PHYLOGENOMIC PLACEMENT OF ANCIENT POLYPLOIDY EVENTS WITHIN THE POALES AND AGAVOIDEAE (ASPARAGALES)

by

MICHAEL RAMON MCKAIN

(Under the Direction of James H. Leebens-Mack)

ABSTRACT

Polyploidy has been an important component to the evolution of angiosperms. Recent studies have shown that an ancient polyploid (paleopolyploid) event can be traced to the lineage leading to the diversification of all angiosperms, and it has long been known that recurring polyploid events can be found throughout the angiosperm tree of life. With the advent of high-throughput sequencing, the prominent place of paleopolyploid events in the evolutionary history of angiosperms has become increasingly clear. Polyploidy is thought to spur both diversification and trait innovation through the duplication and reworking of gene networks. Understanding the evolutionary impact of paleopolyploidy within the angiosperms requires knowing when these events occurred during angiosperm evolution. This study utilizes a high-throughput phylogenomic approach to identify the timing of paleopolyploid events by comparing the origin of paralogous genes within a gene family to a known species tree. Transcriptome data derived from taxa in lineages with previously little to no genomic data, were utilized to assess the timing of duplication events within hundreds of gene families. Previously described paleopolyploid events in the history of grasses, identified through analyses of syntenic blocks within Poaceae genomes, were placed on the Poales phylogeny and the implications of these events were

considered. Additionally, a previously unverified paleopolyploidy event was found to have occurred in a common ancestor of all members of the Asparagales and commelinids (including Poales, Zingiberales, Commelinales, Arecales and Dasypogonales). The phylogeny of the Asparagaceae subfamily Agavoideae was resolved using whole chloroplast genomes, and two previously unknown paleopolyploid events were described within the context of that phylogeny. The potential effects of these paleopolyploid events on the evolution of the "*Yucca-Agave*" bimodal karyotype were discussed. This study demonstrates the utility of large transcriptomic sequencing projects and phylogenomic analyses of gene families to identify novel polyploid events and place them within an evolutionary context.

INDEX WORDS: Agavoideae, Asparagaceae, paleopolyploidy, phylogenomics, Poales

PHYLOGENOMIC PLACEMENT OF ANCIENT POLYPLOIDY EVENTS WITHIN THE POALES AND AGAVOIDEAE (ASPARAGALES)

by

MICHAEL RAMON MCKAIN

BA, Wabash College 2007

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2012

PHYLOGENOMIC PLACEMENT OF ANCIENT POLYPLOIDY EVENTS WITHIN THE POALES AND AGAVOIDEAE (ASPARAGALES)

by

MICHAEL RAMON MCKAIN

Major Professor: James H. Leebens-Mack

Committee: Luis E. Eguiarte
J. L. Hamrick
Russell L. Malmberg
Kathleen C. Parker
Wendy B. Zