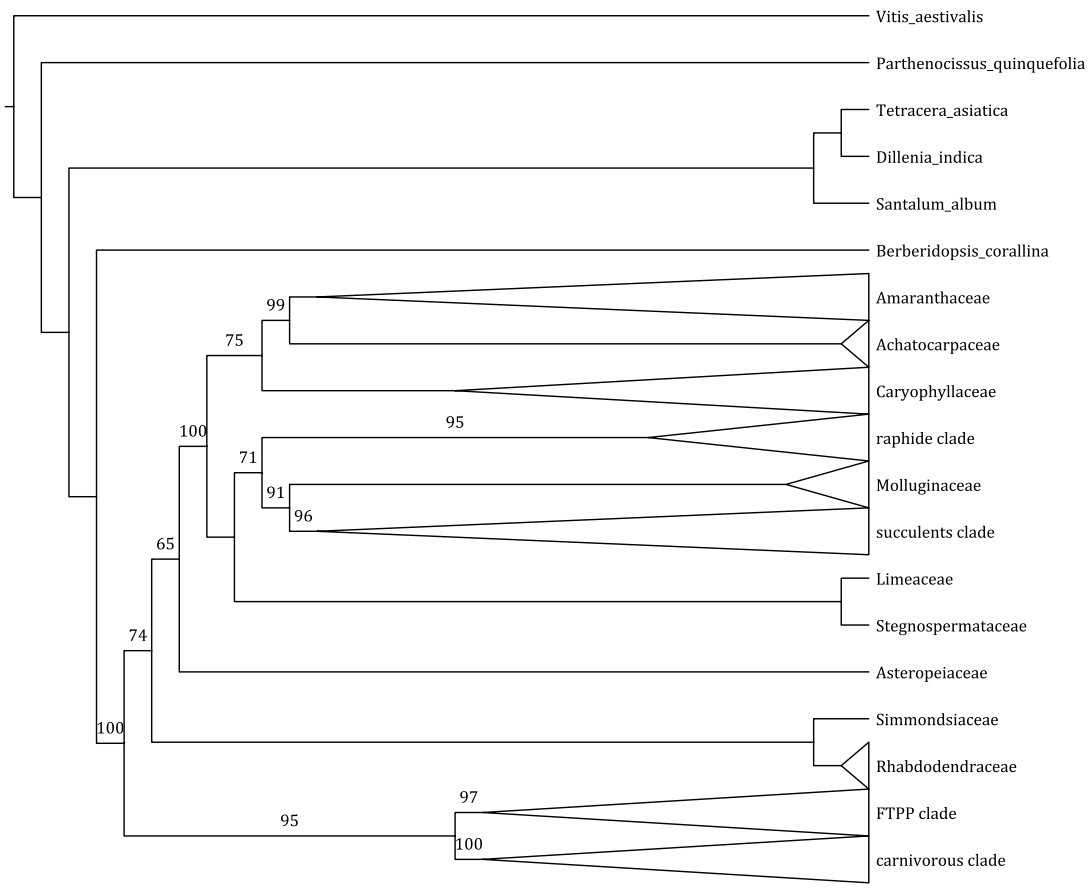

Figure A.2 Summary of the MP strict consensus tree based on *matK*/*trnK* intron data with expanded taxon sampling (652 taxa with 38% missing data). Percent bootstrap values greater than 50% are noted on branches.

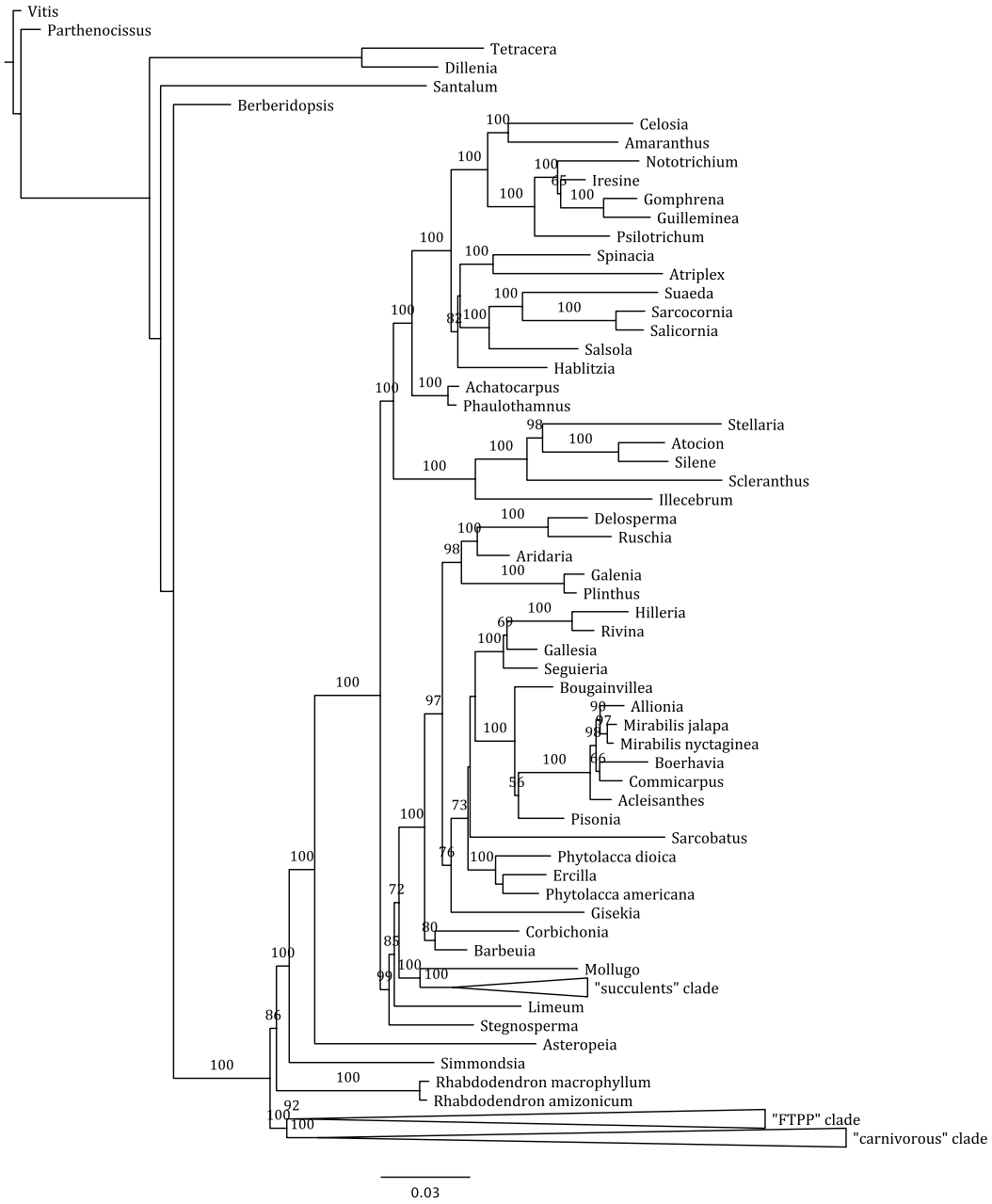

Figure A.3a ML tree based on the 5 genomic regions (*rbcL*, *atpB*, *ndhF*, *matK*, and *trnK* intron) for 136 Caryophyllales taxa. Percent bootstrap values greater than 50% are noted on braches. Expanded details for the "AAC" and "raphide" clades.

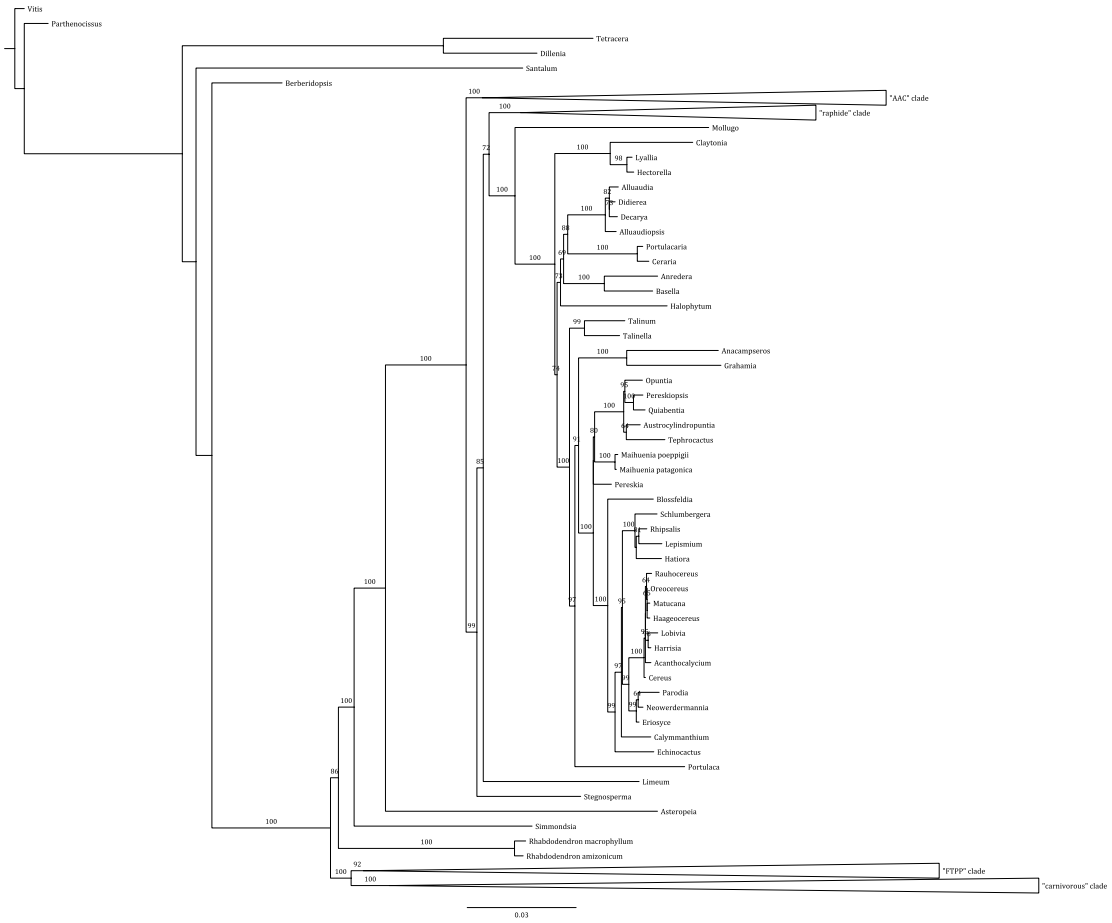

Figure A.3b ML tree based on the 5 genomic regions (*rbcL*, *atpB*, *ndhF*, *matK*, and *trnK* intron) for 136 Caryophyllales taxa. Percent bootstrap values greater than 50% are noted on braches. Expanded details for the "succulents" clade.

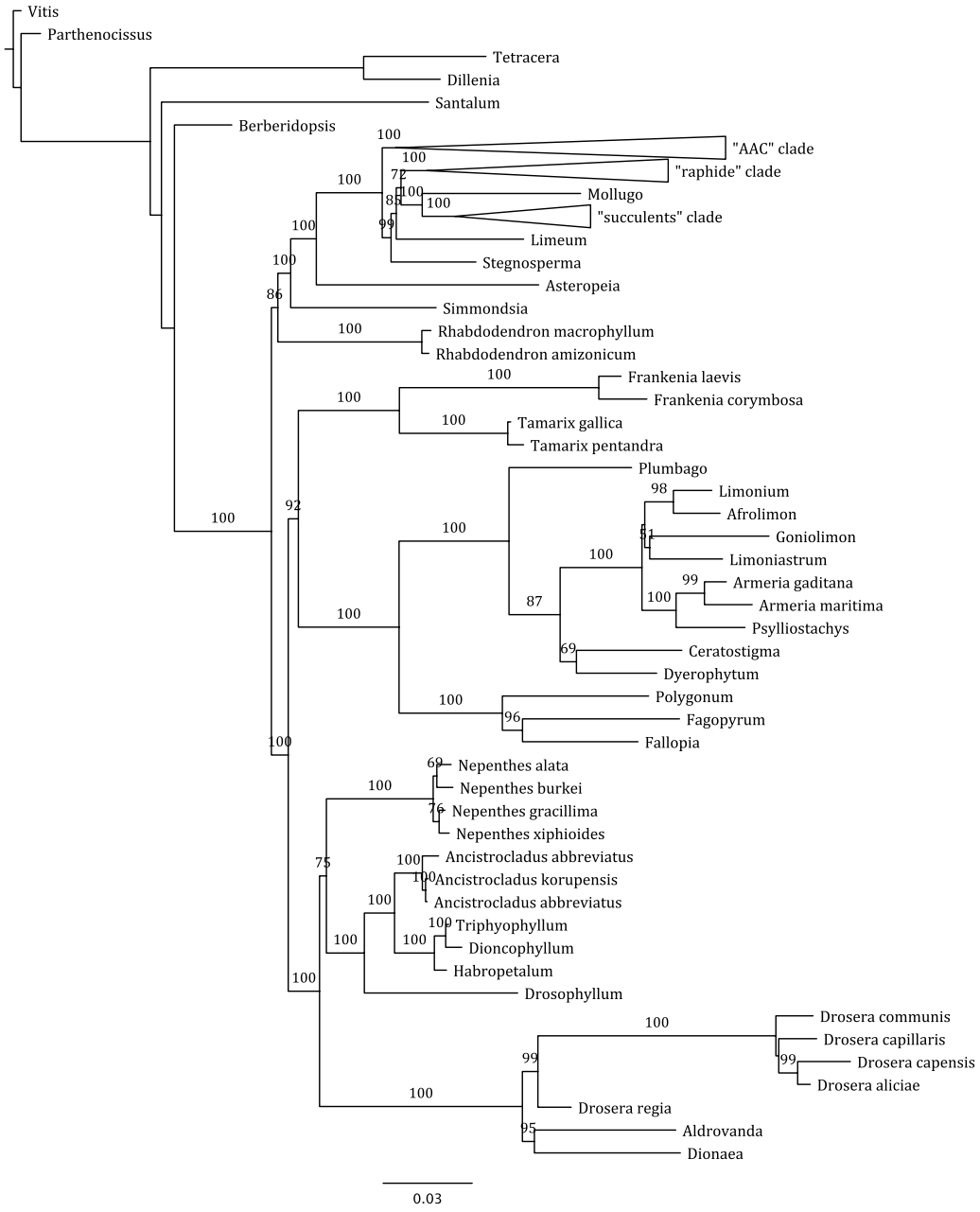

Figure A.3c ML tree based on the 5 genomic regions (*rbcL*, *atpB*, *ndhF*, *matK*, and *trnK* intron) for 136 Caryophyllales taxa. Percent bootstrap values greater than 50% are noted on braches. Expanded details for the "FTPP" and "carnivorous" clades.

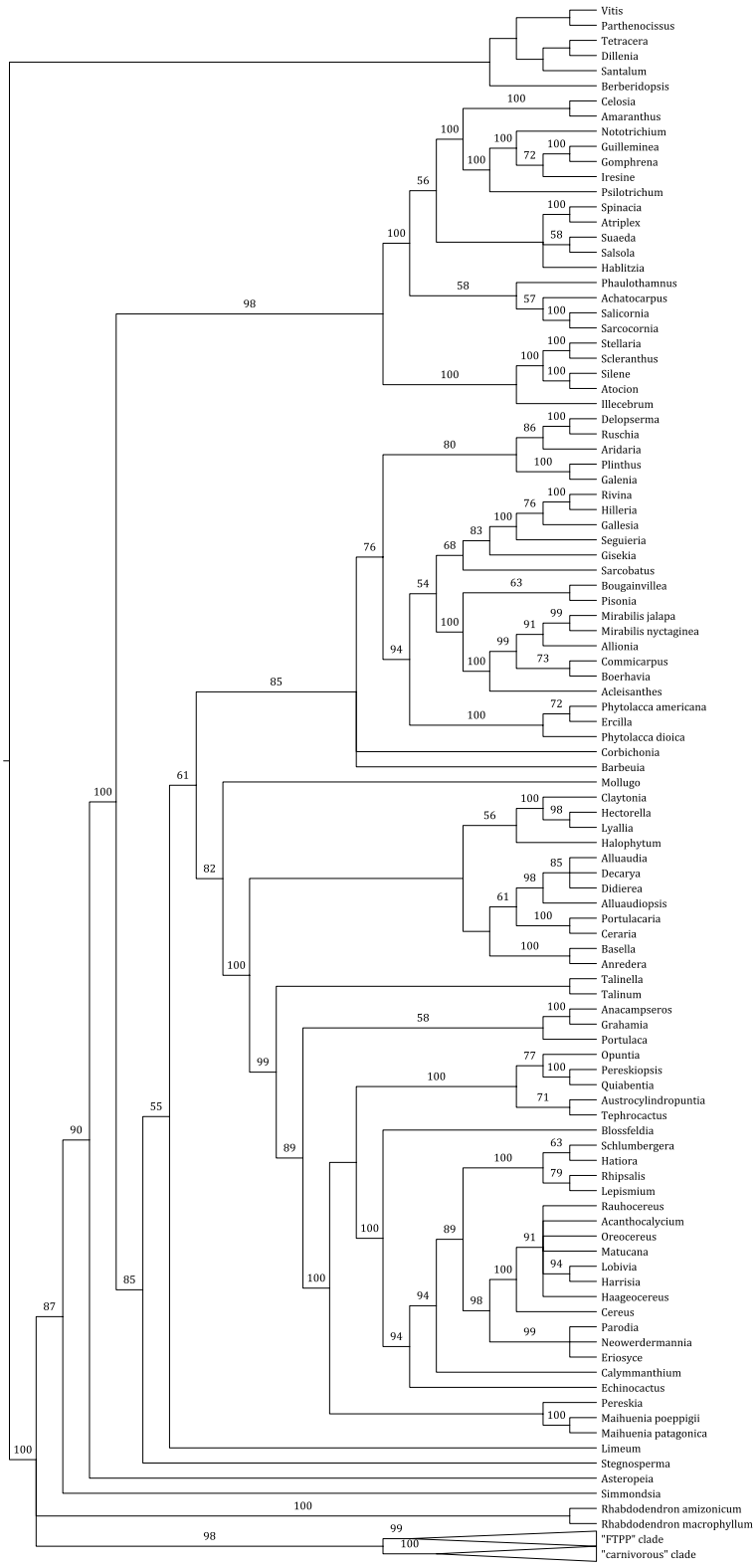

Figure A.4a MP strict consensus tree based on the dataset of five genomic regions (*rbcL*, *atpB*, *ndhF*, *matK*, and *trnK* intron) for 136 taxa (5GR-136; 46% missing data).

Percent bootstrap values greater than 50% are noted on branches. The FTPP and carnivorous clades have been collapsed.

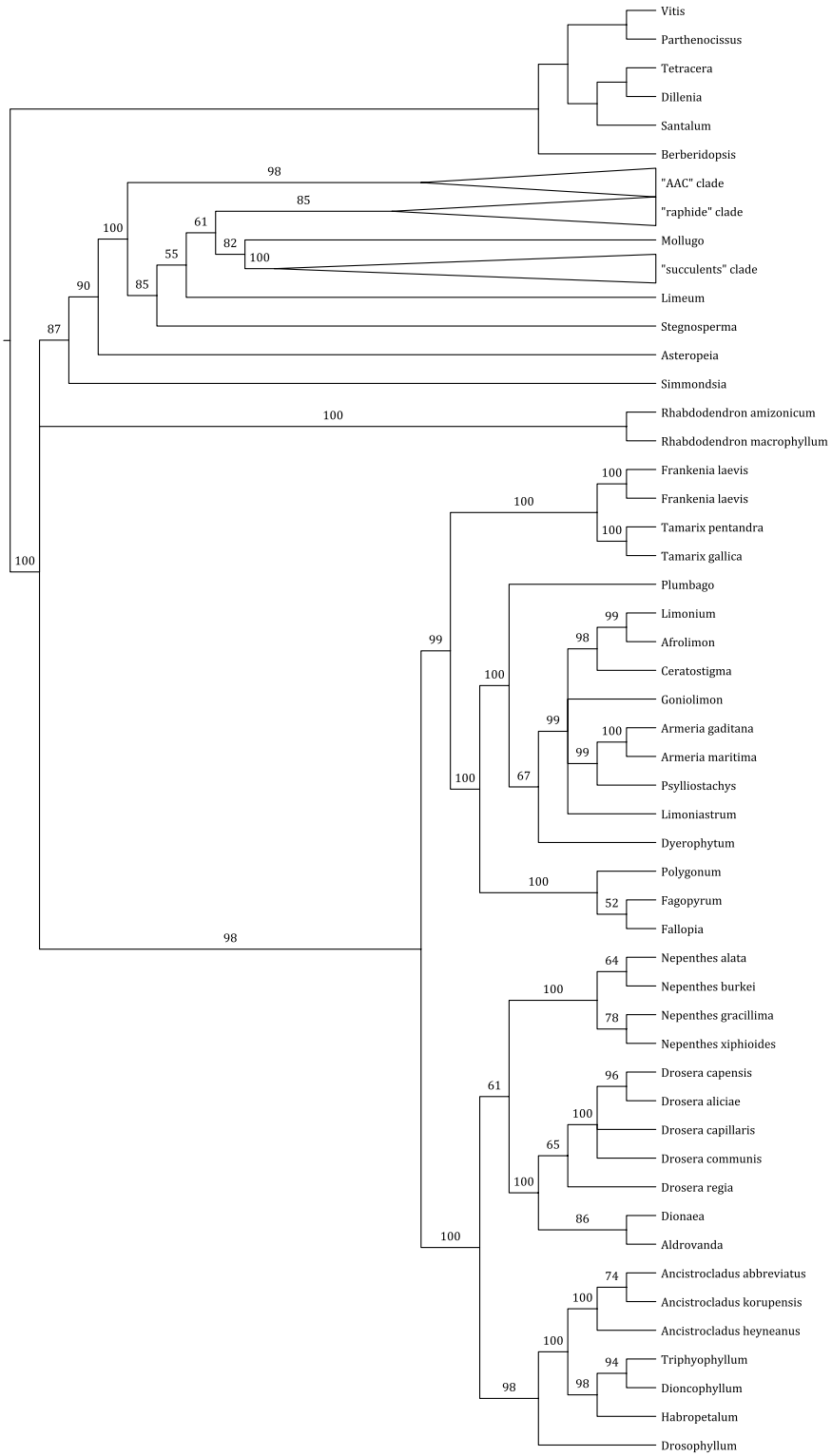

Figure A.4b MP strict consensus tree based on the dataset of five genomic regions (*rbcL*, *atpB*, *ndhF*, *matK*, and *trnK* intron) for 136 taxa (5GR-136; 46% missing data).

Percent bootstrap values greater than 50% are noted on branches. FTPP and carnivorous clades are expanded.

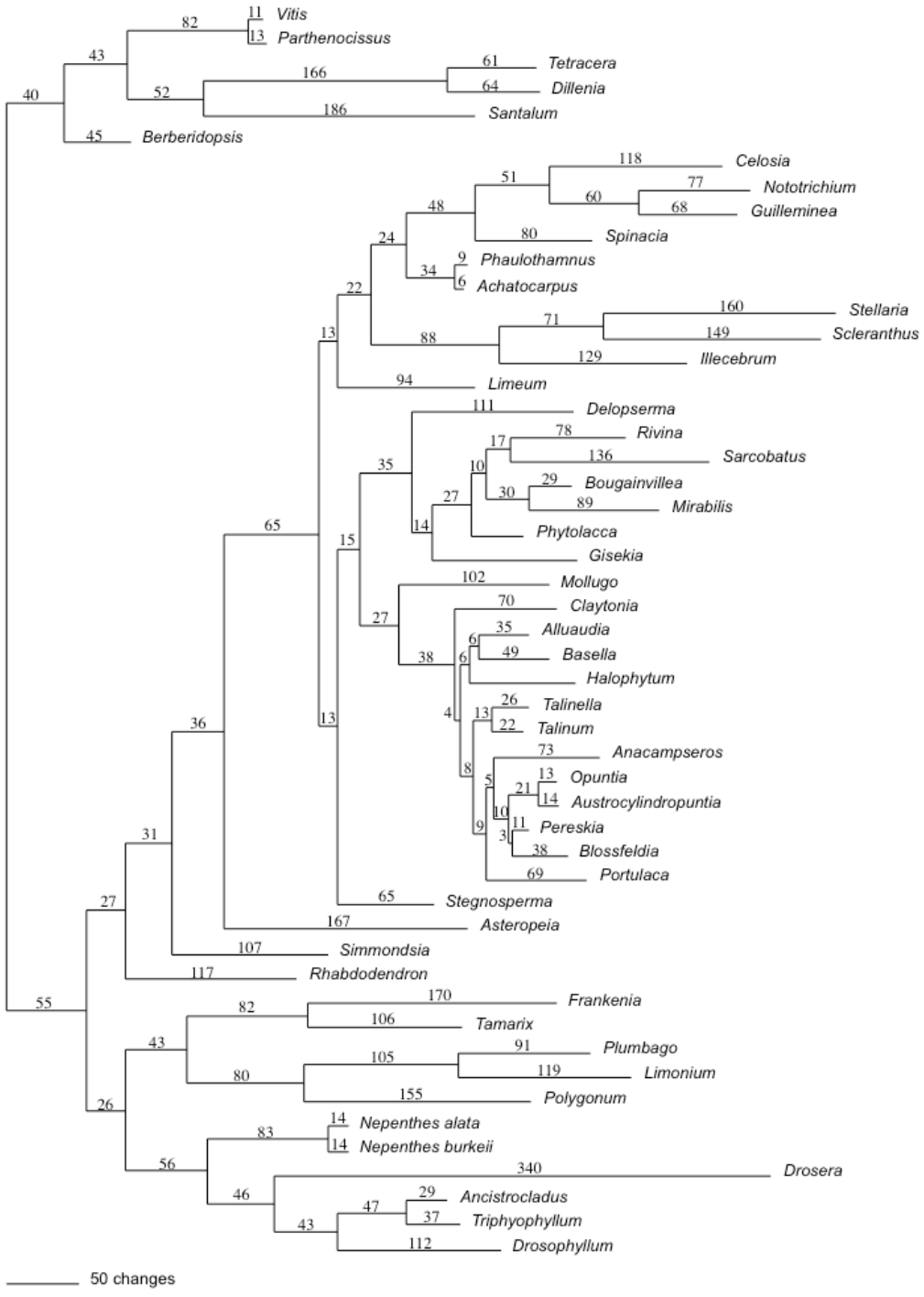

Figure A.5 ML tree based on the *matK*/*trnK* intron dataset for 51 Caryophyllales taxa. Branch lengths are noted on the branches.

Appendix B

Supplemental Table 1 for Chapter 2

Table B.1 Species used, their taxonomic affiliation, GenBank numbers, and information on sources of material for the *matK*/*trnK* dataset with 51 taxa (MT-51). "Complete" and "partial" indicates whether the sequences obtained from this study or GenBank were for the whole genomic region or for portions of it.

		GenBank No./Source	
Family	Genus and species	*matK*	*trnK* introns
Outgroup Taxa			
Berberidopsidaceae	*Berberidopsis corallina*	**this study** – partial (Qiu 97042, IND) **AY042554** – partial (Cuénoud et al. 2002)	**this study** – complete (Qiu 97042, IND)
Dilleniaceae	*Dillenia indica*	**this study** – complete (M. J. Moore, 340 FLAS)	**this study** – complete (M. J. Moore, 340 FLAS)
	Tetracera asiatica	**this study** – partial (Prance 30760, K) **AY042665** – partial (Cuénoud et al. 2002)	**this study** – complete (Prance 30760, K)
Santalaceae	*Santalum album*	**this study** – partial (D. Harbaugh 65, UC) **AY042650** –	**this study** – complete (D. Harbaugh 65, UC)

		partial (Cuénoud et al. 2002)	
Vitaceae	*Parthenocissus quinquefolia*	**this study** – complete (K. Hilu 502, VPI)	**this study** – complete (K. Hilu 502, VPI)
	Vitis aestivalis	**this study** – partial (T. Wieboldt 11,696, VPI) **AF274635** – partial (Fishbein et al. 2001)	**this study** – complete (T. Wieboldt 11,696, VPI)
Ingroup Taxa			
Achatocarpaceae	*Achatocarpus praecox*	**AY514845** – complete (Müller and Borsch 2005)	**AY514845** – complete (Müller and Borsch 2005)
	Phaulothamnus spinescens	**AY514846** – complete (Müller and Borsch 2005)	**AY514846** – complete (Müller and Borsch 2005)
Aizoaceae	*Delopserma napiforme*	**this study** – complete (S. Brockington 700, FLMNH)	**this study** – complete (S. Brockington 700, FLMNH)
Amaranthaceae	*Celosia trigyna*	**AY514811** – complete (Müller and Borsch 2005)	**AY514811** – complete (Müller and Borsch 2005)
	Guilleminea densa	**AY514803** – complete	**AY514803** – complete

		(Müller and Borsch 2005)	(Müller and Borsch 2005)
	Nototrichium sandwicense	**AY514817** – complete (Müller and Borsch 2005)	**AY514817** – complete (Müller and Borsch 2005)
	Spinacia oleracea	**AJ400848** – complete (Schmitz-Linneweber et al. 2001)	**AJ400848** – complete (Schmitz-Linneweber et al. 2001)
Ancistrocladaceae	*Ancistrocladus abbreviatus*	**AF315939** – complete (Meimberg et al. 2001)	**AF315939** – complete (Meimberg et al. 2001)
Asteropeiaceae	*Asteropeia micraster*	**this study** – partial (01849928, M. B. G.[a]) **AY042549** - partial (Cuénoud et al. 2002)	**this study** – complete (01849928, M. B. G.)
Basellaceae	*Basella alba*	**this study** – complete (Qiu 02055, MASS)	**this study** – complete (Qiu 02055, MASS)
Cactaceae	*Austrocylindropuntia subulata*	**AY875364** – complete (Edwards et al. 2005)	**AY875364** – complete (Edwards et al. 2005)
	Blossfeldia liliputana	**AY875366** – complete	**AY875366** – complete

		(Edwards et al. 2005)	(Edwards et al. 2005)
	Opuntia quimilo	**AY015279** – complete (Nyffeler 2002)	**AY015279** – complete (Nyffeler 2002)
	Pereskia aculeata	**AY875355** – complete (Edwards et al. 2005)	**AY875355** – complete (Edwards et al. 2005)
Caryophyllaceae	*Illecebrum verticillatum*	**AY514849** – complete (Müller and Borsch 2005)	**AY514849** – complete (Müller and Borsch 2005)
	Scleranthus perennis	**AY514847** – complete (Müller and Borsch 2005)	**AY514847** – complete (Müller and Borsch 2005)
	Stellaria media	**this study** – partial (A. Hinkle 398, VPI) **AY936299** – partial (Fior et al. 2006)	**this study** – complete (A. Hinkle 398, VPI)
Didiereaceae	*Alluaudia ascendens*	**this study** – partial (A. Hinkle 360, VPI) **AY042541** – partial (Cuénoud et al. 2002)	**this study** – complete (A. Hinkle 360, VPI)
Dioncophyllaceae	*Triphyophyllum*	**AF315940** –	**AF315940** –

	peltatum	complete (Meimberg et al. 2001)	complete (Meimberg et al. 2001)
Droseraceae	*Drosera capensis*	**this study** – partial (A. Hinkle 357, VPI) **AY096122** - partial (Cameron et al. 2002)	**this study** – complete (A. Hinkle 357, VPI)
Drosophyllaceae	*Drosophyllum lusitanicum*	**AY514860** – complete (Müller and Borsch 2005)	**AY514860** – complete (Müller and Borsch 2005)
Frankeniaceae	*Frankenia laevis*	**AY514853** – complete (Müller and Borsch 2005)	**AY514853** – complete (Müller and Borsch 2005)
Gisekiaceae	*Gisekia Africana*	**this study** – partial (J. Kornas 3231, K and S. Brockington 701, FLMNH) **AY042591** - partial (Cuénoud et al. 2002)	**this study** – complete (J. Kornas 3231, K and S. Brockington 701, FLMNH)
Halphytaceae	*Halophytum ameghinoi*	**AY514852** – complete (Müller and Borsch 2005)	**AY514852** – complete (Müller and Borsch 2005)
Molluginaceae	*Limeum africanum*	**this study** – complete (M. J. Moore	**this study** – complete (M. J. Moore

		11905, FLAS)	11905, FLAS)
	Mollugo verticellata	**this study** – complete (T. Wieboldt 11,751, VPI)	**this study** – complete (T. Wieboldt 11,751, VPI)
Nepenthaceae	*Nepenthes alata*	**AF315891** – complete (Meimberg et al. 2001)	**AF315891** – complete (Meimberg et al. 2001)
	Nepenthes burkei	**DQ840247** – complete (Meimberg and Heubl 2006)	**DQ840247** – complete (Meimberg and Heubl 2006)
Nyctaginaceae	*Bouganvillea glabra*	**this study** – complete (S. Crawley 1, VPI)	**this study** – complete (S. Crawley 1, VPI)
	Mirabilis jalapa	**this study** – partial (S. Crawley 2, VPI) **AY042614** – partial (Cuénoud et al. 2002)	**this study** – complete (S. Crawley 2, VPI)
Phytolaccaceae	*Phytolacca americana*	**this study** – complete (A. Hinkle 328, VPI)	**this study** – complete (A. Hinkle 328, VPI)
	Rivina humilis	**AY514850** – complete (Müller and Borsch	**AY514850** – complete (Müller and Borsch

		2005)	2005)
Plumbaginaceae	*Limonium latifolium*	**AY514861** – complete (Müller and Borsch 2005)	**AY514861** – complete (Müller and Borsch 2005)
	Plumbago auriculata	**this study** – complete (M. J. Moore 306, FLAS)	**this study** – complete (M. J. Moore 306, FLAS)
Polygonaceae	*Polygonum cespitosum*	**this study** – complete (A. Hinkle 376, VPI)	**this study** – complete (A. Hinkle 376, VPI)
Portulacaceae	*Anacampseros vulcanensis*	**AY514851** – complete (Müller and Borsch 2005)	**AY514851** – complete (Müller and Borsch 2005)
	Claytonia megarhiza	**this study** – partial (M. W. Chase 10985, K)	**this study** – partial (M. W. Chase 10985, K)
		AY042569 – partial (Cuénoud et al. 2002)	**AY764103** – partial (O'Quinn and Hufford 2005)
		AY764103 – partial (O'Quinn and Hufford 2005)	
	Portulaca oleracea	**AY875349** – complete (Edwards et al.	**AY875349** – complete (Edwards et al.

		2005)	2005)
	Talinella sp__AC45_1	**AY514859** – complete (Müller and Borsch 2005)	**AY514859** – complete (Müller and Borsch 2005)
	Talinum paniculatum	**AY015274** – complete (Nyffeler 2002)	**AY015274** – complete (Nyffeler 2002)
Rhabdodendraceae	*Rhabdodendron amizonicum*	**this study** – complete (Ribiero 1187, K)	**this study** – complete (Ribiero 1187, K)
Sarcobataceae	*Sarcobatus vermiculatus*	**this study** – partial (2994987 M. B. G.) **AY042652** – partial (Cuénoud et al. 2002)	**this study** – complete (2994987 M. B. G.)
Simmondsiaceae	*Simmondsia chinensis*	**AY514854** – complete (Müller and Borsch 2005)	**AY514854** – complete (Müller and Borsch 2005)
Stegnospermataceae	*Stegnosperma halmifolium*	**HQ878442**– complete (Soltis et al. Accepted)	**HQ878442**– complete (Soltis et al. Accepted)
Tamaricaceae	*Tamarix pentandra*	**this study** – partial (K. Hilu 501, VPI) **AY042663** – partial (Cuénoud et al. 2002)	**this study** – complete (K. Hilu 501, VPI)

References:

Cameron, K.M., Wurdack, K.J., Jobson, R.W., 2002. Molecular Evidence for the Common Origin of Snap-Traps among Carnivorous Plants. American Journal of Botany 89, 1503 - 1509.

Cuénoud, P., Savolainen, V., Chatrou, L.W., Powell, M.P., Grayer, R.J., Chase, M.W., 2002. Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL, atpB,* and *matK* DNA sequences. American Journal of Botany 89, 132 - 144.

Edwards, E.J., Nyffeler, R., Donoghue, M.J., 2005. Basal Cactus Phylogeny: Implications of *Pereskia* (Cactaceae) Paraphyly for the Transition to the Cactus Life Form. American Journal of Botany 92, 1177 - 1188.

Fior, S., Karis, P.O., Casazza, G., Minuto, L., Sala, F., 2006. Molecular phylogeny of the Caryophyllaceae (Caryophyllales) inferred from chloroplast *matK* and nuclear rDNA ITS sequences. American Journal of Botany 93, 399 - 411.

Fishbein, M., Hibsch-Jetter, C., Soltis, D.E., Hufford, L., 2001. Phylogeny of Saxifragales (angiosperms, eudicots): analysis of a rapid, ancient radiation. Systematic Biology 50, 817 - 847.

Meimberg, H., Heubl, G., 2006. Introduction of a Nuclear Marker for Phylogenetic Analysis of Nepenthaceae. Plant Biology 8, 831 - 840.

Meimberg, H., Wistuba, A., Dittrich, P., Heubl, G., 2001. Molecular Phylogeny of Nepenthaceae Based on Cladistic Analysis of Plastid trnK Intron Sequence Data. Plant Biology 3, 164 - 175.

Müller, K.F., Borsch, T., 2005. Phylogenetics of Amaranthaceae Based on *matK/trnK* Sequence Data - Evidence from Parsimony, Likelihood, and Bayesian Analyses. Annals of the Missouri Botanical Gardens 92, 66 - 102.

Nyffeler, R., 2002. Phylogenetic relationships in the cactus family (Cactaceae) based on evidence from *trnK/matK* and *trnL-trnF* sequences. American Journal of Botany 89, 312 - 326.

Schmitz-Linneweber, C., Maier, R.M., Alcaraz, J.-P., Cottet, A., Herrmann, R.G., Regis, M., 2001. The Plastid chromosome of spinach (Spinacia oleracea): complete

nucleotide sequence and gene organization. Plant Molecular Biology 45, 307 - 315.

Soltis, D., Smith, S., Cellinese, N., Wurdack, K., Tank, D., Brockington, S., Refulio-Rodriguez, N., Walker, J., Moore, M., Carlsward, B., Bell, C., Latvis, M., Crawley, S., Black, C., Diouf, D., Xi, Z., Gitzendanner, M., Sytsma, K., Qiu, Y.-L., Hilu, K., Davis, C., Sanderson, M., Olmstead, R., Judd, W., Donoghue, M., Soltis, P., Accepted. Angiosperm phylogeny: 17 genes, 640 taxa. American Journal of Botany.

Appendix C

Supplemental Table 2 for Chapter 2

Table C.1 Species used, their taxonomic affiliation, GenBank numbers, and reference information for the 5 gene, 136 taxon data matrix (5GR-136). "Complete" and "partial" indicates whether the sequences obtained from GenBank were for the whole genomic region or for portions of it. In some instances sequence information for the same species was not available, in those cases, a different species from the same genus was used as a "placeholder". Please note that * indicates that information for that entry can be found in Table B.1.

Family	GenBank No./Source				
	matK	*trnK* introns	*rbcL*	*atpB*	*ndhF*
Outgroup Taxa					
Berberidopsidaceae	*Berberidopsis corallina**	*B. corallina**	*B. corallina* **EU002274** – complete (Wang et al., 2009)	*B. corallina* **EU002158** – complete (Wang et al., 2009)	*B. corallina* **EU002201** – complete (Wang et al., 2009)
Dilleniaceae	*Dillenia indica**	*D. indica**	*D. indica* **L01903** – complete (Albert et al.)	*D. retusa* **AF095732** – complete (Savolainen et al., 2000a)	*D. philippinensis* **AY425045** – partial (Davis and Chase, 2004)
	*Tetracera asiatica**	*T. asiatica**	*T. asiatica* **AJ235796** – complete (Savolainen et al., 2000a)	*T. asiatica* **AJ235622** – complete (Savolainen et al., 2000a)	*T. asiatica* **AJ236277** – complete (Albach et al., 2001)
Santalaceae	*Santalum album**	*S. album**	*S. album* **L26077** – complete (Nickrent and Soltis)	*S. album* **AJ235592** – complete (Savolainen et al., 2000a)	No Sequence
Vitaceae	*Parthenocissus*	*P. quinquefolia**	*P. quinquefolia*	No Sequence	No Sequence

	*quinquefolia**		**AJ402985** – partial (Savolainen et al.)		
	*Vitis aestivalis**	*V. aestivalis**	*V. aestivalis* **L01960** – complete (Albert et al., 1992)	*V. aestivalis* **AJ235643** – complete (Savolainen et al.)	*V. vinifera* **NC_007957** – complete (Jansen et al., 2006)
Ingroup Taxa					
Achatocarpaceae	*Achatocarpus praecox**	*A. praecox**	*A. praecox* **AY270142** – complete (Kadereit et al., 2003)	No Sequence	*A. praecox* **AY858609** – complete (Hohmann et al., 2006)
	*Phaulothamnus spinescens**	*P. spinescens**	*P. spinescens* **M97887** – complete (Manhart and Rettig, unpublished)	No Sequence	*P. spinescens* **AY858610** – complete (Hohmann et al., 2006)
Aizoaceae	*Aridaria noctiflora* **AY042619** – partial (Cuénoud et al.)	No Sequence	No Sequence	No Sequence	No Sequence
	Corbichonia decumbens **AY042572** – partial (Cuénoud et al., 2002)	No Sequence	*C. decumbens* **AF132096** – complete (Clement and Mabry, unpublished)	No Sequence	No Sequence
	*Delosperma napiforme**	*D. napiforme**	*D. echinatum* **AJ235778** – complete (Savolainen et al., 2000a)	*D. echinatum* **AJ235452** – complete (Savolainen et al., 2000a)	*D. cooperi* **DQ855864** – complete (Nyffeler, 2007)
	Galenia pubescens **AY042589** –partial	No Sequence	*G. pubescens* **AF132099** –	No Sequence	No Sequence

	(Cuénoud et al., 2002)		complete (Clement and Mabry, unpublished)		
	Plinthus cryptocarpus **AY042633** – partial (Cuénoud et al., 2002)	No Sequence	No Sequence	No Sequence	No Sequence
	Ruschia schollii **AY042649** – partial (Cuénoud et al., 2002)	No Sequence	*R. brakdamensis* **AM234795** – complete (Forest et al., 2007)	No Sequence	No Sequence
Amaranthaceae	*Amaranthus greggii* **AY514808** – complete (Müller and Borsch, 2005b)	*A. greggii* **AY514808** – complete (Müller and Borsch, 2005b)	*A. hypochondriacus* **X51964** – complete (Michalowski et al., 1990)	*A. hypochondriacus* **AJ235388** – complete (Savolainen et al., 2000a)	*A. quitensis* **AF194822** – complete (Applequist and Wallace, 2001)
	Atriplex truncata **AY514830** – complete (Müller and Borsch, 2005b)	*A. truncata* **AY514830** – complete (Müller and Borsch, 2005b)	*A. spongiosa* **AY270060** – complete (Kadereit et al., 2003)	No Sequence	*A. spongiosa* **AY858615** – complete (Hohmann et al., 2006)
	*Celosia trigyna**	*C. trigyna**	*C. argentea* **AY270072** – complete (Kadereit et al., 2003)	*C. argentea* **AF209559** – complete (Soltis et al., 1999)	*C. argentea* **AY959890** – complete (Applequist and Pratt, 2005)
	Gomphrena fuscipellita **AM887525** – complete (Ortuno et al., unpublished)	*G. fuscipellita* **AM887525** – complete (Ortuno et al., unpublished)	*G. elegans* **AY270088** – complete (Kadereit et al., 2003)	No Sequence	No Sequence
	*Guilleminea densa**	*G. densa**	*G. densa* **AY270091** – complete	No Sequence	No Sequence

		(Kadereit et al., 2003)		
Hablitzia tamnoides **AY514825** – complete (Müller and Borsch, 2005b)	*H. tamnoides* **AY514825** – complete (Müller and Borsch, 2005b)	*H. tamnoides* **AY270092** – complete (Kadereit et al., 2003)	No Sequence	*H. tamnoides* **AY858629** – complete (Hohmann et al., 2006)
Iresine cassiniiformis **AM887489** – complete (Borsch et al., unpublished)	*I. cassiniiformis* **AM887489** – complete (Borsch et al., unpublished)	*I. palmeri* **AY270101** – complete (Kadereit et al., 2003)	No Sequence	No Sequence
*Nototrichium sandwicense**	*N. sandwicense**	*N. humile* **AY270111** – complete (Kadereit et al., 2003)	No Sequence	No Sequence
Psilotrichum ferrugineum **AY998108** – complete (Müller and Borsch, 2005a)	*P. ferrugineum* **AY998108** – complete (Müller and Borsch, 2005a)	No Sequence	No Sequence	No Sequence
Salicornia sp. Akhani s.n. **DQ499403** – partial (Kapralov et al., 2006)	*S. sp. Akhani s.n.* **DQ499403** – partial (Kapralov et al., 2006)	*S. meyeriana* **AM234802** – complete (Forest et al., 2007)	No Sequence	*S. sp. Freitag 13/2001* **AY858620** – partial (Hohmann et al., 2006)
Salsola kali **AY514843** – complete (Müller and Borsch, 2005b)	*S. kali* **AY514843** – complete (Müller and Borsch, 2005b)	*S. kali* **AY270129** – complete (Kadereit et al., 2003)	No Sequence	*S. kali* **DQ097401** – complete (Hohmann et al., 2006)
Sarcocornia fruticosa **DQ468645** – partial (Pagliano et al., unpublished)	No Sequence	*S. utahensis* **AY270126** – complete (Kadereit et al., 2003)	No Sequence	No Sequence

	*Spinacia oleracea**	*S. oleracea**	*S. oleracea* **AJ400848** – complete (Zurawski et al., 1982)	*S. oleracea* **AJ400848** – complete (Zurawski et al., 1982)	*S. oleracea* **AJ400848** – complete (Zurawski et al., 1982)
	Suaeda sp. AC66 **AY514841** – complete (Müller and Borsch)	*S. sp. AC66* **AY514841** – complete (Müller and Borsch)	*S. maritima* **AY270137** – complete (Kadereit et al., 2003)	No Sequence	*S. maritima* **DQ097400** – partial (Hohmann et al., 2006)
Ancistrocladaceae	*Ancistrocladus abbreviatus**	*A. abbreviatus**	No Sequence	No Sequence	No Sequence
	Ancistrocladus heyneanus **AF204841** – partial (Meimberg et al., 2000)	No Sequence	No Sequence	No Sequence	No Sequence
	Ancistrocladus korupensis **AF204839** – partial (Meimberg et al., 2000)	No Sequence	*A. korupensis* **Z97636** – complete (Fay et al., 1997)	*A. korupensis* **AF209526** – complete (Soltis et al., 1999)	No Sequence
Asteropeiaceae	*Asteropeia micraster**	*A. micraster**	*A. micraster* **AF206737** – complete (Soltis et al., 1999)	*A. micraster* **AF209533** – complete (Soltis et al., 1999)	No Sequence
Barbeuiaceae	*Barbeuia madagascariensis* **AY042552** – partial (Cuénoud et al., 2002)	No Sequence	No Sequence	No Sequence	No Sequence
Basellaceae	*Anredera cordifolia* **AY042547** – partial (Cuénoud et al., 2002)	No Sequence	*A. cordifolia* **AY270147** – complete (Kadereit et al., 2003)	No Sequence	No Sequence
	*Basella alba**	*B. alba**	*B. alba* **M62564**	No Sequence	*B. alba*

			– complete (Rettig et al., 1992)		**AF194834** – complete (Applequist and Wallace, 2001)
Cactaceae	*Acanthocalycium glaucum* **AY015325** – complete (Nyffeler, 2002)	*A. glaucum* **AY015325** – complete (Nyffeler, 2002)	No Sequence	No Sequence	No Sequence
	*Austrocylindropuntia subulata**	*A. subulata**	*A. subulata* **AY875235** – complete (Edwards et al., 2005)	No Sequence	*A. vestita* **DQ855878** – complete (Nyffeler, 2007)
	*Blossfeldia liliputana**	*B. liliputana**	*B. liliputana* **AY875232** – complete (Edwards et al., 2005)	No Sequence	No Sequence
	Calymmanthium substerile **AY015291** – complete (Nyffeler, 2002)	*C. substerile* **AY015291** – complete (Nyffeler, 2002)	*C. substerile* **AY875230** – complete (Edwards et al., 2005)	No Sequence	No Sequence
	Ceraria fruticulosa **AY875371** – complete (Edwards et al., 2005)	*C. fruticulosa* **AY875371** – partial (Edwards et al., 2005)	*C. fruticulosa* **AY875218** – complete (Edwards et al., 2005)	No Sequence	*C. fruticulosa* **AF194841** – complete (Applequist and Wallace, 2001)
	Cereus alacriportanus **AY015313** – complete (Nyffeler, 2002)	*C. alacriportanus* **AY015313** – complete (Nyffeler, 2002)	*C. fernambucensis* **AY875240** – complete (Edwards et al., 2005)	No Sequence	No Sequence
	Echinocactus platyacanthus **AY015287** –	*E. platyacanthus* **AY015287** – complete	*E. platyacanthus* **AY875215** – complete	No Sequence	No Sequence

complete (Nyffeler, 2002)	(Nyffeler, 2002)	(Edwards et al., 2005)		
Eriosyce napina **AY015339** – complete (Nyffeler, 2002)	*E. napina* **AY015339** – complete (Nyffeler, 2002)	No Sequence	No Sequence	No Sequence
Grahamia coahuilensis **AY875374** – complete (Edwards et al., 2005)	No Sequence	*G. coahuilensis* **AY875246** – complete (Edwards et al., 2005)	No Sequence	*G. frutescens* **DQ855871** – complete (Nyffeler, 2007)
Haageocereus pseudomelanostele **AY015329** – complete (Nyffeler, 2002)	*H. pseudomelanostele* **AY015329** – complete (Nyffeler, 2002)	No Sequence	No Sequence	No Sequence
Harrisia pomanensis **AY015324** – complete (Nyffeler, 2002)	*H. pomanensis* **AY015324** – complete (Nyffeler, 2002)	No Sequence	No Sequence	No Sequence
Hatiora salicornioides **AY015341** – complete (Nyffeler, 2002)	*H. salicornioides* **AY015341** – complete (Nyffeler, 2002)	No Sequence	No Sequence	No Sequence
Lepismium cruciforme **AY015344** – complete (Nyffeler, 2002)	*L. cruciforme* **AY015344** – complete (Nyffeler, 2002)	No Sequence	No Sequence	No Sequence
Lobivia pentlandii **AY015323** – complete (Nyffeler, 2002)	*L. pentlandii* **AY015323** – complete (Nyffeler, 2002)	No Sequence	No Sequence	No Sequence
Maihuenia patagonica **AY015281** –	*M. patagonica* **AY015281** – complete	*M. patagonica* **AY875245** – complete	No Sequence	*M. patagonica* **DQ855877** – complete

complete (Nyffeler, 2002)	(Nyffeler, 2002)	(Edwards et al., 2005)		(Nyffeler, 2007)
Maihuenia poeppigii **AY015282** – complete (Nyffeler, 2002)	*M. poeppigii* **AY015282** – complete (Nyffeler, 2002)	*M. poeppigii* **AY875216** – complete (Edwards et al., 2005)	No Sequence	*M. poeppigii* **AF206714** – partial (Applequist and Wallace, 2001)
Matucana intertexta **AY015327** – complete (Nyffeler, 2002)	*M. intertexta* **AY015327** – complete (Nyffeler, 2002)	No Sequence	No Sequence	No Sequence
Neowerdermannia vorwerkii **AY015340** – complete (Nyffeler, 2002)	*N. vorwerkii* **AY015340** – complete (Nyffeler, 2002)	No Sequence	No Sequence	No Sequence
*Opuntia quimilo**	*O. quimilo**	*O. dillenii* **AY875233** – complete (Edwards et al., 2005)	No Sequence	No Sequence
Oreocereus celsianus **AY015328** – complete (Nyffeler, 2002)	*O. celsianus* **AY015328** – complete (Nyffeler, 2002)	No Sequence	No Sequence	No Sequence
Parodia ottonis **AY015335** – complete (Nyffeler, 2002)	*P. ottonis* **AY015335** – complete (Nyffeler, 2002)	No Sequence	No Sequence	No Sequence
*Pereskia aculeata**	*P. aculeata**	*P. aculeata* **AY875229** – complete (Edwards et al., 2005)	*P. aculeata* **AF209648** – complete (Soltis et al., 1999)	*P. aculeata* **AF194852** – complete (Applequist and Wallace, 2001)
Pereskiopsis deguetii **AY015280** – complete (Nyffeler, 2002)	*P. deguetii* **AY015280** – complete (Nyffeler, 2002)	*P. porteri* **AY875243** – complete (Edwards et al.,	No Sequence	No Sequence

			2005)		
	Quiabentia zehntneri **AY875372** – complete (Edwards et al., 2005)	*Q. zehntneri* **AY875372** – partial (Edwards et al., 2005)	*Q. verticillata* **AY875239** – complete (Edwards et al., 2005)	No Sequence	*Q. verticillata* **AF194858** – complete (Applequist and Wallace, 2001)
	Rauhocereus riosaniensis **AY015326** – complete (Nyffeler, 2002)	*R. riosaniensis* **AY015326** – complete (Nyffeler, 2002)	No Sequence	No Sequence	No Sequence
	Rhipsalis floccose **AY015342** – complete (Nyffeler, 2002)	*R. floccose* **AY015342** – complete (Nyffeler, 2002)	No Sequence	No Sequence	No Sequence
	Schlumbergera truncata **AY015343** – complete (Nyffeler, 2002)	*S. truncata* **AY015343** – complete (Nyffeler, 2002)	*S. truncata* **M83543** – complete (Manhart et al., unpublished)	No Sequence	No Sequence
	Tephrocactus articulatus **AY875367** – complete (Edwards et al., 2005)	*T. articulatus* **AY875367** – partial (Edwards et al., 2005)	*T. articulatus* **AY875248** – complete (Edwards et al., 2005)	No Sequence	No Sequence
Caryophyllaceae	*Atocion rupestris* **EF547242** – complete (Mower et al., 2007)	*A. rupestris* **EF547242** – complete (Mower et al., 2007)	No Sequence	No Sequence	No Sequence
	*Illecebrum verticillatum**	*I. verticillatum**	*I. verticillatum* **AY270143** – complete (Kadereit et al., 2003)	No Sequence	No Sequence
	*Scleranthus perennis**	*S. perennis**	*S. annuus* **AY270145** – complete	No Sequence	*S. biflorus* **AY090633** – partial (Smissen

			(Kadereit et al., 2003)		et al., 2002)
	Silene latifolia **EF547239** – complete (Mower et al., 2007)	*S. latifolia* **EF547239** – complete (Mower et al., 2007)	*S. latifolia* **EF418555** – complete (Kapralov and Filatov, 2007)	*S. nutans* **AJ235601** – complete (Savolainen et al., 2000a)	*S. latifolia* **DQ841751** – partial (Houliston and Olson, 2006)
	*Stellaria media**	*S. media**	*S. media* **M62570** – complete (Rettig et al., unpublished)	*S. media* **AF209680** – partial (Soltis et al., 1999)	*S. media* **AY090630** – partial (Smissen et al., 2002)
Didiereaceae	*Alluaudia ascendens**	*A. ascendens**	*A. procera* **M62563** – complete (Rettig et al., 1992)	No Sequence	*A. humbertii* **AF194832** – complete (Applequist and Wallace, 2001)
	Alluaudiopsis fiherenensis **AY042542** – partial (Cuénoud et al., 2002	No Sequence	No Sequence	No Sequence	No Sequence
	Decarya madagascariensis **AY042574** – partial (Cuénoud et al., 2002)	No Sequence	No Sequence	No Sequence	*D. madagascariensis* **AF194844** – complete (Applequist and Wallace, 2001)
	Didierea trollii **AY042576** – partial (Cuénoud et al., 2002)	No Sequence	No Sequence	No Sequence	*D. trollii* **AF194845** – complete (Applequist and Wallace, 2001)
Dioncophyllaceae	*Dioncophyllum tholloni* **AF204844** – partial (Meimberg et al., 2000)	No Sequence	No Sequence	No Sequence	No Sequence

	Habropetalum dawei **AF204845** – partial (Meimberg et al., 2000)	No Sequence	No Sequence	No Sequence	No Sequence
	*Triphyophyllum peltatum**	*T. peltatum**	*T. peltatum* **Z97637** – complete (Fay et al., 1997)	No Sequence	No Sequence
Droseraceae	*Aldrovanda vesiculosa* **AY096120** – partial (Cameron et al., 2002)	No Sequence	*A. vesiculosa* **AY096106** – complete (Cameron et al., 2002)	*A. vesiculosa* **AY096108** – complete (Cameron et al., 2002)	No Sequence
	Dionaea muscipula **AF204847** – partial (Meimberg et al., 2000)	No Sequence	*D. muscipula* **L01904** – complete (Albert et al.)	*D. muscipula* **AY096112** – complete (Cameron et al., 2002)	No Sequence
	Drosera aliciae **AF204849** – partial (Meimberg et al., 2000)	No Sequence	*D. aliciae* **AB072516** – partial (Rivadavia et al., 2003)	No Sequence	No Sequence
	*Drosera capensis**	*D. capensis**	*D. capensis* **L01909** – complete (Albert et al., 1992)	*D. capensis* **AY096110** – complete (Cameron et al., 2002)	No Sequence
	Drosera capillaris **AF204850** – partial (Meimberg et al., 2000)	No Sequence	*D. capillaris* **AB355693** – partial (Hoshi et al., unpublished)	No Sequence	No Sequence
	Drosera communis **AY042579** – partial (Cuénoud et al., 2002)	No Sequence	No Sequence	No Sequence	No Sequence
	Drosera regia	No Sequence	*D. regia* **L01914**	*D. regia*	No Sequence

	AF204848 – partial (Meimberg et al., 2000)		– complete (Albert et al., 1992)	**AY096111** – complete (Cameron et al., 2002)	
Drosophyllaceae	*Drosophyllum lusitanicum**	*D. lusitanicum**	*D. lusitanicum* **L01907** – complete (Albert et al., 1992)	*D. lusitanicum* **AY096113** – complete (Cameron et al., 2002)	No Sequence
Frankeniaceae	*Frankenia corymbosa* **AY042587** – partial (Cuénoud et al., 2002)	No Sequence	No Sequence	No Sequence	No Sequence
	*Frankenia laevis**	*F. laevis**	*F. pulverulenta* **Z97638** – Complete (Fay et al., 1997)	*F. pulverulenta* **AJ235476** – complete (Savolainen et al., 2000a)	No Sequence
Gisekiaceae	*Gisekia africana**	*G. africana**	*G. pharnacioides* **M97890** – complete (Manhart and Rettig, unpublished)	No Sequence	No Sequence
Halophytaceae	*Halophytum ameghinoi**	*H. ameghinoi**	*H. ameghinoi* **AJ403024** – partial (Savolainen et al., 2000b)	No Sequence	No Sequence
Molluginaceae	*Limeum africanum**	*L. africanum**	*L. sp. Hoot 983* **AF093727** – complete (Hoot et al., 1999)	*L. sp. Hoot 983* **AF093385** – complete (Hoot et al., 1999)	No Sequence
	*Mollugo verticellata**	*M. verticellata**	*M. verticellata* **M62566** –	*M. verticellata* **AF209631** –	*M. verticellata* **AF194827** –

			complete (Rettig et al., 1992)	complete (Soltis et al., 1999)	complete (Applequist and Wallace, 2001)
Nepenthaceae	*Nepenthes alata**	*N. alata**	*N. alata* **L01936** – complete (Albert et al., 1992)	*N. alata* **AJ235542** – complete (Savolainen et al., 2000a)	No Sequence
	*Nepenthes burkei**	*N. burkei**	No Sequence	No Sequence	No Sequence
	Nepenthes gracillima **DQ007086** – complete (Meimberg et al.)	*N. gracillima* **DQ007086** – complete (Meimberg et al.)	No Sequence	No Sequence	No Sequence
	Nepenthes xiphioides **DQ007080** – complete (Meimberg et al.)	*N. xiphioides* **DQ007080** – complete (Meimberg et al.)	No Sequence	No Sequence	No Sequence
Nyctaginaceae	*Acleisanthes somalensis* **AY042655** – partial (Cuénoud et al., 2002)	No Sequence	No Sequence	No Sequence	*A. lanceolata* **EF079509** – complete (Douglas and Manos, 2007)
	Allionia incarnata **AY042540** – partial (Cuénoud et al., 2002)	No Sequence	No Sequence	No Sequence	*A. violacea* **AF194823** – complete (Applequist and Wallace)
	Boerhavia coccinea **AY042558** – partial (Cuénoud et al., 2002)	No Sequence	No Sequence	No Sequence	*B. coccinea* **EF079525** – complete (Douglas and Manos, 2007)
	*Bougainvillea glabra**	*B. glabra**	*B. glabra* **M88340** – complete (Manhart	*B. glabra* **AJ235415** – complete (Savolainen et	*B. alba* **AF194825** – complete (Applequist and

			unpublished)	al., 2000a)	Wallace, 2001)
	Commicarpus raynalii **AY042571** – partial (Cuénoud et al., 2002)	No Sequence	No Sequence	No Sequence	*C. coctoris* **EF079535** – complete (Douglas and Manos, 2007)
	*Mirabilis jalapa**	*M. jalapa**	*M. jalapa* **M62565** – complete (Rettig et al.)	*M. jalapa* **AF209629** – partial (Soltis et al., 1999)	*M. jalapa* **AF194826** – complete (Applequist and Wallace, 2001)
	Mirabilis nyctaginea **AY042624** – partial (Cuénoud et al., 2002)	No Sequence	No Sequence	No Sequence	No Sequence
	Pisonia umbellifera **AY042632** – partial (Cuénoud et al., 2002)	No Sequence	No Sequence	No Sequence	*P. capitata* **EF079537** – complete (Douglas and Manos, 2007)
Phytolaccaceae	*Ercilla volubilis* **AY042583** – partial (Cuénoud et al., 2002)	No Sequence	*E. volubilis* **AJ235800** – complete (Savolainen et al., 2000a)	*E. volubilis* **AJ235464** – complete (Savolainen et al., 2000a)	No Sequence
	Gallesia integrifolia **AY042590** – partial (Cuénoud et al., 2002)	No Sequence	No Sequence	No Sequence	No Sequence
	Hilleria latifolia **AY042601** – partial (Cuénoud et al., 2002)	No Sequence	No Sequence	No Sequence	No Sequence
	*Phytolacca americana**	*P. americana**	*P. americana* **M62567** – complete (Rettig et al., 1992)	*P. americana* **AF093391** – complete (Hoot et al., 1999)	*P. americana* **AF130229** – complete (Olmstead et al.,

					2000)
	Phytolacca dioica **AY042631** – partial (Cuénoud et al., 2002)	No Sequence	No Sequence	*P. dioica* **AJ235558** – complete (Savolainen et al., 2000a)	*P. acinosa* **AF194828** – complete (Applequist and Wallace, 2001)
	*Rivina humilis**	*R. humilis**	*R. humilis* **M62569** – complete (Rettig et al., 1992)	No Sequence	*R. humilis* **EF079516** – complete (Douglas and Manos, 2007)
	Seguieria aculeata **AY042654** – partial (Cuénoud et al., 2002)	No Sequence	No Sequence	No Sequence	No Sequence
Plumbaginaceae	*Afrolimon purpuratum* **AY042537** – partial (Cuénoud et al., 2002)	No Sequence	No Sequence	No Sequence	No Sequence
	Armeria gaditana **AF204856** – partial (Meimberg et al., 2000)	No Sequence	*A. splendens* **Y16908** – complete (Lledo et al., 1998)	No Sequence	No Sequence
	Armeria maritima **AY042548** – partial (Cuénoud et al., 2002)	No Sequence	*A. bottendorfensis* **Z97640** – complete (Fay et al., 1997)	No Sequence	No Sequence
	Ceratostigma minus **AY042566** – partial (Cuénoud et al., 2002)	No Sequence	*C. minus* **Z97641** – complete (Fay et al., 1997)	No Sequence	No Sequence
	Dyerophytum africanum **AY042581** – partial (Cuénoud et al.,	No Sequence	*D. africanum* **AJ312253** – complete (Lledo et al., 2001)	No Sequence	No Sequence

	2002)				
	Goniolimon tataricum **AF204855** – partial (Meimberg et al., 2000)	No Sequence	*G. speciosum* **AJ312254** – complete (Lledo et al., 2001)	No Sequence	No Sequence
	Limoniastrum monopetalum **AY042609** – partial (Cuénoud et al., 2002)	No Sequence	*L. monopetalum* **Z97642** – complete (Fay et al., 1997)	No Sequence	No Sequence
	*Limonium latifolium**	*L. latifolium**	*L. rigualii* **Z97645** – complete (Fay et al., 1997)	*L. arborescens* **AF209620** – complete (Soltis et al., 1999)	No Sequence
	*Plumbago auriculata**	*P. auriculata**	*P. auriculata* **M77701** – complete (Giannasi et al., 1992)	*P. auriculata* **EU002166** – complete (Wang et al., 2009)	*P. auriculata* **EU002252** – complete (Wang et al., 2009)
	Psylliostachys suworowii **AY042639** – partial (Cuénoud et al., 2002)	No Sequence	*P. suworowii* **Y16907** – complete (Lledo et al., 1998)	No Sequence	No Sequence
Polygonaceae	*Fagopyrum esculentum* **AB093087** – complete (Yamane et al., 2003)	*F. esculentum* **NC_010776** – complete (Logacheva et al., 2008)	*F. esculentum* **NC_010776** – complete (Logacheva et al., 2008)	*F. esculentum* **NC_010776** – complete (Logacheva et al., 2008)	*F. esculentum* **NC_010776** – complete (Logacheva et al., 2008)
	Fallopia multiflora var. ciliinervis **EU024768** – complete (Yu and Li, unpublished)	No Sequence	*F. multiflora* **FM883616** – complete (Galasso, unpublished)	No Sequence	*F. japonica* **EF438048** – partial (Sanchez and Kron)
	*Polygonum cespitosum**	*P. cespitosum**	*P. weyrichii* **AF297145** – complete (Frye	*P. sachalinense* **AJ235569** – complete	*P. forrestii* **EF438051** – partial (Sanchez

			and Kron, 2003)	(Savolainen et al., 2000a)	and Kron)
Portulacaceae	*Anacampseros vulcanensis**	*A. vulcanensis**	*A. papyracea* **AM235079** – complete (Forest et al., 2007)	No Sequence	*A. retusa* **AF194833** – complete (Applequist and Wallace, 2001)
	*Claytonia megarhiza**	*C. megarhiza**	*C. perfoliata* **AF132093** – complete (Clement and Mabry, unpublished)	No Sequence	*C. virginica* **AF194856** – complete (Applequist and Wallace, 2001)
	Hectorella caespitosa **EF551350** – partial (Wagstaff and Hennion, 2007)	*H. caespitosa* **EF551350** – partial (Wagstaff and Hennion, 2007)	*H. caespitosa* **DQ267193** – complete (Applequist et al., 2006)	No Sequence	*H. caespitosa* **DQ093963** – complete (Applequist et al., 2006)
	Lyallia kerguelensis **EF551349** – partial (Wagstaff and Hennion, 2007)	*L. kerguelensis* **EF551349** – partial (Wagstaff and Hennion, 2007)	*L. kerguelensis* **EF551348** – complete (Wagstaff and Hennion, 2007)	No Sequence	No Sequence
	*Portulaca oleracea**	*P. oleracea**	*P. grandiflora* **M62568** – complete (Rettig et al., 1992)	*P. grandiflora* **AF209659** – partial (Soltis et al., 1999)	*P. oleracea* **AF194867** – complete (Applequist and Wallace, 2001)
	Portulacaria afra **AY875368** – complete (Edwards et al., 2005)	*P. afra* **AY875368** – partial (Edwards et al., 2005)	*P. afra* **AM235080** – complete (Forest et al., 2007)	No Sequence	*P. afra* **AF194857** – complete (Applequist and Wallace, 2001)
	*Talinella sp__AC45_1**	*T. sp__AC45_1**	No Sequence	No Sequence	*T. pachypoda* **DQ855868** – complete (Nyffeler, 2007)

	*Talinum paniculatum**	*T. paniculatum**	*T. paniculatum* **AY875214** – complete (Edwards et al., 2005)	No Sequence	*T. paniculatum* **DQ855866** – complete (Nyffeler, 2007)
Rhabdodendraceae	*Rhabdodendron amizonicum**	*R. amizonicum**	*R. amizonicum* **Z97649** – complete (Fay et al., 1997)	*R. amizonicum* **AJ235578** – complete (Savolainen et al., 2000a)	No Sequence
	Rhabdodendron macrophyllum **AY042642** – partial (Cuénoud et al., 2002)	No Sequence	No Sequence	No Sequence	No Sequence
Sarcobataceae	*Sarcobatus vermiculatus**	*S. vermiculatus**	*S. vermiculatus* **AF132088** – complete (Clement and Mabry, unpublished)	No Sequence	*S. vermiculatus* **EF079555** – partial (Douglas and Manos, 2007)
Simmondsiaceae	*Simmondsia chinensis**	*S. chinensis**	*S. chinensis* **AF093732** – complete (Hoot et al., 1999)	*S. chinensis* **AF093401** – complete (Hoot et al., 1999)	No Sequence
Stegnospermataceae	*Stegnosperma halmifolium**	*S. halmifolium**	*S. halmifolium* **M62571** – complete (Rettig et al., 1992)	No Sequence	*S. cubense* **EF079554** – complete (Douglas and Manos, 2007)
Tamaricaceae	*Tamarix gallica* **AF204861** – partial (Meimberg et al., 2000)	No Sequence	*T. parviflora* **AY099901** – complete (Gaskin et al., 2004)	No Sequence	No Sequence
	*Tamarix pentandra**	*T. pentandra**	*T. pentandra* **Z97650** –	*T. pentandra* **AF209684** –	No Sequence

	complete (Fay et al., 1997)	complete (Soltis et al., 1999)

References:

Albach, D.C., Soltis, P.S., Soltis, D.E., Olmstead, R.G., 2001. Phylogenetic Analysis of Asterids Based on Sequences of Four Genes. Annals of the Missouri Botanical Garden 88, 163-212.

Albert, V.A., Williams, S.E., Chase, M.W., 1992. Carnivorous Plants: Phylogeny and Structural Evolution. Science 257, 1491 - 1495.

Applequist, W.L., Pratt, D.B., 2005. The Malagasy endemic *Dendroportulaca* (Portulacaceae) is referable to *Deeringia* (Amaranthaceae): molecular and morphological evidence. Taxon 54, 681-687.

Applequist, W.L., Wagner, W.L., Zimmer, E.A., Nepokroeff, M., 2006. Molecular Evidence Resolving the Systematic Position of *Hectorella* (Portulacaceae). Systematic Botany 31, 310-319.

Applequist, W.L., Wallace, R.S., 2001. Phylogeny of the Portulacaceous Cohort Based on *ndhF* Sequence Data. Systematic Botany 26, 406-419.

Cameron, K.M., Wurdack, K.J., Jobson, R.W., 2002. Molecular Evidence for the Common Origin of Snap-Traps Among Carnivorous Plants. American Journal of Botany 89, 1503-1509.

Cuénoud, P., Savolainen, V., Chatrou, L.W., Powell, M.P., Grayer, R.J., Chase, M.W., 2002. Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL, atpB,* and *matK* DNA sequences. American Journal of Botany 89, 132 - 144.

Davis, C.C., Chase, M.W., 2004. Elatinaceae are sister to Malpighiaceae; Peridiscaceae belong to Saxifragales. American Journal of Botany 91, 262-273.

Douglas, N.A., Manos, P.S., 2007. Molecular phylogeny of Nyctaginaceae: taxonomy, biogeography, and characters associated with a radiation of xerophytic genera in North America. American Journal of Botany 94, 856-872.

Edwards, E.J., Nyffeler, R., Donoghue, M.J., 2005. Basal Cactus Phylogeny: Implications of *Pereskia* (Cactaceae) Paraphyly for the Transition to the Cactus Life Form. American Journal of Botany 92, 1177 - 1188.

Fay, M.F., Cameron, K.M., Prance, G., T., Lledo, M.D., Chase, M.W., 1997. Familial relationships of *Rhabdodendron* (*Rhabdodendraceae*): plastid *rbc*L sequences indicate a caryophyllid placement. Kew Bulletin 54, 923 - 932.

Forest, F., Grenyer, R., Rouget, M., Davies, J., Cowling, R.M., Faith, D.P., Balmford, A., Manning, J.C., Procheş, Ş., van der Bank, M., Reeves, G., Hedderson, T.A.J., Savolainen, V., 2007. Preserving the evolutionary potential of floras in biodiversity hotspots. Nature 445, 757-760.

Frye, A.S.L., Kron, K.A., 2003. rbcL Phylogeny of Character Evolution in Polygonaceae. Systematic Botany 28, 326-332.

Gaskin, J.F., Ghahremani-nejad, F., Zhang, D.-y., Londo, J.P., 2004. A Systematic Overview of Frankeniaceae and Tamaricaceae from Nuclear rDNA and Plastid Sequence Data. Annals of the Missouri Botanical Garden 91, 401-409.

Giannasi, D.E., Zurawski, G., Learn, G., Clegg, M.T., 1992. Evolutionary Relationships of the Caryophyllidae Based on Comparative rbcL Sequences. Systematic Botany 17, 1-15.

Hohmann, S., Kadereit, J.W., Kadereit, G., 2006. Understanding Mediterranean-Californian disjunctions: molecular evidence from Chenopodiaceae-Betoideae. Taxon 55, 67-78.

Hoot, S.B., Magallon, S., Crane, P.R., 1999. Phylogeny of Basal Eudicots Based on Three Molecular Data Sets: atpB, rbcL, and 18s Nuclear Ribosomal DNA Sequences. Annals of the Missouri Botanical Garden 86, 1-32.

Houliston, G.J., Olson, M.S., 2006. Nonneutral Evolution of Organelle Genes in *Silene vulgaris*. Genetics 174, 1983-1994.

Jansen, R.K., Kaittanis, C., Saski, C., Lee, S.-B., Tomkins, J., Alverson, A.J., Daniell, H., 2006. Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evolutionary Biology 6, 32.

Kadereit, G., Borsch, T., Weising, K., Freitag, H., 2003. Phylogeny of Amaranthaceae and Chenopodiaceae and the evolution of C_4 photosynthesis. International Journal of Plant Sciences 164, 959-986.

Kapralov, M.V., Akhani, H., Voznesenskaya, E.V., Edwards, G., Franceschi, V., Roalson, E.H., 2006. Phylogenetic Relationships in the Salicornioideae / Suaedoideae / Salsoloideae s.l. (Chenopodiaceae) Clade and a Clarification of the Phylogenetic Position of *Bienertia* and *Alexandra* Using Multiple DNA Sequence Datasets. Systematic Botany 31, 571-585.

Kapralov, M.V., Filatov, D.A., 2007. Widespread positive selection in the photosynthetic Rubisco enzyme. BMC Evolutionary Biology 7, 73.

Lledo, M.D., Crespo, M.B., Cameron, K.M., Fay, M.F., Chase, M.W., 1998. Systematics of Plumbaginaceae Based upon Cladistic Analysis of rbcL Sequence Data. Systematic Botany 23, 21-29.

Lledo, M.D., Karis, P.O., Crespo, M.B., Fay, M.F., Chase, M.W., 2001. Phylogenetic position and taxonomic status of the genus *Aegialitis* and subfamilies Staticoideae and Plumbaginoideae (Plumbaginaceae): evidence from plastid DNA sequences and morphology. Plant Systematics and Evolution 229, 107-124.

Logacheva, M.D., Samigullin, T.H., Dhingra, A., Penin, A.A., 2008. Comparative chloroplast genomics and phylogenetics of *Fagopyrum esculentum* ssp. *ancestrale* - A wild ancestor of cultivated buckwheat. BMC Plant Biology 8, 59.

Meimberg, H., Dittrich, P., Bringmann, G., Schlauer, J., Heubl, G., 2000. Molecular Phylogeny of Caryophyllidae s.l. Based on *MatK* Sequences with Special Emphasis on Carnivorous Taxa. Plant Biology 2, 218-228.

Meimberg, H., Thalhammer, S., Brachmann, A., Heubl, G., 2006. Comparative analysis of a translocated copy of the *trnK* intron in carnivorous family Nepenthaceae. Molecular Phylogenetics and Evolution 39, 478-490.

Michalowski, C.B., Bohnert, H.J., Klessig, D.F., Berry, J.O., 1990. Nucleotide sequence of rbcL from *Amaranthus hypochondriacus* chloroplasts. Nucleic Acids Research 18, 2187.

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

Müller, K., Borsch, T., 2005a. Multiple origins of a unique pollen feature: stellate pore ornamentation in Amaranthaceae. Grana 44, 266-281.

Müller, K.F., Borsch, T., 2005b. Phylogenetics of Amaranthaceae Based on *matK*/*trnK* Sequence Data - Evidence from Parsimony, Likelihood, and Bayesian Analyses. Annals of the Missouri Botanical Gardens 92, 66 - 102.

Nickrent, D.L., Soltis, D.E., 1995. A Comparison of Angiosperm Phylogenies from Nuclear 18S rDNA and rbcL Sequences. Annals of the Missouri Botanical Garden 82, 208-234.

Nyffeler, R., 2002. Phylogenetic relationships in the cactus family (Cactaceae) based on evidence from *trnK*/*matK* and *trnL-trnF* sequences. American Journal of Botany 89, 312 - 326.

Nyffeler, R., 2007. The Closest Relatives of Cacti: Insights from Phylogenetic Analyses of Chloroplast and Mitochondrial Sequences with Special Emphasis on Relationships in the Tribe Anacampseroteae. American Journal of Botany 94, 89 - 101.

Olmstead, R.G., Kim, K.-J., Jansen, R.K., Wagstaff, S.J., 2000. The Phylogeny of the Asteridae sensu lato Based on Chloroplast *ndhF* Gene Sequences. Molecular Phylogenetics and Evolution 16, 96-112.

Rettig, J.H., Wilson, H.D., Manhart, J.R., 1992. Phylogeny of the Caryophyllales - gene sequence data. Taxon 41, 201-209.

Rivadavia, F., Kondo, K., Kato, M., Hasebe, M., 2003. Phylogeny of the sundews, *Drosera* (Droseraceae), based on chloroplast *rbcL* and nuclear 18S ribosomal DNA Sequences. American Journal of Botany 90, 123-130.

Sanchez, A., Kron, K.A., 2008. Phylogenetics of Polygonaceae with an Emphasis on the Evolution of Eriogonoideae. Systematic Botany 33, 87-96.

Savolainen, V., Chase, M.W., Hoot, S.B., Morton, C.M., Soltis, D.E., Bayer, C., Fay, M.F., DeBruijn, A.Y., Sullivan, S., Qiu, Y.-L., 2000a. Phylogenetics of Flowering

Plants Based on Combined Analysis of Plastid *atpB* and *rbcL* Gene Sequences. Systematic Biology 49, 306 - 362.

Savolainen, V., Fay, M.F., Albach, D.C., Backlund, A., van der Bank, M., Cameron, K.M., Johnson, S.A., Lledo, M.D., Pintaud, J.C., Powell, M., Sheahan, M.C., Soltis, D.E., Soltis, P.S., Weston, P., Whitten, W.M., Wurdack, K.J., Chase, M.W., 2000b. Phylogeny of the eudicots: a nearly complete familial analysis based on *rbcL* gene sequences. Kew Bulletin 55, 257-309.

Smissen, R.D., Clement, J.C., Garnock-Jones, P.J., Chambers, G.K., 2002. Subfamilial relationships within Caryophyllaceae as inferred from 5' *ndhF* sequences. American Journal of Botany 89, 1336-1341.

Soltis, P.S., Soltis, D.E., Chase, M.W., 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402, 402-404.

Wagstaff, S.J., Hennion, F., 2007. Evolution and biogeography of *Lyallia* and *Hectorella* (Portulacaceae), geographically isolated sisters from the Southern Hemisphere. Antarctic Science 19, 417-426.

Wang, H., Moore, M.J., Soltis, P.S., Bell, C., Brockington, S.F., Alexandre, R., Davis, C.C., Latvis, M., Manchester, S.R., Soltis, D.E., 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc Natl Acad Sci U S A 106, 3853-3858.

Yamane, K., Yasui, Y., Ohnishi, O., 2003. Intraspecific cpDNA variations of diploid and tetraploid perennial buckwheat, *Fagopyrum cymosum* (Polygonaceae). American Journal of Botany 90, 339-346.

Zurawski, G., Bohnert, H.J., Whitfeld, P.R., Bottomley, W., 1982. Nucleotide sequence of the gene for the M_r 32,000 thylakoid membrane protein from *Spinacia oleracea* and *Nicotiana debneyi* predicts a totally conserved primary translation product of M_r 38,950. Proc Natl Acad Sci U S A 79, 7699-7703.

Appendix D

Supplemental Table 3 for Chapter 2

Table D.1 Species used, their taxonomic affiliation, GenBank numbers, and reference information for the *matK*/*trnK* data added to the MT-51 dataset (see Table B.1) resulting in the *matK*/*trnK* 652 taxon dataset (MT-652). "Complete" and "partial" indicates whether the sequences obtained from GenBank were for the whole genomic region or for portions of it.

		GenBank No./Source	
Family	**Genus and species**	***matK***	***trnK* introns**
Aizoaceae	*Aridaria_noctiflora*	**AY042619** – partial (Cuénoud et al., 2002)	No Sequence
	Delosperma_cooperi	**DQ855843** – partial (Nyffeler, 2007)	No Sequence
	Delosperma_echinatum	**AY042575** – partial (Cuénoud et al., 2002)	No Sequence
	Galenia_pubescens	**AY042589** – partial (Cuénoud et al., 2002)	No Sequence
	Plinthus_cryptocarpus	**AY042633** – partial (Cuénoud et al., 2002)	No Sequence
	Ruschia_schollii	**AY042649** – partial (Cuénoud et al., 2002)	No Sequence
	Trichodiadema_barbatum	**AY042666** – partial (Cuénoud et al., 2002)	No Sequence
Amaranthaceae	*Achyranthes_arborescens*	**AY042534** – partial (Cuénoud et al., 2002)	No Sequence
	Achyranthes_aspera	**AY514815** – complete (Müller and Borsch, 2005b)	**AY514815** – complete (Müller and Borsch, 2005b)
	Achyropsis_leptostachya	**AY998117** – complete (Müller and Borsch, 2005a)	**AY998117** – complete (Müller and Borsch, 2005a)
	Acroglochin_chenopodioides	**AY514826** – complete (Müller and Borsch, 2005b)	**AY514826** – complete (Müller and Borsch, 2005b)
	Aerva_javanica	**AY514793** – complete (Müller and Borsch, 2005b)	**AY514793** – complete (Müller and Borsch, 2005b)
	Aerva_leucura	**DQ317308** – complete	**DQ317308** – complete

	(Müller and Borsch, 2005a)	(Müller and Borsch, 2005a)
Aerva_sanguinolenta	**AM887482** – complete (Sage et al., 2007)	**AM887482** – complete (Sage et al., 2007)
Agriophyllum_squarrosum	**AY514827** – complete (Müller and Borsch, 2005b)	**AY514827** – complete (Müller and Borsch, 2005b)
Allenrolfea_vaginata	**AY514828** – complete (Müller and Borsch, 2005b)	**AY514828** – complete (Müller and Borsch, 2005b)
Allmaniopsis_fruticulosa	**AY998116** – partial (Müller and Borsch, 2005a)	**AY998116** – partial (Müller and Borsch, 2005a)
Alternanthera_altacruzensis	**AM887483** – complete (Sage et al., 2007)	**AM887483** – complete (Sage et al., 2007)
Alternanthera_caracasana	**AF542595** – partial (Hilu et al., 2003)	No Sequence
Alternanthera_flavescens	**AM887484** – complete (Sage et al., 2007)	**AM887484** – complete (Sage et al., 2007)
Alternanthera_microphylla	**AM887485** – complete (Sage et al., 2007)	**AM887485** – complete (Sage et al., 2007)
Alternanthera_pungens	**AY514795** – complete (Müller and Borsch, 2005b)	**AY514795** – complete (Müller and Borsch, 2005b)
Alternanthera_sessilis	**AY514796** – complete (Müller and Borsch, 2005b)	**AY514796** – complete (Müller and Borsch, 2005b)
Amaranthus_acutilobus	**AY042544** – partial (Cuénoud et al., 2002)	No Sequence
Amaranthus_asplundii	**AM887486** – complete (Sage et al., 2007)	**AM887486** – complete (Sage et al., 2007)
Amaranthus_caudatus	**AY514809** – complete (Müller and Borsch, 2005b)	**AY514809** – complete (Müller and Borsch, 2005b)
Amaranthus_greggii	**AY514808** – complete (Müller and Borsch, 2005b)	**AY514808** – complete (Müller and Borsch, 2005b)
Amaranthus_hybridus	**EF590393** – partial (Kress and Erickson, 2007)	No Sequence
Amaranthus_paniculatus	**AF204866** – partial (Meimberg et al., 2000)	No Sequence
Amaranthus_praetermissus	**AM887487** – complete (Sage et al., 2007)	**AM887487** – complete (Sage et al., 2007)
Amaranthus_spinosus	**EF590394** – partial (Kress and Erickson, 2007)	No Sequence

Amaranthus_viridis	**AM887488** – complete (Sage et al., 2007)	**AM887488** – complete (Sage et al., 2007)
Aphanisma_blitoides	**AY514844** – complete (Müller and Borsch, 2005b)	**AY514844** – complete (Müller and Borsch, 2005b)
Arthraerua_leubnitziae	**AY998115** – complete (Müller and Borsch, 2005a)	**AY998115** – complete (Müller and Borsch, 2005a)
Arthrocnemum_glaucum	**AY996303** – partial (Murakeözy et al., 2007)	No Sequence
Arthrocnemum_macrostachyum	**DQ465003** – partial (Pagliano et al, unpublished)	No Sequence
Atriplex_patula	**AY042550** – partial (Cuénoud et al., 2002)	No Sequence
Atriplex_tatarica	**AY936329** – partial (Fior et al., 2006)	No Sequence
Atriplex_truncata	**AY514830** – complete (Müller and Borsch, 2005b)	**AY514830** – complete (Müller and Borsch, 2005b)
Axyris_hybrida	**AY042551** – partial (Cuénoud et al., 2002)	No Sequence
Bassia_hirsuta	**AY514831** – complete (Müller and Borsch, 2005b)	**AY514831** – complete (Müller and Borsch, 2005b)
Bassia_scoparia	**AY042604** – partial (Cuénoud et al., 2002)	No Sequence
Beta_trigyna	**AY042555** – partial (Cuénoud et al., 2002)	No Sequence
Beta_vulgaris	**AY514832** – complete (Müller and Borsch, 2005b)	**AY514832** – complete (Müller and Borsch, 2005b)
Bienertia_cycloptera	**AY514833** – complete (Müller and Borsch, 2005b)	**AY514833** – complete (Müller and Borsch, 2005b)
Blackiella_inflata	**AY042557** – partial (Cuénoud et al., 2002)	No Sequence
Blutaparon_vermiculare	**AY514798** – complete (Müller and Borsch, 2005b)	**AY514798** – complete (Müller and Borsch, 2005b)
Bosea_cypria	**AY042559** – partial (Cuénoud et al., 2002)	No Sequence
Bosea_yervamora	**AY514810** – complete (Müller and Borsch, 2005b)	**AY514810** – complete (Müller and Borsch, 2005b)
Calicorema_capitata	**AY514807** – complete	**AY514807** – complete

	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)
Calicorema_squarrosa	**AY998114** – complete	**AY998114** – complete
	(Müller and Borsch, 2005a)	(Müller and Borsch, 2005a)
Camphorosma_monspeliaca	**AY514829** – complete	**AY514829** – complete
	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)
Centemopsis_kirkii	**AM887526** – complete	**AM887526** – complete
	(Borsch et al., unpublished)	(Borsch et al., unpublished)
Centemopsis_micrantha	**AY998105** – complete	**AY998105** – complete
	(Müller and Borsch, 2005a)	(Müller and Borsch, 2005a)
Centemopsis_trinervis	**AY998107** – complete	**AY998107** – partial
	(Müller and Borsch, 2005a)	(Müller and Borsch, 2005a)
Chamissoa_altissima	**AY514857** – complete	**AY514857** – complete
	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)
Charpentiera_obovata	**AY514855** – complete	**AY514855** – complete
	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)
Chenopodium_acuminatum	**AY514836** – complete	**AY514836** – complete
	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)
Chenopodium_bonus_henricus	**AY042567** – partial	No Sequence
	(Cuénoud et al., 2002)	
Chenopodium_botrys	**AY514835** – complete	**AY514835** – complete
	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)
Cyathula_achyranthoides	**AY514862** – complete	**AY514862** – complete
	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)
Deeringia_amaranthoides	**AY514814** – complete	**AY514814** – complete
	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)
Froelichia_floridana	**AY514799** – complete	**AY514799** – complete
	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)
Froelichia_gracilis	**AY042588** – partial	No Sequence
	(Cuénoud et al., 2002)	
Gomphrena_ferruginea	**AM887524** – complete	**AM887524** – complete
	(Ortuno et al., unpublished)	(Ortuno et al., unpublished)
Gomphrena_fuscipellita	**AM887525** – complete	**AM887525** – complete
	(Ortuno et al., unpublished)	(Ortuno et al., unpublished)
Gomphrena_haageana	**AY514800** – complete	**AY514800** – complete
	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)
Gomphrena_mandonii	**AY514801** – complete	**AY514801** – complete
	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)

Gomphrena_pulchella	**AY514802** – complete (Müller and Borsch, 2005b)	**AY514802** – complete (Müller and Borsch, 2005b)
Hablitzia_tamnoides	**AY042598** – partial (Cuénoud et al., 2002)	No Sequence
Halimione_portulacoides	**DQ468648** – partial (Pagliano et al., unpublished)	No Sequence
Halocnemum_strobilaceum	**DQ468643** – partial (Pagliano et al., unpublished)	No Sequence
Hebanthe_occidentalis	**AY514821** – complete (Müller and Borsch, 2005b)	**AY514821** – complete (Müller and Borsch, 2005b)
Hermbstaedtia_glauca	**AY514812** – complete (Müller and Borsch, 2005b)	**AY514812** – complete (Müller and Borsch, 2005b)
Iresine_alternifolia	**AM887490** – complete (Borsch et al., unpublished)	**AM887490** – complete (Borsch et al., unpublished)
Iresine_cassiniiformis	**AM887489** – complete (Borsch et al., unpublished)	**AM887489** – complete (Borsch et al., unpublished)
Iresine_diffusa_f_lindenii	**AY514805** – complete (Müller and Borsch, 2005b)	**AY514805** – complete (Müller and Borsch, 2005b)
Iresine_herbstii	**AY042603** – partial (Cuénoud et al., 2002)	No Sequence
Iresine_palmeri	**AY514804** – complete (Müller and Borsch, 2005b)	**AY514804** – complete (Müller and Borsch, 2005b)
Kyphocarpa_angustifolia	**AY998111** – complete (Müller and Borsch, 2005a)	**AY998111** – complete (Müller and Borsch, 2005a)
Kyphocarpa_trichinoides	**AY998106** – complete (Müller and Borsch, 2005a)	**AY998106** – partial (Müller and Borsch, 2005a)
Maireana_sedifolia	**AY042613** – partial (Cuénoud et al., 2002)	No Sequence
Marcelliopsis_splendens	**AY998112** – complete (Müller and Borsch, 2005a)	**AY998112** – complete (Müller and Borsch, 2005a)
Mechowia_grandiflora	**AY998113** – complete (Müller and Borsch, 2005a)	**AY998113** – complete (Müller and Borsch, 2005a)
Nitrophila_occidentalis	**AY514840** – complete (Müller and Borsch, 2005b)	**AY514840** – complete (Müller and Borsch, 2005b)
Nothosaerva_brachiata	**AY514806** – complete (Müller and Borsch, 2005b)	**AY514806** – complete (Müller and Borsch, 2005b)
Nototrichium_humile	**AY514816** – complete	**AY514816** – complete

	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)
Oreobliton_thesioides	**AY875638** – complete	**AY875638** – complete
	(Müller and Borsch, 2005a)	(Müller and Borsch, 2005a)
Pandiaka_angustifolia	**AY514818** – complete	**AY514818** – complete
	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)
Patellifolia_procumbens	**AY875637** – complete	**AY875637** – complete
	(Müller and Borsch, 2005a)	(Müller and Borsch, 2005a)
Pedersenia_cardenasii	**AM887491** – complete	**AM887491** – complete
	(Borsch et al., unpublished)	(Borsch et al., unpublished)
Pfaffia_fruticulosa	**AM887492** – complete	**AM887492** – complete
	(Borsch et al., unpublished)	(Borsch et al., unpublished)
Pleuropetalum_sprucei	**AF542596** – partial	No Sequence
	(Hilu et al., 2003)	
Polycnemum_majus	**AY514839** – complete	**AY514839** – complete
	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)
Polycnemum_verrucosum	**AY514838** – complete	**AY514838** – complete
	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)
Pseudoplantago_friesii	**AY514820** – complete	**AY514820** – complete
	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)
Psilotrichum_africanum	**AY514822** – complete	**AY514822** – complete
	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)
Psilotrichum_ferrugineum	**AY998108** – complete	**AY998108** – complete
	(Müller and Borsch, 2005a)	(Müller and Borsch, 2005a)
Psilotrichum_gnaphalobryum	**AY042638** – partial	No Sequence
	(Cuénoud et al., 2002)	
Psilotrichum_sericeum	**AY998109** – complete	**AY998109** – complete
	(Müller and Borsch, 2005a)	(Müller and Borsch, 2005a)
Ptilotus_manglesii	**AY042640** – partial	No Sequence
	(Cuénoud et al., 2002)	
Ptilotus_obovatus	**AY514823** – complete	**AY514823** – complete
	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)
Pupalia_lappacea	**AY514858** – complete	**AY514858** – complete
	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)
Rhagodia_baccata	**AY042643** – partial	No Sequence
	(Cuénoud et al., 2002)	
Salicornia_brachystachya	**DQ864702** – partial	No Sequence
	(Pagliano et al., unpublished)	

Salicornia_disarticulata	**AY996304** – partial (Murakeözy et al., 2007)	No Sequence
Salicornia_dolichostachya	**DQ468650** – partial (Pagliano et al., unpublished)	No Sequence
Salicornia_emerici	**DQ468652** – partial (Pagliano et al., unpublished)	No Sequence
Salicornia_fragilis	**AY996312** – partial (Murakeözy et al., 2007)	No Sequence
Salicornia_obscura	**DQ468651** – partial (Pagliano et al., unpublished)	No Sequence
Salicornia_patula	**DQ468644** – partial (Pagliano et al., unpublished)	No Sequence
Salicornia_ramosissima	**AY996317** – partial (Murakeözy et al., 2007)	No Sequence
Salicornia_veneta	**DQ468649** – partial (Pagliano et al., unpublished)	No Sequence
Salsola_kali	**AY514843** – complete (Müller and Borsch, 2005b)	**AY514843** – complete (Müller and Borsch, 2005b)
Sarcocornia_fruticosa	**DQ468645** – partial (Pagliano et al., unpublished)	No Sequence
Sarcocornia_perennis	**DQ468646** – partial (Pagliano et al., unpublished)	No Sequence
Sericocoma_avolans	**AY998103** – complete (Müller and Borsch, 2005a)	**AY998103** – complete (Müller and Borsch, 2005)
Sericocoma_heterochiton	**AY998104** – complete (Müller and Borsch, 2005a)	**AY998104** – complete (Müller and Borsch, 2005a)
Sericorema_sericea	**AY998110** – complete (Müller and Borsch, 2005a)	**AY998110** – partial (Müller and Borsch, 2005a)
Sericostachys_scandens	**AY514819** – complete (Müller and Borsch, 2005b)	**AY514819** – complete (Müller and Borsch, 2005b)
Suaeda_maritima	**DQ468647** – partial (Pagliano et al., unpublished)	No Sequence
Suaeda_vera	**AY042658** – partial (Cuénoud et al., 2002)	No Sequence
Tidestromia_lanuginosa	**AY514797** – complete (Müller and Borsch, 2005b)	**AY514797** – complete (Müller and Borsch, 2005b)
Xerosiphon_aphyllus	**AM887523** – complete	**AM887523** – complete

		(Borsch et al., unpublished)	(Borsch et al., unpublished)
Ancistrocladaceae	*Ancistrocladus_hamatus*	**AF204842** – partial (Meimberg et al., 2000)	No Sequence
	Ancistrocladus_heyneanus	**AF204841** – partial (Meimberg et al., 2000)	No Sequence
	Ancistrocladus_korupensis	**AY042546** – partial (Cuénoud et al., 2002)	No Sequence
Barbeuiaceae	*Barbeuia_madagascariensis*	**AY042552** – partial (Cuénoud et al., 2002)	No Sequence
Basellaceae	*Anredera_cordifolia*	**AY042547** – partial (Cuénoud et al., 2002)	No Sequence
Cactaceae	*Acanthocalycium_glaucum*	**AY015325** – complete (Nyffeler, 2002)	**AY015325** – complete (Nyffeler, 2002)
	Acanthocereus_pentagonus	**AY015295** – complete (Nyffeler, 2002)	**AY015295** – complete (Nyffeler, 2002)
	Armatocereus_godingianus	**AY015296** – complete (Nyffeler, 2002)	**AY015296** – complete (Nyffeler, 2002)
	Astrophytum_myriostigma	**AY015288** – complete (Nyffeler, 2002)	**AY015288** – complete (Nyffeler, 2002)
	Austrocactus_bertinii	**AY015300** – complete (Nyffeler, 2002)	**AY015300** – complete (Nyffeler, 2002)
	Austrocylindropuntia_vestita	**AY015278** – complete (Nyffeler, 2002)	**AY015278** – complete (Nyffeler, 2002)
	Aztekium_ritteri	**AY015290** – complete (Nyffeler, 2002)	**AY015290** – complete (Nyffeler, 2002)
	Brasiliopuntia_brasiliensis	**AY875370** – complete (Edwards et al., 2005)	**AY875370** – partial (Edwards et al., 2005)
	Browningia_hertlingiana	**AY015315** – complete (Nyffeler, 2002)	**AY015315** – complete (Nyffeler, 2002)
	Calymmanthium_substerile	**AY015291** – complete (Nyffeler, 2002)	**AY015291** – complete (Nyffeler, 2002)
	Castellanosia_caineana	**AY015298** – complete (Nyffeler, 2002)	**AY015298** – complete (Nyffeler, 2002)
	Cereus_alacriportanus	**AY015313** – complete (Nyffeler, 2002)	**AY015313** – complete (Nyffeler, 2002)
	Coleocephalocereus_fluminensis	**AY015318** – complete (Nyffeler, 2002)	**AY015318** – complete (Nyffeler, 2002)

Copiapoa_bridgesii	**AY015293** – complete (Nyffeler, 2002)	**AY015293** – complete (Nyffeler, 2002)
Copiapoa_laui	**AY015294** – complete (Nyffeler, 2002)	**AY015294** – complete (Nyffeler, 2002)
Copiapoa_solaris	**AY015292** – complete (Nyffeler, 2002)	**AY015292** – complete (Nyffeler, 2002)
Corryocactus_brevistylus	**AY015302** – complete (Nyffeler, 2002)	**AY015302** – complete (Nyffeler, 2002)
Corryocactus_tenuiculus	**AY015303** – complete (Nyffeler, 2002)	**AY015303** – complete (Nyffeler, 2002)
Disocactus_amazonicus	**AY015312** – complete (Nyffeler, 2002)	**AY015312** – complete (Nyffeler, 2002)
Echinocactus_platyacanthus	**AY015287** – complete (Nyffeler, 2002)	**AY015287** – complete (Nyffeler, 2002)
Echinocereus_pentalophus	**AY015307** – complete (Nyffeler, 2002)	**AY015307** – complete (Nyffeler, 2002)
Echinopsis_chiloensis	**AY015322** – complete (Nyffeler, 2002)	**AY015322** – complete (Nyffeler, 2002)
Echinopsis_pentlandii	**AY015323** – complete (Nyffeler, 2002)	**AY015323** – complete (Nyffeler, 2002)
Eriosyce_aurata	**AY015336** – complete (Nyffeler, 2002)	**AY015336** – complete (Nyffeler, 2002)
Eriosyce_islayensis	**AY015337** – complete (Nyffeler, 2002)	**AY015337** – complete (Nyffeler, 2002)
Eriosyce_napina	**AY015339** – complete (Nyffeler, 2002)	**AY015339** – complete (Nyffeler, 2002)
Eriosyce_subgibbosa	**AY015338** – complete (Nyffeler, 2002)	**AY015338** – complete (Nyffeler, 2002)
Escontria_chiotilla	**AY015308** – complete (Nyffeler, 2002)	**AY015308** – complete (Nyffeler, 2002)
Eulychnia_iquiquensis	**AY015301** – complete (Nyffeler, 2002)	**AY015301** – complete (Nyffeler, 2002)
Frailea_gracillima	**AY015285** – complete (Nyffeler, 2002)	**AY015285** – complete (Nyffeler, 2002)
Frailea_phaeodisca	**AY015286** – complete (Nyffeler, 2002)	**AY015286** – complete (Nyffeler, 2002)
Gymnocalycium_denudatum	**AY015317** – complete	**AY015317** – complete

	(Nyffeler, 2002)	(Nyffeler, 2002)
Haageocereus_pseudomelanostele	**AY015329** – complete (Nyffeler, 2002)	**AY015329** – complete (Nyffeler, 2002)
Harrisia_pomanensis	**AY015324** – complete (Nyffeler, 2002)	**AY015324** – complete (Nyffeler, 2002)
Hatiora_salicornioides	**AY015341** – complete (Nyffeler, 2002)	**AY015341** – complete (Nyffeler, 2002)
Hylocereus_peruvianus	**AY015310** – complete (Nyffeler, 2002)	**AY015310** – complete (Nyffeler, 2002)
Lepismium_cruciforme	**AY015344** – complete (Nyffeler, 2002)	**AY015344** – complete (Nyffeler, 2002)
Leptocereus_leonii	**AY015297** – complete (Nyffeler, 2002)	**AY015297** – complete (Nyffeler, 2002)
Maihuenia_patagonica	**AY015281** – complete (Nyffeler, 2002)	**AY015281** – complete (Nyffeler, 2002)
Maihuenia_poeppigii	**AY015282** – complete (Nyffeler, 2002)	**AY015282** – complete (Nyffeler, 2002)
Maihueniopsis_subterranea	**EU834746** – complete (Nyffeler and Eggli, 2010)	No Sequence
Mammillaria_haageana	**AY015289** – complete (Nyffeler, 2002)	**AY015289** – complete (Nyffeler, 2002)
Matucana_intertexta	**AY015327** – complete (Nyffeler, 2002)	**AY015327** – complete (Nyffeler, 2002)
Micranthocereus_albicephalus	**AY015314** – complete (Nyffeler, 2002)	**AY015314** – complete (Nyffeler, 2002)
Neoraimondia_arequipensis	**AY015299** – complete (Nyffeler, 2002)	**AY015299** – complete (Nyffeler, 2002)
Neowerdermannia_vorwerkii	**AY015340** – complete (Nyffeler, 2002)	**AY015340** – complete (Nyffeler, 2002)
Opuntia_dillenii	**AY875369** – complete (Edwards et al., 2005)	**AY875369** – partial (Edwards et al., 2005)
Opuntia_fragilis	**EF590413** – partial (Kress and Erickson, 2007)	No Sequence
Opuntia_microdasys	**AY042622** – partial (Cuénoud et al., 2002)	No Sequence
Oreocereus_celsianus	**AY015328** – complete (Nyffeler, 2002)	**AY015328** – complete (Nyffeler, 2002)

Pachycereus_schottii	**AY015309** – complete (Nyffeler, 2002)	**AY015309** – complete (Nyffeler, 2002)
Parodia_buenekeri	**AY015331** – complete (Nyffeler, 2002)	**AY015331** – complete (Nyffeler, 2002)
Parodia_haselbergii	**AY015330** – complete (Nyffeler, 2002)	**AY015330** – complete (Nyffeler, 2002)
Parodia_maassii	**AY015333** – complete (Nyffeler, 2002)	**AY015333** – complete (Nyffeler, 2002)
Parodia_magnifica	**AY015332** – complete (Nyffeler, 2002)	**AY015332** – complete (Nyffeler, 2002)
Parodia_microsperma	**AY015334** – complete (Nyffeler, 2002)	**AY015334** – complete (Nyffeler, 2002)
Parodia_ottonis	**AY015335** – complete (Nyffeler, 2002)	**AY015335** – complete (Nyffeler, 2002)
Pereskia_aureiflora	**AY875354** – complete (Edwards et al., 2005)	**AY875354** – complete (Edwards et al., 2005)
Pereskia_bleo	**AY875359** – complete (Edwards et al., 2005)	**AY875359** – complete (Edwards et al., 2005)
Pereskia_diaz_romeroana	**AY875353** – complete (Edwards et al., 2005)	**AY875353** – partial (Edwards et al., 2005)
Pereskia_grandifolia_var_grandifolia	**AY875362** – complete (Edwards et al., 2005)	**AY875362** – complete (Edwards et al., 2005)
Pereskia_guamacho	**AY015275** – complete (Nyffeler, 2002)	**AY015275** – complete (Nyffeler, 2002)
Pereskia_horrida	**AY875356** – complete (Edwards et al., 2005)	**AY875356** – complete (Edwards et al., 2005)
Pereskia_lychnidiflora	**AY875358** – complete (Edwards et al., 2005)	**AY875358** – complete (Edwards et al., 2005)
Pereskia_marcanoi	**AY875360** – complete (Edwards et al., 2005)	**AY875360** – complete (Edwards et al., 2005)
Pereskia_nemorosa	**AY875350** – complete (Edwards et al., 2005)	**AY875350** – partial (Edwards et al., 2005)
Pereskia_portulacifolia	**AY875361** – complete (Edwards et al., 2005)	**AY875361** – complete (Edwards et al., 2005)
Pereskia_quisqueyana	**AY875352** – complete (Edwards et al., 2005)	**AY875352** – partial (Edwards et al., 2005)
Pereskia_sacharosa	**AY875363** – complete	**AY875363** – complete

	(Edwards et al., 2005)	(Edwards et al., 2005)
Pereskia_stenantha	**AY015276** – complete (Nyffeler, 2002)	**AY015276** – complete (Nyffeler, 2002)
Pereskia_zinniiflora	**AY015277** – complete (Nyffeler, 2002)	**AY015277** – complete (Nyffeler, 2002)
Pereskiopsis_deguetii	**AY015280** – complete (Nyffeler, 2002)	**AY015280** – complete (Nyffeler, 2002)
Pfeiffera_ianthothele	**AY015304** – complete (Nyffeler, 2002)	**AY015304** – complete (Nyffeler, 2002)
Pfeiffera_miyagawae	**AY015305** – complete (Nyffeler, 2002)	**AY015305** – complete (Nyffeler, 2002)
Pfeiffera_monacantha	**AY015306** – complete (Nyffeler, 2002)	**AY015306** – complete (Nyffeler, 2002)
Quiabentia_verticillata	**AY042641** – partial (Cuénoud et al., 2002)	No Sequence
Quiabentia_zehntneri	**AY875372** – complete (Edwards et al., 2005)	**AY875372** – partial (Edwards et al., 2005)
Rauhocereus_riosaniensis	**AY015326** – complete (Nyffeler, 2002)	**AY015326** – complete (Nyffeler, 2002)
Rhipsalis_floccosa	**AY015342** – complete (Nyffeler, 2002)	**AY015342** – complete (Nyffeler, 2002)
Rhipsalis_teres	**AY042645** – partial (Cuénoud et al., 2002)	No Sequence
Samaipaticereus_corroanus	**AY015321** – complete (Nyffeler, 2002)	**AY015321** – complete (Nyffeler, 2002)
Schlumbergera_truncata	**AY015343** – complete (Nyffeler, 2002)	**AY015343** – complete (Nyffeler, 2002)
Selenicereus_boeckmannii	**AY015311** – complete (Nyffeler, 2002)	**AY015311** – complete (Nyffeler, 2002)
Stetsonia_coryne	**AY015320** – complete (Nyffeler, 2002)	**AY015320** – complete (Nyffeler, 2002)
Tacinga_funalis	**AY042660** – partial (Cuénoud et al., 2002)	No Sequence
Tephrocactus_articulatus	**AY875367** – complete (Edwards et al., 2005)	**AY875367** – partial (Edwards et al., 2005)
Uebelmannia_pectinifera	**AY015319** – complete (Nyffeler, 2002)	**AY015319** – complete (Nyffeler, 2002)

Caryophyllaceae	*Acanthophyllum_sordidum*	**AY936324** – partial (Fior et al., 2006)	No Sequence
	Agrostemma_githago	**AY042539** – partial (Cuénoud et al., 2002)	No Sequence
	Allochrusa_versicolor	**AY936323** – partial (Fior et al., 2006)	No Sequence
	Arenaria_balearica	**DQ901521** – partial (Fior and Karis, 2007)	No Sequence
	Arenaria_bertolonii	**DQ901520** – partial (Fior and Karis, 2007)	No Sequence
	Arenaria_ciliata	**DQ901542** – partial (Fior and Karis, 2007)	No Sequence
	Arenaria_digyna	**AY936304** – partial (Fior et al., 2006)	No Sequence
	Arenaria_koriniana	**AY936318** – partial (Fior et al., 2006)	No Sequence
	Arenaria_lanuginosa	**DQ901544** – partial (Fior and Karis, 2007)	No Sequence
	Arenaria_moehringioides	**DQ901543** – partial (Fior and Karis, 2007)	No Sequence
	Arenaria_musciformis	**DQ901545** – partial (Fior and Karis, 2007)	No Sequence
	Arenaria_nevadensis	**AY936303** – partial (Fior et al., 2006)	No Sequence
	Arenaria_pogonantha	**AY936300** – partial (Fior et al., 2006)	No Sequence
	Arenaria_serpyllifolia	**AY936302** – partial (Fior et al., 2006)	No Sequence
	Arenaria_syreistschikowii	**AY936317** – partial (Fior et al., 2006)	No Sequence
	Arenaria_tetraquetra	**AF400155** – partial (Cuénoud et al., 2002)	No Sequence
	Arenaria_trichophora	**AY936301** – partial (Fior et al., 2006)	No Sequence
	Atocion_rupestre	**EF547242** – complete (Mower et al., 2007)	**EF547242** – complete (Mower et al., 2007)
	Bufonia_tenuifolia	**AY936289** – partial	No Sequence

	(Fior et al., 2006)	
Bufonia_wilkommiana	**AY936290** – partial	No Sequence
	(Fior et al., 2006)	
Cerastium_arvense	**AY936295** – partial	No Sequence
	(Fior et al., 2006)	
Cerastium_fontanum	**AY936296** – partial	No Sequence
	(Fior et al., 2006)	
Corrigiola_litoralis	**AY936331** – partial	No Sequence
	(Fior et al., 2006)	
Dianthus_furcatus	**AY936320** – partial	No Sequence
	(Fior et al., 2006)	
Dianthus_seguieri	**AY936321** – partial	No Sequence
	(Fior et al., 2006)	
Drypis_spinosa	**AY936293** – partial	No Sequence
	(Fior et al., 2006)	
Gypsophila_altissima	**AY042597** – partial	No Sequence
	(Cuénoud et al., 2002)	
Gypsophila_elegans	**AY936327** – partial	No Sequence
	(Fior et al., 2006)	
Gypsophila_repens	**AY936326** – partial	No Sequence
	(Fior et al., 2006)	
Herniaria_baetica	**AY936283** – partial	No Sequence
	(Fior et al., 2006)	
Holosteum_umbellatum	**AY936297** – partial	No Sequence
	(Fior et al., 2006)	
Honckenya_peploides	**AY042602** – partial	No Sequence
	(Cuénoud et al., 2002)	
Loeflingia_hispanica	**AY936288** – partial	No Sequence
	(Fior et al., 2006)	
Lychnis_coronaria	**AY042612** – partial	No Sequence
	(Cuénoud et al., 2002)	
Lychnis_flos_jovis	**AY936313** – partial	No Sequence
	(Fior et al., 2006)	
Minuartia_geniculata	**AY936307** – partial	No Sequence
	(Fior et al., 2006)	
Minuartia_graminifolia	**AY936316** – partial	No Sequence
	(Fior et al., 2006)	

Minuartia_lanceolata	**AY936292** – partial (Fior et al., 2006)	No Sequence
Minuartia_laricifolia	**AY936294** – partial (Fior et al., 2006)	No Sequence
Minuartia_picta	**AY936319** – partial (Fior et al., 2006)	No Sequence
Moehringia_bavarica	**DQ901529** – partial (Fior and Karis, 2007)	No Sequence
Moehringia_ciliata	**DQ901538** – partial (Fior and Karis, 2007)	No Sequence
Moehringia_concarenae	**DQ901530** – partial (Fior and Karis, 2007)	No Sequence
Moehringia_dielsiana	**DQ901533** – partial (Fior and Karis, 2007)	No Sequence
Moehringia_diversifolia	**DQ901540** – partial (Fior and Karis, 2007)	No Sequence
Moehringia_fontqueri	**DQ901514** – partial (Fior and Karis, 2007)	No Sequence
Moehringia_glaucovirens	**DQ901534** – partial (Fior and Karis, 2007)	No Sequence
Moehringia_glochidisperma	**DQ901519** – partial (Fior and Karis, 2007)	No Sequence
Moehringia_grisebachii	**DQ901523** – partial (Fior and Karis, 2007)	No Sequence
Moehringia_insubrica	**DQ901539** – partial (Fior and Karis, 2007)	No Sequence
Moehringia_intermedia	**DQ901531** – partial (Fior and Karis, 2007)	No Sequence
Moehringia_intricata	**AY936305** – partial (Fior et al., 2006)	No Sequence
Moehringia_jankae	**DQ901524** – partial (Fior and Karis, 2007)	No Sequence
Moehringia_lateriflora	**DQ901525** – partial (Fior and Karis, 2007)	No Sequence
Moehringia_lebrunii	**DQ901541** – partial (Fior and Karis, 2007)	No Sequence
Moehringia_markgrafii	**DQ901532** – partial	No Sequence

	(Fior and Karis, 2007)	
Moehringia_minutiflora	**DQ901526** – partial	No Sequence
	(Fior and Karis, 2007)	
Moehringia_muscosa	**AY936306** – partial	No Sequence
	(Fior et al., 2006)	
Moehringia_papulosa	**DQ901535** – partial	No Sequence
	(Fior and Karis, 2007)	
Moehringia_pendula	**DQ901528** – partial	No Sequence
	(Fior and Karis, 2007)	
Moehringia_pentandra	**DQ901522** – partial	No Sequence
	(Fior and Karis, 2007)	
Moehringia_sedoides	**DQ901537** – partial	No Sequence
	(Fior and Karis, 2007)	
Moehringia_stellarioides	**DQ901527** – partial	No Sequence
	(Fior and Karis, 2007)	
Moehringia_tejedensis	**DQ901518** – partial	No Sequence
	(Fior and Karis, 2007)	
Moehringia_trinervia	**AY042615** – partial	No Sequence
	(Cuénoud et al., 2002)	
Ortegia_hispanica	**AY936286** – partial	No Sequence
	(Fior et al., 2006)	
Paronychia_echinulata	**AY936285** – partial	No Sequence
	(Fior et al., 2006)	
Paronychia_kapela	**AY936284** – partial	No Sequence
	(Fior et al., 2006)	
Petrocoptis_pyrenaica	**AY936314** – partial	No Sequence
	(Fior et al., 2006)	
Polycarpon_tetraphyllum	**AY936287** – partial	No Sequence
	(Fior et al., 2006)	
Sagina_pilifera	**AY936291** – partial	No Sequence
	(Fior et al., 2006)	
Saponaria_ocymoides	**AY042651** – partial	No Sequence
	(Cuénoud et al., 2002)	
Saponaria_officinalis	**AY936325** – partial	No Sequence
	(Fior et al., 2006)	
Schiedea_adamantis	**DQ907802** – partial	No Sequence
	(Kapralov and Filatov, 2006)	

Schiedea_aprokremnos	**DQ907803** – partial (Kapralov and Filatov, 2006)	No Sequence
Schiedea_globosa	**DQ907804** – partial (Kapralov and Filatov, 2006)	No Sequence
Schiedea_haleakalensis	**DQ907805** – partial (Kapralov and Filatov, 2006)	No Sequence
Schiedea_helleri	**DQ907806** – partial (Kapralov and Filatov, 2006)	No Sequence
Schiedea_hookerii	**DQ907807** – partial (Kapralov and Filatov, 2006)	No Sequence
Schiedea_jacobii	**DQ907808** – partial (Kapralov and Filatov, 2006)	No Sequence
Schiedea_kaalae	**DQ907809** – partial (Kapralov and Filatov, 2006)	No Sequence
Schiedea_kauaiensis	**DQ907810** – partial (Kapralov and Filatov, 2006)	No Sequence
Schiedea_kealiae	**DQ907811** – partial (Kapralov and Filatov, 2006)	No Sequence
Schiedea_laui	**DQ907812** – partial (Kapralov and Filatov, 2006)	No Sequence
Schiedea_ligustrina	**DQ907813** – partial (Kapralov and Filatov, 2006)	No Sequence
Schiedea_lydgatei	**DQ907814** – partial (Kapralov and Filatov, 2006)	No Sequence
Schiedea_mannii	**DQ907815** – partial (Kapralov and Filatov, 2006)	No Sequence
Schiedea_membranacea	**DQ907816** – partial (Kapralov and Filatov, 2006)	No Sequence
Schiedea_menziesii	**DQ907817** – partial (Kapralov and Filatov, 2006)	No Sequence
Schiedea_nuttallii	**DQ907818** – partial (Kapralov and Filatov, 2006)	No Sequence
Schiedea_obovata	**DQ907819** – partial (Kapralov and Filatov, 2006)	No Sequence
Schiedea_pentandra	**DQ907820** – partial (Kapralov and Filatov, 2006)	No Sequence
Schiedea_perlmannii	**DQ907821** – partial	No Sequence

	(Kapralov and Filatov, 2006)	
Schiedea_salicaria	**DQ907822** – partial	No Sequence
	(Kapralov and Filatov, 2006)	
Schiedea_sarmentosa	**DQ907823** – partial	No Sequence
	(Kapralov and Filatov, 2006)	
Schiedea_spergulina	**DQ907824** – partial	No Sequence
	(Kapralov and Filatov, 2006)	
Schiedea_stellarioides	**DQ907825** – partial	No Sequence
	(Kapralov and Filatov, 2006)	
Schiedea_trinervis	**DQ907826** – partial	No Sequence
	(Kapralov and Filatov, 2006)	
Schiedea_verticillata	**DQ907827** – partial	No Sequence
	(Kapralov and Filatov, 2006)	
Schiedea_viscosa	**DQ907828** – partial	No Sequence
	(Kapralov and Filatov, 2006)	
Silene_campanula	**AY936311** – partial	No Sequence
	(Fior et al., 2006)	
Silene_diclinis	**EF646877** – partial	No Sequence
	(Muir and Filatov, 2007)	
Silene_dioica	**EF646900** – partial	No Sequence
	(Muir and Filatov, 2007)	
Silene_douglasii	**EF547238** – complete	**EF547238** – complete
	(Mower et al., 2007)	(Mower et al., 2007)
Silene_heuffelii	**EF646876** – partial	No Sequence
	(Muir and Filatov, 2007)	
Silene_italica	**AY936312** – partial	No Sequence
	(Fior et al., 2006)	
Silene_latifolia	**AY707959** – partial	No Sequence
	(Kejnovsky et al., 2006)	
Silene_marizii	**EF646873** – partial	No Sequence
	(Muir and Filatov, 2007)	
Silene_noctiflora	**EF547240** – complete	**EF547240** – complete
	(Mower et al., 2007)	(Mower et al., 2007)
Silene_nutans	**AF542598** – partial	No Sequence
	(Hilu et al., 2003)	
Silene_otites	**AY514848** – complete	**AY514848** – complete
	(Müller and Borsch, 2005b)	(Müller and Borsch, 2005b)

	Silene_rothmaleri	**AY042656** – partial (Cuénoud et al., 2002)	No Sequence
	Silene_scouleri	**EF547243** – complete (Mower et al., 2007)	**EF547243** – complete (Mower et al., 2007)
	Silene_uniflora	**DQ841761** – partial (Houliston and Olson, 2006)	No Sequence
	Silene_virginica	**EF547244** – complete (Mower et al., 2007)	**EF547244** – complete (Mower et al., 2007)
	Silene_vulgaris	**EU749399** – partial (Fazekas et al., 2008)	No Sequence
	Spergula_arvensis	**AY936310** – partial (Fior et al., 2006)	No Sequence
	Spergularia_marina	**AY936309** – partial (Fior et al., 2006)	No Sequence
	Spergularia_rubra	**AY936308** – partial (Fior et al., 2006)	No Sequence
	Stellaria_nemorum	**AY936298** – partial (Fior et al., 2006)	No Sequence
	Vaccaria_hispanica	**AY042669** – partial (Cuénoud et al., 2002)	No Sequence
	Velezia_rigida	**AY936322** – partial (Fior et al., 2006)	No Sequence
Didiereaceae	*Decarya_madagascariensis*	**AY042574** – partial (Cuénoud et al., 2002)	No Sequence
	Didierea_trollii	**AY042576** – partial (Cuénoud et al., 2002)	No Sequence
Dioncophyllaceae	*Dioncophyllum_tholloni*	**AF204844** – partial (Meimberg et al., 2000)	No Sequence
	Habropetalum_dawei	**AF204845** – partial (Meimberg et al., 2000)	No Sequence
Droseraceae	*Aldrovanda_vesiculosa*	**AY096120** – partial (Cameron et al., 2002)	No Sequence
	Dionaea_muscipula	**AY042578** – partial (Cameron et al., 2002)	No Sequence
	Drosera_adelae	**AY096121** – partial (Cameron et al., 2002)	No Sequence
	Drosera_aliciae	**AF204849** – partial	No Sequence

		(Meimberg et al., 2000)	
	Drosera_capillaris	**AF204850** – partial (Meimberg et al., 2000)	No Sequence
	Drosera_communis	**AY042579** – partial (Cuénoud et al., 2002)	No Sequence
	Drosera_regia	**AF204848** – partial (Meimberg et al., 2000)	No Sequence
Frankeniaceae	*Frankenia_corymbosa*	**AY042587** – partial (Cuénoud et al., 2002)	No Sequence
Molluginaceae	*Corbichonia_decumbens*	**AY042572** – partial (Cuénoud et al., 2002)	No Sequence
	Glinus_lotoides	**AY042592** – partial (Cuénoud et al., 2002)	No Sequence
	Glischrothamnus_ulei	**AY042593** – partial (Cuénoud et al., 2002)	No Sequence
	Lophiocarpus_burchellii	**AY042611** – partial (Cuénoud et al., 2002)	No Sequence
	Suessenguthiella_scleranthoides	**AY042659** – partial (Cuénoud et al., 2002)	No Sequence
Nepenthaceae	*Nepenthes_adnata*	**AF315866** – complete (Meimberg et al., 2001)	**AF315866** – complete (Meimberg et al., 2001)
	Nepenthes_albomarginata	**AF315908** – complete (Meimberg et al., 2001)	**AF315908** – complete (Meimberg et al., 2001)
	Nepenthes_ampullaria	**AF315888** – complete (Meimberg et al., 2001)	**AF315888** – complete (Meimberg et al., 2001)
	Nepenthes_aristolochioides	**DQ007088** – complete (Meimberg et al., 2006)	**DQ007088** – complete (Meimberg et al., 2006)
	Nepenthes_bellii	**AF315926** – complete (Meimberg et al., 2001)	**AF315926** – complete (Meimberg et al., 2001)
	Nepenthes_bicalcarata	**DQ007089** – complete (Meimberg et al., 2006)	**DQ007089** – partial (Meimberg et al., 2006)
	Nepenthes_bongso	**AF315865** – complete (Meimberg et al., 2001)	**AF315865** – complete (Meimberg et al., 2001)
	Nepenthes_boschiana	**AF315903** – complete (Meimberg et al., 2001)	**AF315903** – complete (Meimberg et al., 2001)
	Nepenthes_burbidgeae	**AF315921** – complete (Meimberg et al., 2001)	**AF315921** – complete (Meimberg et al., 2001)

Nepenthes_clipeata	**AF315878** – complete (Meimberg et al., 2001)	**AF315878** – complete (Meimberg et al., 2001)
Nepenthes_danseri	**DQ007087** – complete (Meimberg et al., 2006)	**DQ007087** – complete (Meimberg et al., 2006)
Nepenthes_densiflora	**AF315927** – complete (Meimberg et al., 2001)	**AF315927** – complete (Meimberg et al., 2001)
Nepenthes_diatas	**AF315915** – complete (Meimberg et al., 2001)	**AF315915** – complete (Meimberg et al., 2001)
Nepenthes_distillatoria	**AF204838** – partial (Meimberg et al., 2000)	No Sequence
Nepenthes_dubia	**AF315869** – complete (Meimberg et al., 2001)	**AF315869** – complete (Meimberg et al., 2001)
Nepenthes_edwardsiana	**DQ840248** – complete (Meimberg and Heubl, 2006)	**DQ840248** – complete (Meimberg and Heubl, 2006)
Nepenthes_ephippiata	**AF315906** – complete (Meimberg et al., 2001)	**AF315906** – complete (Meimberg et al., 2001)
Nepenthes_eustachya	**AF315867** – complete (Meimberg et al., 2001)	**AF315867** – complete (Meimberg et al., 2001)
Nepenthes_eymae	**AF315930** – complete (Meimberg et al., 2001)	**AF315930** – complete (Meimberg et al., 2001)
Nepenthes_faizaliana	**AF315917** – complete (Meimberg et al., 2001)	**AF315917** – complete (Meimberg et al., 2001)
Nepenthes_fusca	**AF315936** – complete (Meimberg et al., 2001)	**AF315936** – complete (Meimberg et al., 2001)
Nepenthes_glabrata	**AF315928** – complete (Meimberg et al., 2001)	**AF315928** – complete (Meimberg et al., 2001)
Nepenthes_gracilis	**AF315937** – complete (Meimberg et al., 2001)	**AF315937** – complete (Meimberg et al., 2001)
Nepenthes_gracillima	**DQ007086** – complete (Meimberg et al., 2006)	**DQ007086** – complete (Meimberg et al., 2006)
Nepenthes_gymnamphora	**AF315864** – complete (Meimberg et al., 2001)	**AF315864** – complete (Meimberg et al., 2001)
Nepenthes_hamata	**AF315914** – complete (Meimberg et al., 2001)	**AF315914** – complete (Meimberg et al., 2001)
Nepenthes_hirsuta	**AF315889** – complete (Meimberg et al., 2001)	**AF315889** – complete (Meimberg et al., 2001)
Nepenthes_inermis	**AF315870** – complete	**AF315870** – complete

	(Meimberg et al., 2001)	(Meimberg et al., 2001)
Nepenthes_insignis	**AF315882** – complete	**AF315882** – complete
	(Meimberg et al., 2001)	(Meimberg et al., 2001)
Nepenthes_khasiana	**AF204836** – partial	No Sequence
	(Meimberg et al., 2000)	
Nepenthes_lamii	**AF315905** – complete	**AF315905** – complete
	(Meimberg et al., 2001)	(Meimberg et al., 2001)
Nepenthes_lavicola	**AF315935** – complete	**AF315935** – complete
	(Meimberg et al., 2001)	(Meimberg et al., 2001)
Nepenthes_longifolia	**AF315871** – complete	**AF315871** – complete
	(Meimberg et al., 2001)	(Meimberg et al., 2001)
Nepenthes_lowii	**AF315876** – complete	**AF315876** – complete
	(Meimberg et al., 2001)	(Meimberg et al., 2001)
Nepenthes_macfarlanei	**AF204832** – partial	No Sequence
	(Meimberg et al., 2000)	
Nepenthes_macrophylla	**AF315931** – complete	**AF315931** – complete
	(Meimberg et al., 2001)	(Meimberg et al., 2001)
Nepenthes_macrovulgaris	**AF315934** – complete	**AF315934** – complete
	(Meimberg et al., 2001)	(Meimberg et al., 2001)
Nepenthes_madagascariensis	**AF204835** – partial	No Sequence
	(Meimberg et al., 2000)	
Nepenthes_mapuluensis	**AF315918** – complete	**AF315918** – complete
	(Meimberg et al., 2001)	(Meimberg et al., 2001)
Nepenthes_masoalensis	**AF315884** – complete	**AF315884** – complete
	(Meimberg et al., 2001)	(Meimberg et al., 2001)
Nepenthes_maxima	**AF315913** – complete	**AF315913** – complete
	(Meimberg et al., 2001)	(Meimberg et al., 2001)
Nepenthes_merrilliana	**AF315912** – complete	**AF315912** – complete
	(Meimberg et al., 2001)	(Meimberg et al., 2001)
Nepenthes_mikei	**AF315911** – complete	**AF315911** – complete
	(Meimberg et al., 2001)	(Meimberg et al., 2001)
Nepenthes_mira	**DQ007085** – complete	**DQ007085** – complete
	(Meimberg et al., 2006)	(Meimberg et al., 2006)
Nepenthes_mirabilis	**AF315920** – complete	**AF315920** – complete
	(Meimberg et al., 2001)	(Meimberg et al., 2001)
Nepenthes_muluensis	**AF315933** – complete	**AF315933** – complete
	(Meimberg et al., 2001)	(Meimberg et al., 2001)

Nepenthes_murudensis	**DQ007084** – complete (Meimberg et al., 2006)	**DQ007084** – complete (Meimberg et al., 2006)
Nepenthes_neoguineensis	**AF315896** – complete (Meimberg et al., 2001)	**AF315896** – complete (Meimberg et al., 2001)
Nepenthes_northiana	**AF315901** – complete (Meimberg et al., 2001)	**AF315901** – complete (Meimberg et al., 2001)
Nepenthes_ovata	**AF315873** – complete (Meimberg et al., 2001)	**AF315873** – complete (Meimberg et al., 2001)
Nepenthes_pectinata	**AF315909** – complete (Meimberg et al., 2001)	**AF315909** – complete (Meimberg et al., 2001)
Nepenthes_pervillei	**AF204837** – partial (Meimberg et al., 2000)	No Sequence
Nepenthes_pilosa	**AF315919** – complete (Meimberg et al., 2001)	**AF315919** – complete (Meimberg et al., 2001)
Nepenthes_rafflesiana	**AF315910** – complete (Meimberg et al., 2001)	**AF315910** – complete (Meimberg et al., 2001)
Nepenthes_rajah	**AF315880** – complete (Meimberg et al., 2001)	**AF315880** – complete (Meimberg et al., 2001)
Nepenthes_ramispina	**DQ007083** – complete (Meimberg et al., 2006)	**DQ007083** – complete (Meimberg et al., 2006)
Nepenthes_reinwardtiana	**AF315907** – complete (Meimberg et al., 2001)	**AF315907** – complete (Meimberg et al., 2001)
Nepenthes_rhombicaulis	**AF315874** – complete (Meimberg et al., 2001)	**AF315874** – complete (Meimberg et al., 2001)
Nepenthes_sanguinea	**AF315923** – complete (Meimberg et al., 2001)	**AF315923** – complete (Meimberg et al., 2001)
Nepenthes_sibuyanensis	**DQ840246** – complete (Meimberg and Heubl, 2006)	**DQ840246** – complete (Meimberg and Heubl, 2006)
Nepenthes_singalana	**DQ007082** – complete (Meimberg et al., 2006)	**DQ007082** – complete (Meimberg et al., 2006)
Nepenthes_spathulata	**DQ007081** – complete (Meimberg et al., 2006)	**DQ007081** – partial (Meimberg et al., 2006)
Nepenthes_spectabilis	**AF315868** – complete (Meimberg et al., 2001)	**AF315868** – complete (Meimberg et al., 2001)
Nepenthes_stenophylla	**AF315922** – complete (Meimberg et al., 2001)	**AF315922** – complete (Meimberg et al., 2001)
Nepenthes_sumatrana	**AF315872** – complete	**AF315872** – complete

		(Meimberg et al., 2001)	(Meimberg et al., 2001)
	Nepenthes_talangensis	**AF315924** – complete	**AF315924** – complete
		(Meimberg et al., 2001)	(Meimberg et al., 2001)
	Nepenthes_tentaculata	**AF315932** – complete	**AF315932** – complete
		(Meimberg et al., 2001)	(Meimberg et al., 2001)
	Nepenthes_thorelii	**AF204831** – partial	No Sequence
		(Meimberg et al., 2000)	
	Nepenthes_tobaica	**AF204829** – partial	No Sequence
		(Meimberg et al., 2000)	
	Nepenthes_tomoriana	**AF204830** – partial	No Sequence
		(Meimberg et al., 2000)	
	Nepenthes_treubiana	**AF315893** – complete	**AF315893** – complete
		(Meimberg et al., 2001)	(Meimberg et al., 2001)
	Nepenthes_truncata	**AF315904** – complete	**AF315904** – complete
		(Meimberg et al., 2001)	(Meimberg et al., 2001)
	Nepenthes_veitchii	**AF204828** – partial	No Sequence
		(Meimberg et al., 2000)	
	Nepenthes_ventricosa	**AF204833** – partial	No Sequence
		(Meimberg et al., 2000)	
	Nepenthes_vieillardii	**AB103305** – complete	**AB103305** – partial
		(Kurata et al., 2008)	(Kurata et al., 2008)
	Nepenthes_villosa	**AF315925** – complete	**AF315925** – complete
		(Meimberg et al., 2001)	(Meimberg et al., 2001)
	Nepenthes_xiphioides	**DQ007080** – complete	**DQ007080** – complete
		(Meimberg et al., 2006)	(Meimberg et al., 2006)
Nyctaginaceae	*Acleisanthes_somalensis*	**AY042655** – partial	No Sequence
		(Cuénoud et al., 2002)	
	Allionia_incarnata	**AY042540** – partial	No Sequence
		(Cuénoud et al., 2002)	
	Boerhavia_coccinea	**AY042558** – partial	No Sequence
		(Cuénoud et al., 2002)	
	Commicarpus_raynalii	**AY042571** – partial	No Sequence
		(Cuénoud et al., 2002)	
	Mirabilis_nyctaginea	**AY042624** – partial	No Sequence
		(Cuénoud et al., 2002)	
	Pisonia_umbellifera	**AY042632** – partial	No Sequence
		(Cuénoud et al., 2002)	

Phytolaccaceae	*Agdestis_clematidea*	**AY042538** – partial (Cuénoud et al., 2002)	No Sequence
	Ercilla_volubilis	**AY042583** – partial (Cuénoud et al., 2002)	No Sequence
	Gallesia_integrifolia	**AY042590** – partial (Cuénoud et al., 2002)	No Sequence
	Hilleria_latifolia	**AY042601** – partial (Cuénoud et al., 2002)	No Sequence
	Petiveria_alliacea	**AY042628** – partial (Cuénoud et al., 2002)	No Sequence
	Phytolacca_dioica	**AY042631** – partial (Cuénoud et al., 2002)	No Sequence
	Seguieria_aculeata	**AY042654** – partial (Cuénoud et al., 2002)	No Sequence
Plumbaginaceae	*Afrolimon_purpuratum*	**AY042537** – partial (Cuénoud et al., 2002)	No Sequence
	Armeria_gaditana	**AF204856** – partial (Meimberg et al., 2000)	No Sequence
	Armeria_maritima	**AY042548** – partial (Cuénoud et al., 2002)	No Sequence
	Ceratostigma_minus	**AY042566** – partial (Cuénoud et al., 2002)	No Sequence
	Dyerophytum_africanum	**AY042581** – partial (Cuénoud et al., 2002)	No Sequence
	Goniolimon_tataricum	**AF204855** – partial (Meimberg et al., 2000)	No Sequence
	Limoniastrum_feei	**EU531681** – partial (Kruger et al., unpublished)	No Sequence
	Limoniastrum_monopetalum	**AY042609** – partial (Cuénoud et al., 2002)	No Sequence
	Limonium_cavanillesii	**AY042610** – partial (Cuénoud et al., 2002)	No Sequence
	Limonium_mouretii	**AF204854** – partial (Meimberg et al., 2000)	No Sequence
	Limonium_narbonense	**AF204853** – partial (Meimberg et al., 2000)	No Sequence
	Limonium_oblanceolatum	**AF204852** – partial	No Sequence

		(Meimberg et al., 2000)	
	Limonium_rigualii	**AM889717** – partial (Ford et al., 2009)	No Sequence
	Limonium_rumicifolium	**AF204851** – partial (Meimberg et al., 2000)	No Sequence
	Limonium_thiniense	**AM889718** – partial (Ford et al., 2009)	No Sequence
	Plumbago_europaea	**AY042634** – partial (Cuénoud et al., 2002)	No Sequence
	Plumbago_indica	**AF204857** – partial (Meimberg et al., 2000)	No Sequence
	Psylliostachys_suworowii	**AY042639** – partial (Cuénoud et al., 2002)	No Sequence
Polygonaceae	*Aconogonon_alpinum*	**AF204858** – partial (Meimberg et al., 2000)	No Sequence
	Aconogonon_songoricum	**EU024773** – complete (Yu et al., 2008)	No Sequence
	Afrobrunnichia_erecta	**FJ154489** – partial (Sanchez and Kron, 2009)	No Sequence
	Antigonon_guatimalense	**FJ154491** – partial (Sanchez and Kron, 2009)	No Sequence
	Antigonon_leptopus	**EF437988** – partial (Sanchez and Kron, 2008)	No Sequence
	Bistorta_amplexicaulis	**EF438014** – partial (Sanchez and Kron, 2008)	No Sequence
	Bistorta_vacciniifolia	**AY042556** – partial (Cuénoud et al., 2002)	No Sequence
	Brunnichia_ovata	**FJ154492** – partial (Sanchez and Kron, 2009)	No Sequence
	Chorizanthe_brevicornu	**EF437991** – partial (Sanchez and Kron, 2008)	No Sequence
	Chorizanthe_rigida	**EF437993** – partial (Sanchez and Kron, 2008)	No Sequence
	Coccoloba_peltata	**AY042570** – partial (Cuénoud et al., 2002)	No Sequence
	Coccoloba_swartzii	**EF437995** – partial (Sanchez and Kron, 2008)	No Sequence

Coccoloba_uvifera	**EF437996** – partial (Sanchez and Kron, 2008)	No Sequence
Dedeckera_eurekensis	**EF437997** – partial (Sanchez and Kron, 2008)	No Sequence
Emex_spinosa	**AY042582** – partial (Cuénoud et al., 2002)	No Sequence
Eriogonum_alatum	**EF437998** – partial (Sanchez and Kron, 2008)	No Sequence
Eriogonum_cernuum	**EF437999** – partial (Sanchez and Kron, 2008)	No Sequence
Eriogonum_clavellatum	**EF438000** – partial (Sanchez and Kron, 2008)	No Sequence
Eriogonum_esmeraldense	**EF438002** – partial (Sanchez and Kron, 2008)	No Sequence
Eriogonum_fasciculatum	**EF438004** – partial (Sanchez and Kron, 2008)	No Sequence
Eriogonum_flavum	**AY042584** – partial (Cuénoud et al., 2002)	No Sequence
Eriogonum_hoffmannii	**EF438005** – partial (Sanchez and Kron, 2008)	No Sequence
Eriogonum_inflatum	**EF438006** – partial (Sanchez and Kron, 2008)	No Sequence
Eriogonum_lemmonii	**EF438007** – partial (Sanchez and Kron, 2008)	No Sequence
Fagopyrum_cymosum	**AY042585** – partial (Cuénoud et al., 2002)	No Sequence
Fagopyrum_esculentum	**AB093087** – complete (Yamane et al., 2003)	**AB093087** – complete (Yamane et al., 2003)
Fagopyrum_tataricum	**AB093086** – complete (Yamane et al., 2003)	**AB093086** – complete (Yamane et al., 2003)
Fallopia_convolvulus	**EU749340** – partial (Fazekas et al., 2008)	No Sequence
Fallopia_dentatoalata	**EU024769** – complete (Yu et al., 2008)	No Sequence
Fallopia_dumetorum	**AM503813** – partial (Li et al., 2008)	No Sequence
Fallopia_japonica	**AY042586** – partial	No Sequence

	(Cuénoud et al., 2002)	
Fallopia_multiflora	**EF153691** – partial (Yan et al., 2008)	No Sequence
Fallopia_sachalinensis	**AY042635** – partial (Cuénoud et al., 2002)	No Sequence
Gilmania_luteola	**EF438010** – partial (Sanchez and Kron, 2008)	No Sequence
Johanneshowellia_crateriorum	**EF438011** – partial (Sanchez and Kron, 2008)	No Sequence
Knorringia_sibirica	**EU024771** – complete (Yu et al., 2008)	No Sequence
Muehlenbeckia_axillaris	**AY042617** – partial (Cuénoud et al., 2002)	No Sequence
Muehlenbeckia_tamnifolia	**FJ154499** – partial (Sanchez and Kron, 2009)	No Sequence
Oxyria_digyna	**FJ154500** – partial (Sanchez and Kron, 2009)	No Sequence
Oxyria_sinensis	**AY042625** – partial (Cuénoud et al., 2002)	No Sequence
Persicaria_hydropiper	**EU749342** – partial (Fazekas et al., 2008)	No Sequence
Persicaria_maculosa	**EU749346** – partial (Fazekas et al., 2008)	No Sequence
Persicaria_paniculata	**EF438016** – partial (Sanchez and Kron, 2008)	No Sequence
Persicaria_pensylvanica	**EF438017** – partial (Sanchez and Kron, 2008)	No Sequence
Persicaria_posumbu	**EF438015** – partial (Sanchez and Kron, 2008)	No Sequence
Persicaria_sagittata	**EF438018** – partial (Sanchez and Kron, 2008)	No Sequence
Persicaria_sieboldii	**AB038185** – complete (Kita and Kato, unpublished)	No Sequence
Persicaria_virginiana	**EF438019** – partial (Sanchez and Kron, 2008)	No Sequence
Podopterus_cordifolius	**FJ154494** – partial (Sanchez and Kron, 2009)	No Sequence

Polygonum_aviculare	**EU749338** – partial (Fazekas et al., 2008)	No Sequence
Polygonum_bistorta	**AF204859** – partial (Meimberg et al., 2000)	No Sequence
Polygonum_forrestii	**AY042605** – partial (Cuénoud et al., 2002)	No Sequence
Polygonum_runcinatum	**AY042627** – partial (Cuénoud et al., 2002)	No Sequence
Rheum_coreanum	**AB115687** – partial (Yang et al., 2004)	No Sequence
Rheum_franzenbachii	**AB115689** – partial (Yang et al., 2004)	No Sequence
Rheum_kialense	**AB115692** – partial (Yang et al., 2004)	No Sequence
Rheum_officinale	**AB115686** – partial (Yang et al., 2004)	No Sequence
Rheum_palmatum	**AB115679** – partial (Yang et al., 2004)	No Sequence
Rheum_pinchonii	**AY042644** – partial (Cuénoud et al., 2002)	No Sequence
Rheum_rhaponticum	**AB115688** – partial (Yang et al., 2004)	No Sequence
Rheum_spiciforme	**AB115694** – partial (Yang et al., 2004)	No Sequence
Rheum_tanguticum	**AB115683** – partial (Yang et al., 2004)	No Sequence
Rheum_undulatum	**AB115691** – partial (Yang et al., 2004)	No Sequence
Rumex_acetosella	**EF438022** – partial (Sanchez and Kron, 2008)	No Sequence
Rumex_induratus	**AY042647** – partial (Cuénoud et al., 2002)	No Sequence
Rumex_obtusifolius	**EF438023** – partial (Sanchez and Kron, 2008)	No Sequence
Ruprechtia_chiapensis	**FJ154495** – partial (Sanchez and Kron, 2009)	No Sequence
Ruprechtia_coriacea	**AY042648** – partial	No Sequence

		(Cuénoud et al., 2002)	
	Ruprechtia_fusca	**FJ154496** – partial (Sanchez and Kron, 2009)	No Sequence
	Ruprechtia_laxiflora	**EF438024** – partial (Sanchez and Kron, 2008)	No Sequence
	Triplaris_americana	**AY042668** – partial (Cuénoud et al., 2002)	No Sequence
	Triplaris_poeppigiana	**FJ154497** – partial (Sanchez and Kron, 2009)	No Sequence
	Triplaris_setosa	**FJ154498** – partial (Sanchez and Kron, 2009)	No Sequence
Portulaccaceae	*Anacampseros_karasmontana*	**DQ855859** – complete (Nyffeler, 2007)	No Sequence
	Anacampseros_subnuda	**DQ855861** – complete (Nyffeler, 2007)	No Sequence
	Anacampseros_telephiastrum	**DQ855862** – complete (Nyffeler, 2007)	No Sequence
	Avonia_albissima	**DQ855856** – complete (Nyffeler, 2007)	No Sequence
	Avonia_papyracea	**AY042545** – partial (Cuénoud et al., 2002)	No Sequence
	Avonia_recurvata	**DQ855858** – complete (Nyffeler, 2007)	No Sequence
	Calandrinia_ciliata	**AY764127** – partial (O'Quinn and Hufford, 2005)	**AY764127** – partial (O'Quinn and Hufford, 2005)
	Calandrinia_feltonii	**AY042562** – partial (Cuénoud et al., 2002)	No Sequence
	Ceraria_fruticulosa	**AY875371** – complete (Edwards et al., 2005)	**AY875371** – partial (Edwards et al., 2005)
	Cistanthe_grandiflora	**AY042568** – partial (Cuénoud et al., 2002)	No Sequence
	Claytonia_acutifolia	**AY764097** – partial (O'Quinn and Hufford, 2005)	**AY764097** – partial (O'Quinn and Hufford, 2005)
	Claytonia_arctica	**AY764096** – partial (O'Quinn and Hufford, 2005)	**AY764096** – partial (O'Quinn and Hufford, 2005)
	Claytonia_arenicola	**AY764088** – partial (O'Quinn and Hufford, 2005)	**AY764088** – partial (O'Quinn and Hufford, 2005)

Claytonia_cordifolia	**AY764100** – partial (O'Quinn and Hufford, 2005)	**AY764100** – partial (O'Quinn and Hufford, 2005)
Claytonia_exigua	**AY764089** – partial (O'Quinn and Hufford, 2005)	**AY764089** – partial (O'Quinn and Hufford, 2005)
Claytonia_gypsophiloides	**AY764090** – partial (O'Quinn and Hufford, 2005)	**AY764090** – partial (O'Quinn and Hufford, 2005)
Claytonia_joanneana	**AY764101** – partial (O'Quinn and Hufford, 2005)	**AY764101** – partial (O'Quinn and Hufford, 2005)
Claytonia_lanceolata	**AY764102** – partial (O'Quinn and Hufford, 2005)	**AY764102** – partial (O'Quinn and Hufford, 2005)
Claytonia_nevadensis	**AY764104** – partial (O'Quinn and Hufford, 2005)	**AY764104** – partial (O'Quinn and Hufford, 2005)
Claytonia_ogilviensis	**AY764105** – partial (O'Quinn and Hufford, 2005)	**AY764105** – partial (O'Quinn and Hufford, 2005)
Claytonia_palustris	**AY764106** – partial (O'Quinn and Hufford, 2005)	**AY764106** – partial (O'Quinn and Hufford, 2005)
Claytonia_parviflora	**AY764093** – partial (O'Quinn and Hufford, 2005)	**AY764093** – partial (O'Quinn and Hufford, 2005)
Claytonia_sarmentosa	**AY764107** – partial (O'Quinn and Hufford, 2005)	**AY764107** – partial (O'Quinn and Hufford, 2005)
Claytonia_saxosa	**AY764094** – partial (O'Quinn and Hufford, 2005)	**AY764094** – partial (O'Quinn and Hufford, 2005)
Claytonia_scammaniana	**AY764108** – partial (O'Quinn and Hufford, 2005)	**AY764108** – partial (O'Quinn and Hufford, 2005)
Claytonia_sibirica	**AY764109** – partial (O'Quinn and Hufford, 2005)	**AY764109** – partial (O'Quinn and Hufford, 2005)
Claytonia_tuberosa	**AY764111** – partial (O'Quinn and Hufford, 2005)	**AY764111** – partial (O'Quinn and Hufford, 2005)
Claytonia_umbellata	**AY764112** – partial (O'Quinn and Hufford, 2005)	**AY764112** – partial (O'Quinn and Hufford, 2005)
Claytonia_virginica	**AY764113** – partial (O'Quinn and Hufford, 2005)	**AY764113** – partial (O'Quinn and Hufford, 2005)
Claytonia_washingtoniana	**AY764095** – partial (O'Quinn and Hufford, 2005)	**AY764095** – partial (O'Quinn and Hufford, 2005)
Grahamia_australiana	**DQ855855** – complete (Nyffeler, 2007)	No Sequence
Grahamia_bracteata	**AY015273** – complete	**AY015273** – complete

	(Nyffeler, 2002)	(Nyffeler, 2002)
Grahamia_coahuilensis	**AY875374** – complete (Edwards et al., 2005)	No Sequence
Grahamia_frutescens	**DQ855851** – complete (Nyffeler, 2007)	No Sequence
Grahamia_kurtzii	**DQ855853** – partial (Nyffeler, 2007)	No Sequence
Grahamia_vulcanensis	**AF542597** – partial (Hilu et al., 2003)	No Sequence
Hectorella_caespitosa	**EF551350** – partial (Wagstaff and Hennion, 2007)	**EF551350** – partial (Wagstaff and Hennion, 2007)
Lewisia_cantelovii	**AY042607** – partial (Cuénoud et al., 2002)	No Sequence
Lewisia_columbiana	**AY764126** – partial (O'Quinn and Hufford, 2005)	**AY764126** – partial (O'Quinn and Hufford, 2005)
Lewisia_rediviva	**AY764125** – partial (O'Quinn and Hufford, 2005)	**AY764125** – partial (O'Quinn and Hufford, 2005)
Lyallia_kerguelensis	**EF551349** – partial (Wagstaff and Hennion, 2007)	**EF551349** – partial (Wagstaff and Hennion, 2007)
Montia_bostockii	**AY764114** – partial (O'Quinn and Hufford, 2005)	**AY764114** – partial (O'Quinn and Hufford, 2005)
Montia_chamissoi	**AY764120** – partial (O'Quinn and Hufford, 2005)	**AY764120** – partial (O'Quinn and Hufford, 2005)
Montia_dichotoma	**AY764115** – partial (O'Quinn and Hufford, 2005)	**AY764115** – partial (O'Quinn and Hufford, 2005)
Montia_diffusa	**AY764121** – partial (O'Quinn and Hufford, 2005)	**AY764121** – partial (O'Quinn and Hufford, 2005)
Montia_fontana	**AY764119** – partial (O'Quinn and Hufford, 2005)	**AY764119** – partial (O'Quinn and Hufford, 2005)
Montia_howellii	**AY764117** – partial (O'Quinn and Hufford, 2005)	**AY764117** – partial (O'Quinn and Hufford, 2005)
Montia_linearis	**AY764116** – partial (O'Quinn and Hufford, 2005)	**AY764116** – partial (O'Quinn and Hufford, 2005)
Montia_parvifolia	**AY042616** – partial (Cuénoud et al., 2002)	No Sequence

	Neopaxia_erythrophylla	**AY764123** – partial (O'Quinn and Hufford, 2005)	**AY764123** – partial (O'Quinn and Hufford, 2005)
	Neopaxia_racemosa	**AY764124** – partial (O'Quinn and Hufford, 2005)	**AY764124** – partial (O'Quinn and Hufford, 2005)
	Phemeranthus_multiflorus	**EU834747** – complete (Nyffeler and Eggli, 2010)	No Sequence
	Phemeranthus_teretifolius	**EU834749** – complete (Nyffeler and Eggli, 2010)	No Sequence
	Portulaca_bicolor	**DQ855848** – partial (Nyffeler, 2007)	No Sequence
	Portulaca_fluvialis	**EU834750** – complete (Nyffeler and Eggli, 2010)	No Sequence
	Portulaca_grandiflora	**EU834751** – complete (Nyffeler and Eggli, 2010)	No Sequence
	Portulacaria_afra	**AY042637** – partial (Cuénoud et al., 2002)	No Sequence
	Talinaria_coahuilensis	**AY042661** – partial (Cuénoud et al., 2002)	No Sequence
	Talinum_fruticosum	**DQ855844** – complete (Nyffeler, 2007)	No Sequence
	Talinum_lineare	**EU834752** – complete (Nyffeler and Eggli, 2010)	No Sequence
	Talinum_caffrum	**AY042662** – partial (Cuénoud et al., 2002)	No Sequence
	Talinum_polygaloides	**DQ855845** – complete (Nyffeler, 2007)	No Sequence
	Talinum_portulacifolium	**DQ855847** – complete (Nyffeler, 2007)	No Sequence
	Talinum_punae	**EU834748** – complete (Nyffeler and Eggli, 2010)	No Sequence
Rhabdodendraceae	*Rhabdodendron_macrophyllum*	**AY042642** – partial (Cuénoud et al., 2002)	No Sequence
Tamaricaceae	*Tamarix_gallica*	**AF204861** – partial (Meimberg et al., 2000)	No Sequence

References:

Cameron, K.M., Wurdack, K.J., Jobson, R.W., 2002. Molecular Evidence for the Common Origin of Snap-Traps Among Carnivorous Plants. American Journal of Botany 89, 1503-1509.

Cuénoud, P., Savolainen, V., Chatrou, L.W., Powell, M.P., Grayer, R.J., Chase, M.W., 2002. Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL, atpB,* and *matK* DNA sequences. American Journal of Botany 89, 132 - 144.

Edwards, E.J., Nyffeler, R., Donoghue, M.J., 2005. Basal Cactus Phylogeny: Implications of *Pereskia* (Cactaceae) Paraphyly for the Transition to the Cactus Life Form. American Journal of Botany 92, 1177 - 1188.

Fazekas, A.J., Burgess, K.S., Kesanakurti, P.R., Graham, S.W., Newmaster, S.G., Husband, B.C., Percy, D.M., Hajibabaei, M., Barrett, S.C.H., 2008. Multiple Multilocus DNA Barcodes from the Plastid Genome Discriminate Plant Species Equally Well. PLoS One 3, 2802.

Fior, S., Karis, P.O., 2007. Phylogeny, evolution and systematics of *Moehringia* (Caryophyllaceae) as inferred from molecular and morphological data: a case of homology reassessment. Cladistics 23, 362-372.

Fior, S., Karis, P.O., Casazza, G., Minuto, L., Sala, F., 2006. Molecular phylogeny of the Caryophyllaceae (Caryophyllales) inferred from chloroplast *matK* and nuclear rDNA ITS sequences. American Journal of Botany 93, 399 - 411.

Ford, C.S., Ayres, K.L., Toomey, N., Haider, N., Van Alphen Stahl, J., Kelly, L.J., Wikström, N., Hollingsworth, P.M., Duff, R.J., Hoot, S.B., Cowan, R.S., Chase, M.W., Wilkinson, M.J., 2009. Selection of candidate coding DNA barcoding regions for use on land plants. Botanical Journal of the Linnean Society 159, 1-11.

Hilu, K.W., Borsch, T., Müller, K., Soltis, D.E., Soltis, P.S., Savolainen, V., Chase, M.W., Powell, M.P., Alice, L.A., Evans, R., Sauquet, H., Neinhuis, C., Slotta, T.A.B., Jens, G.R., Campbell, C.S., Chatrou, L.W., 2003. Angiosperm phylogeny based on *matK* sequence information. American Journal of Botany 90, 1758 - 1776.

Houliston, G.J., Olson, M.S., 2006. Nonneutral Evolution of Organelle Genes in *Silene vulgaris*. Genetics 174, 1983-1994.

Kapralov, M.V., Filatov, D.A., 2006. Molecular Adaptation during Adaptive Radiation in the Hawaiian Endemic Genus *Schiedea*. PLoS One 1, 8.

Kejnovsky, E., Kubat, Z., Hobza, R., Lengerova, M., Sato, S., Tabata, S., Fukui, K., Matsunaga, S., Vyskot, B., 2006. Accumulation of chloropolast DNA sequences of the Y chromosome of *Silene latifolia*. Genetica 128, 167-175.

Kress, W.J., Erickson, D.L., 2007. A Two-Locus Global DNA Barcode for Land Plants: The Coding *rbcL* Gene Complements the Non-Coding *trnH-psbA* Spacer Region. PLoS One 2, 508.

Kurata, K., Jaffré, T., Setoguchi, H., 2008. Genetic diversity and geographical structure of the pitcher plant *Nepenthes vieillardii* in New Caledonia: A chloroplast DNA haplotype analysis. American Journal of Botany 95, 1632-1644.

Li, M., Wunder, J., Bissoli, G., Scarponi, E., Gazzani, S., Barbaro, E., Saedler, H., Varotto, C., 2008. Development of COS genes as universally amplifiable markers for phylogenetic reconstructions of closely related plant species. Cladistics 24, 727-745.

Meimberg, H., Dittrich, P., Bringmann, G., Schlauer, J., Heubl, G., 2000. Molecular Phylogeny of Caryophyllidae s.l. Based on *MatK* Sequences with Special Emphasis on Carnivorous Taxa. Plant Biology 2, 218-228.

Meimberg, H., Heubl, G., 2006. Introduction of a Nuclear Marker for Phylogenetic Analysis of Nepenthaceae. Plant Biology 8, 831 - 840.

Meimberg, H., Thalhammer, S., Brachmann, A., Heubl, G., 2006. Comparative analysis of a translocated copy of the *trnK* intron in carnivorous family Nepenthaceae. Molecular Phylogenetics and Evolution 39, 478-490.

Meimberg, H., Wistuba, A., Dittrich, P., Heubl, G., 2001. Molecular Phylogeny of Nepenthaceae Based on Cladistic Analysis of Plastid trnK Intron Sequence Data. Plant Biology 3, 164 - 175.

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

Muir, G., Filatov, D., 2007. A Selective Sweep in the Chloroplast DNA of Dioecious Silene (Section *Elisanthe*). Genetics 177, 1239-1247.

Müller, K., Borsch, T., 2005a. Multiple origins of a unique pollen feature: stellate pore ornamentation in Amaranthaceae. Grana 44, 266-281.

Müller, K.F., Borsch, T., 2005b. Phylogenetics of Amaranthaceae Based on *matK/trnK* Sequence Data - Evidence from Parsimony, Likelihood, and Bayesian Analyses. Annals of the Missouri Botanical Gardens 92, 66 - 102.

Murakeözy, E.P., Aïnouche, A., Meudec, A., Deslandes, E., Poupart, N., 2007. Phylogenetic relationships and genetic diversity of the Salicornieae (Chenopodiaceae) native to the Atlantic coasts of France. Plant Systematics and Evolution 264, 217-237.

Nyffeler, R., 2002. Phylogenetic relationships in the cactus family (Cactaceae) based on evidence from *trnK/matK* and *trnL-trnF* sequences. American Journal of Botany 89, 312 - 326.

Nyffeler, R., 2007. The Closest Relatives of Cacti: Insights from Phylogenetic Analyses of Chloroplast and Mitochondrial Sequences with Special Emphasis on Relationships in the Tribe Anacampseroteae. American Journal of Botany 94, 89 - 101.

Nyffeler, R., Eggli, U., 2010. Disintegrating Portulacaceae: A new familial classification of the suborder Portulacineae (Caryophyllales) based on molecular and morphological data. Taxon 59, 227-240.

O'Quinn, R., Hufford, L., 2005. Molecular Systematics of Montieae (Portulacaceae): Implications for Taxonomy, Biogeography and Ecology. Systematic Botany 30, 314 - 331.

Sage, R.F., Sage, T.L., Pearcy, R.W., Borsch, T., 2007. The Taxonomic Distribution of C4 Photosynthesis in Amaranthaceae Sensu Stricto. American Journal of Botany 94, 1992-2003.

Sanchez, A., Kron, K.A., 2008. Phylogenetics of Polygonaceae with an Emphasis on the Evolution of Eriogonoideae. Systematic Botany 33, 87-96.

Sanchez, A., Kron, K.A., 2009. Phylogenetic relationships of *Afrobrunnichia* Hutch. & Dalziel (Polygonaceae) based on three chloroplast genes and ITS. Taxon 58, 781-792.

Wagstaff, S.J., Hennion, F., 2007. Evolution and biogeography of *Lyallia* and *Hectorella* (Portulacaceae), geographically isolated sisters from the Southern Hemisphere. Antarctic Science 19, 417-426.

Yamane, K., Yasui, Y., Ohnishi, O., 2003. Intraspecific cpDNA variations of diploid and tetraploid perennial buckwheat, *Fagopyrum cymosum* (Polygonaceae). American Journal of Botany 90, 339-346.

Yan, P., Pang, Q., Jiao, X., Zhao, X., Shen, Y., Zhao, S., 2008. Genetic variation and identification of cultivated Fallopia multiflora and its wild relatives by using chloroplast matK and 18S rRNA gene sequences. Planta Med 74, 1504-1509.

Yang, D.-Y., Fushimi, H., Cai, S.-Q., Komatsu, K., 2004. Molecular Analysis of *Rheum* Species used as Rhei Rhizoma Based on the Chloroplast *matK* Gene Sequence and Its Application for Identification. Biological Pharmaceuticals Bulletin 27, 375-383.

Yu, W.-G., Fan, S.-J., Xu, C.-M., Zhu, L.-T., Hou, Y.-T., Lin, F.-Y., Li, F.-Z., 2008. Systematic Position of *Reynoutria* and *Polygonum sibiricum* inferred from sequences of chloroplast *trn*L-F and *mat*K. Journal of Systematics and Evolution 46, 676-681.

Appendix E

Supplemental Figures for Chapter 3

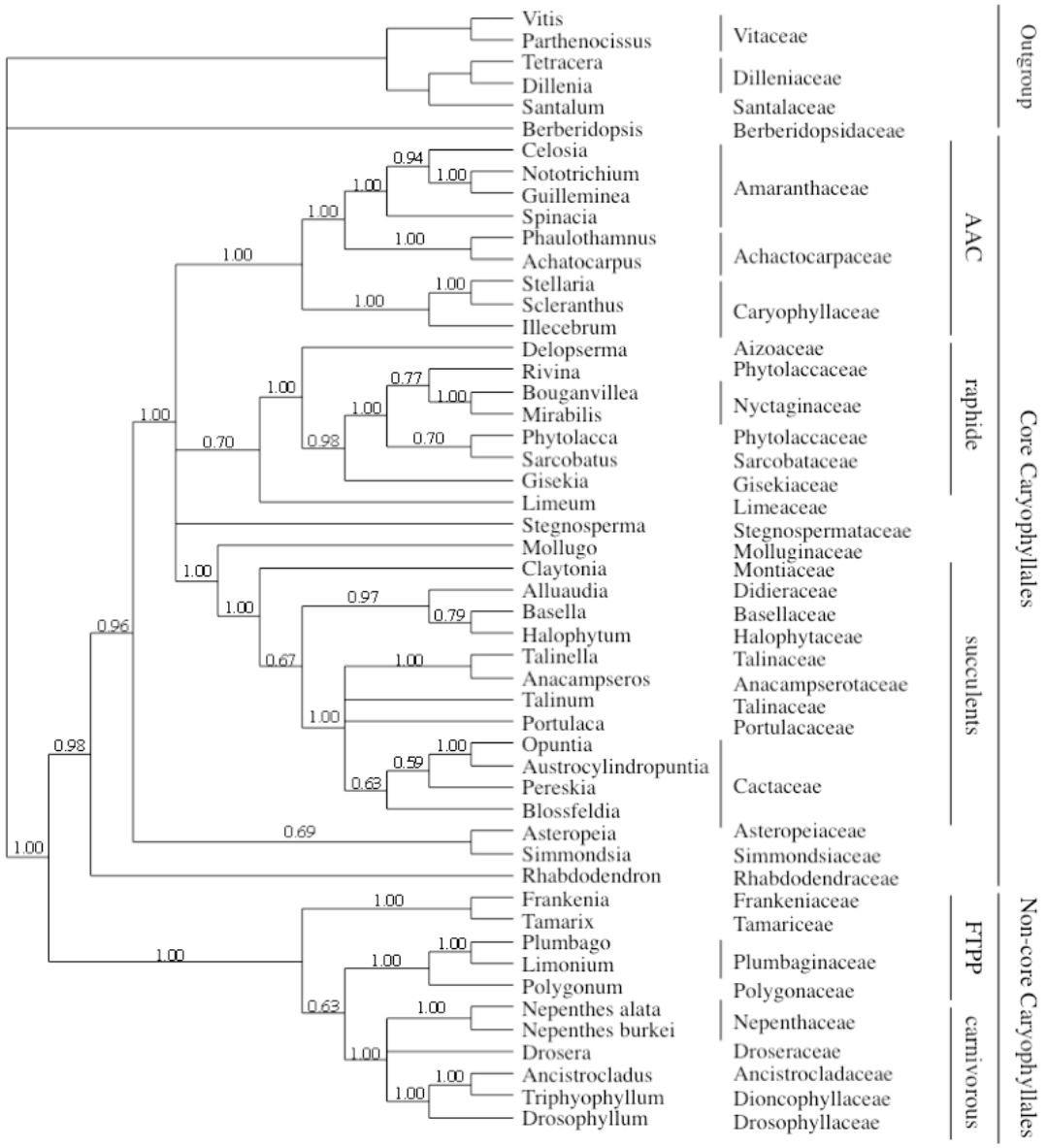

Figure E.1 The 50% majority rule tree for the Caryophyllales derived from Bayesian Inference of substitutions and indels from *trnK* intron. Posterior probability values are noted on the nodes.

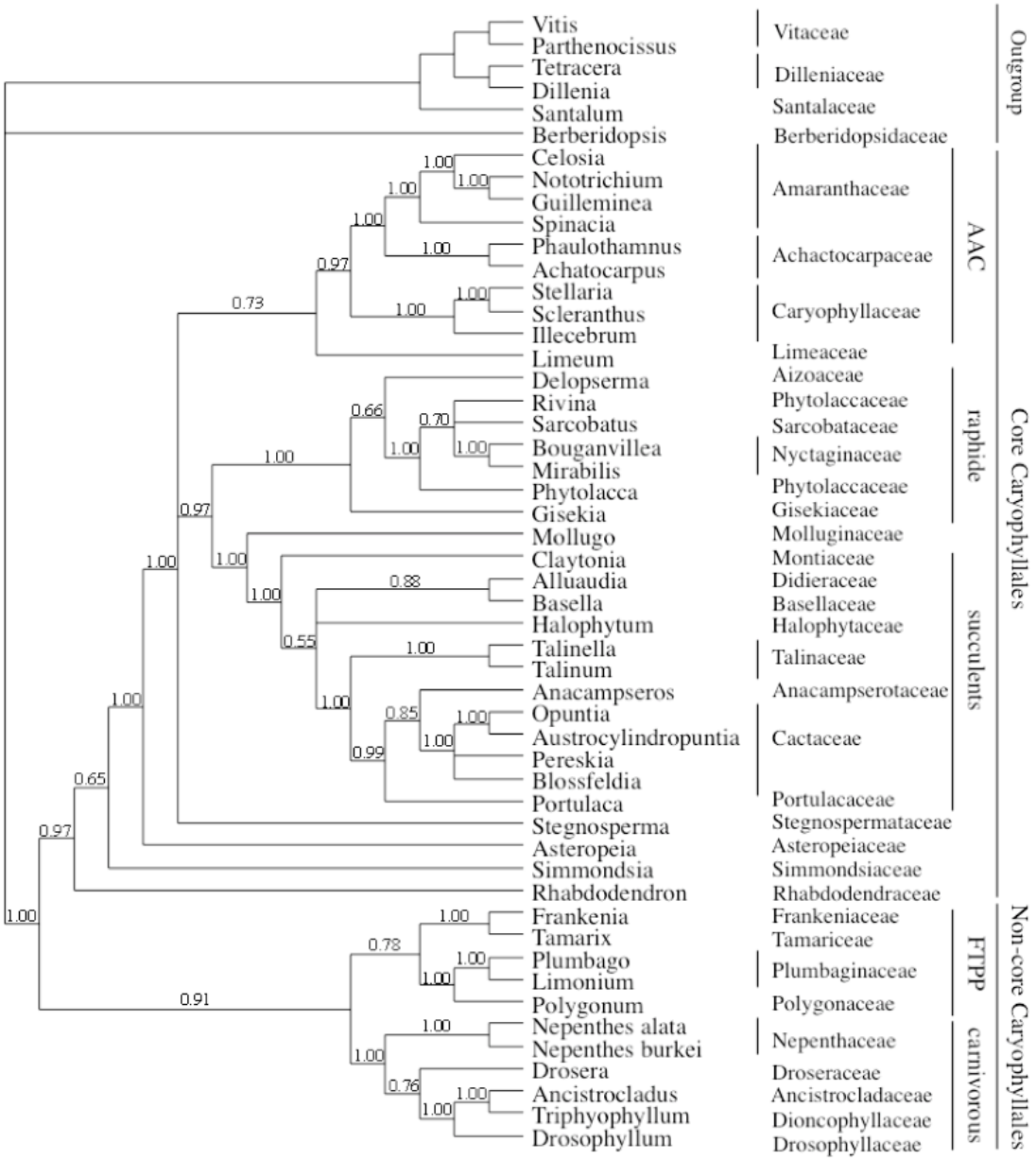

Figure E.2 The 50% majority rule tree for the Caryophyllales derived from Bayesian Inference of substitutions and indels from *matK* ORF. Posterior probability values are noted on the nodes.

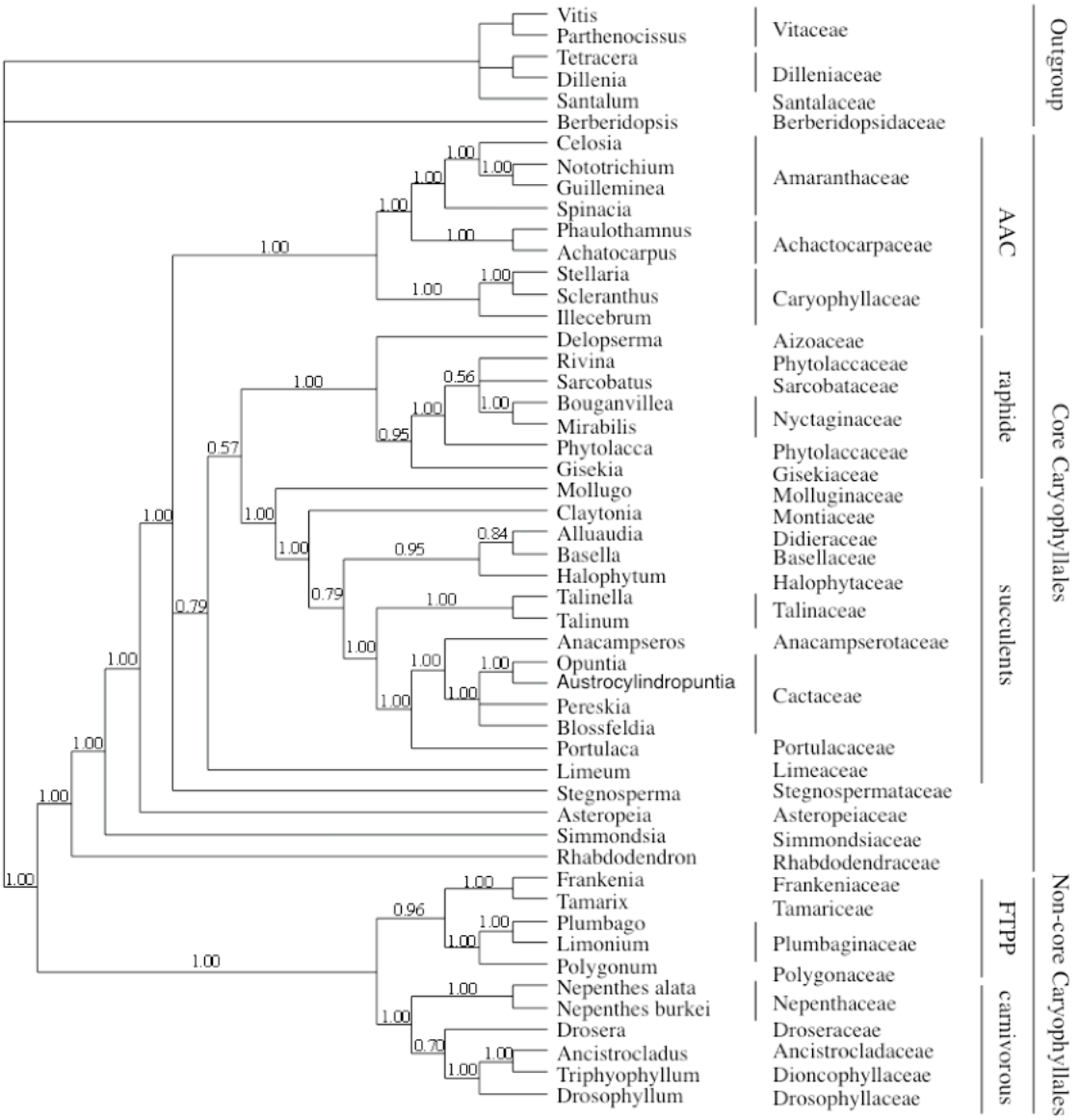

Figure E.3 The 50% majority rule tree for the Caryophyllales derived from Bayesian Inference of substitutions and indels from combined *matK*/*trnK* intron. Posterior probability values are noted on the nodes.

Appendix F

Annotated list of Figures:

Figure 2.1 Summary of the ML tree of the Caryophyllales based on total evidence (plastid IR region plus eleven other genomic regions) from Brockington et al. (2009; Fig. 1). Percent bootstrap values greater than 50% are noted on branches.

Figure 2.2 ML tree based on the *matK*/*trnK* intron dataset for 51 Caryophyllales taxa (0.3% missing data). Percent bootstrap values greater than 50% are noted on branches. Note the misplacement of the Limeaceae compared to the Brockington et al. (2009) tree (Fig. 2.1).

Figure 2.3 Comparison of bootstrap support for the major Caryophyllales nodes in the ML analyses. **a** Support for major nodes obtained using *matK*/*trnK* intron with limited taxon sampling (MT-51), expanded taxon sampling (MT-652), and in the Brockington et al. (2009) study. **b** Support for major nodes using two genomic regions (*matK*/*trnK* intron; MT-136), five genomic regions (5GR-136), and the Brockington et al. (2009) study. **c** Support obtained from the unconstrained five genomic regions (5GR-136), each of the three constrained datasets (3/5GR, 4/5GR, and 5/5GR), and the Brockington et al (2009) study.

Figure 2.4 Summary of the ML tree based on *matK*/*trnK* intron data with expanded taxon sampling (652 taxa with 38% missing data). Percent bootstrap values greater than 50% are noted on branches and branch lengths are noted below the branches in italics.

Figure 2.5 Summary of the ML tree based on *matK*/*trnK* intron data for 136 taxa (MT-136; 21% missing data). Percent bootstrap values greater than 50% are noted on branches.

Figure 2.6 Family level detail of the ML tree based on the dataset of five genomic regions (*rbcL*, *atpB*, *ndhF*, *matK*, and *trnK* intron) for 136 taxa (5GR-136; 46% missing data). Percent bootstrap values greater than 50% are noted on branches. Note the substantial increase in bootstrap support for the placement of the problematic lineages Simmondsiaceae, Limeaceae, and Stegnospermataceae compared with the MT-136 (Fig. 2.5). The tree depicting full details can be found as Figs. A5a, A5b, and A5c.

Figure 2.7 Summaries of ML trees displaying the phylogenetic impact of constraining the original five genomic region dataset (5GR-136) by retaining taxa based on number of genomic regions available. **a** Unconstrained tree based on the 5GR-136 dataset. **b** A tree constrained by allowing taxa that have at least a partial sequence for 3 of the 5 genomic regions (3/5GR; 98 taxa; 38% missing data). **c** A tree constrained by allowing taxa that have at least partial sequences for 4 of the 5 genomic regions represented (4/5GR; 48 taxa; 17% missing data). **d** A tree constrained by allowing taxa that have all 5 genomic regions represented by at least a partial sequence (5/5GR; 15 taxa; 2% missing data). Percent bootstrap values greater than 50% are noted on branches.

Figure 2.8 Proportion of missing data in the *matK/trnK* 652 taxon dataset (MT-652) and the five genomic region dataset (5GR-136). **a** Distribution of missing data among families. **b** Distribution of missing data among major clades.

Figure 3.1 Distribution of substitution rates across 5' and 3' *trnK* introns and the *matK* gene as calculated in HyPhy using the GTR model of evolution. A diagram of *matK* and *trnK* is included to indicate the approximate position of the sites along these genomic regions.

Figure 3.2 Phylogenetic informativeness profiles for *matK* ORF (red), *trnK* intron (green), and combined *matK*/*trnK* (blue). An ultrametric tree (obtained with RAxML and ultrametrisized with PATHd8) with relative divergence times is shown at top and the profiles of net and per-site phylogenetic informativeness are shown below.

Figure 3.3 Strict consensus tree for the Caryophyllales derived from Maximum Parsimony analysis. Bootstrap values are noted on the nodes. A) Phylogeny based on substitutions and indels from *trnK* intron. B) Phylogeny based on substitutions and indels from *matK* ORF.

Figure 3.4 The 50% majority rule trees derived from Maximum Likelihood analyses. Bootstrap values are noted on the nodes with Posterior Probability values noted below the nodes. The * denotes nodes that differ between the ML and BI topologies (See Figs. E1 and E2). A) Phylogeny based on substitutions and indels from *trnK* intron. B) Phylogeny based on substitutions and indels from *matK* ORF.

Figure 3.5 Phylogeny of Caryophyllales based on substitutions and indels from *matK* ORF/*trnK* intron combined. A) Strict consensus tree derived from Maximum Parsimony analysis. Bootstrap values are noted on the nodes. B) The 50% majority rule tree derived from Maximum Likelihood analysis. Bootstrap values are noted on the nodes with Posterior Probability values noted below the nodes. The * denotes nodes that differ between the ML and BI topologies.

Figure 3.6 Comparing *matK* ORF and *trnK* intron separately and combined (substitutions vs. indels) for the degree of bootstrap support and number of nodes resolved from the Maximum Parsimony analyses.

Figure 4.1 Schematic of the *matK* gene indicating where RNA editing events are located in each of the species. The gene is approximately 500 amino acids long, each marked segement above corresponds to 50 amino acids. A * indicates that the editing event takes place in the second codon position while the Δ denotes a chage in the first codon position. A. *Dioon edule* B. *Zamia fisheri* C. *Picea omorkia* D. *Pinus sylvestris* E. *Pinus thunbergii*.

Figure 4.2 Maximum likelihood trees of the Gnetales and other gymnosperm species. BS support values noted above the branches. **A.** Dataset comprised of DNA sequences

only. **B.** Dataset comprised of both DNA and cDNA sequences for those species shown to undergo RNA editing.

Figure 4.3 Diagram of the 38 base pair (bp) indel in some species of *Ephedra*. **A**. The consensus ATG is 62 bp upstream of the proposed alternate ATG (★). **B**. Amino acid translation of the DNA sequence for *Ephedra* when the consensus ATG (★) is used. Note the presense of stop codons (black boxes) 46 and 54 amino acids downstream of the start. **C**. Amino acid translation of the DNA sequence for *Ephedra* when the proposed alternate ATG (★) is used. Note that in species with the indel, this codes for Leu, not Met.

Figure 4.4 Maximum likelihood trees of the Gnetales and other gymnosperm species. In the datasets used to generate these trees, the alternate ATG for *Ephedra* has been used in place of the consensus ATG. BS support values noted above the branches. **A.** Dataset comprised of DNA sequences only. **B.** Dataset comprised of both DNA and cDNA sequences for those species shown to undergo RNA editing.

Figure A.1 MP strict consensus tree based on the *matK/trnK* intron dataset for 51 Caryophyllales taxa (0.3% missing data). Percent bootstrap values greater than 50% are noted on branches.

Figure A.2 Summary of the MP strict consensus tree based on *matK*/*trnK* intron data with expanded taxon sampling (652 taxa with 38% missing data). Percent bootstrap values greater than 50% are noted on branches.

Figure A.3a ML tree based on the 5 genomic regions (*rbcL, atpB, ndhF, matK,* and *trnK* intron) for 136 Caryophyllales taxa. Percent bootstrap values greater than 50% are noted on braches. Expanded details for the "AAC" and "raphide" clades.

Figure A.3b ML tree based on the 5 genomic regions (*rbcL, atpB, ndhF, matK,* and *trnK* intron) for 136 Caryophyllales taxa. Percent bootstrap values greater than 50% are noted on braches. Expanded details for the "succulents" clade.

Figure A.3c ML tree based on the 5 genomic regions (*rbcL, atpB, ndhF, matK,* and *trnK* intron) for 136 Caryophyllales taxa. Percent bootstrap values greater than 50% are noted on braches. Expanded details for the "FTPP" and "carnivorous" clades.

Figure A.4a MP strict consensus tree based on the dataset of five genomic regions (*rbcL, atpB, ndhF, matK,* and *trnK* intron) for 136 taxa (5GR-136; 46% missing data). Percent bootstrap values greater than 50% are noted on branches. The FTPP and carnivorous clades have been collapsed.

Figure A.4b MP strict consensus tree based on the dataset of five genomic regions (*rbcL, atpB, ndhF, matK,* and *trnK* intron) for 136 taxa (5GR-136; 46% missing data). Percent bootstrap values greater than 50% are noted on branches. FTPP and carnivorous clades are expanded.

Figure A.5 ML tree based on the *matK/trnK* intron dataset for 51 Caryophyllales taxa. Branch lengths are noted on the branches.

Figure E.1 The 50% majority rule tree for the Caryophyllales derived from Bayesian Inference of substitutions and indels from *trnK* intron. Posterior probability values are noted on the nodes.

Figure E.2 The 50% majority rule tree for the Caryophyllales derived from Bayesian Inference of substitutions and indels from *matK* ORF. Posterior probability values are noted on the nodes.

Figure E.3 The 50% majority rule tree for the Caryophyllales derived from Bayesian Inference of substitutions and indels from combined *matK*/*trnK* intron. Posterior probability values are noted on the nodes.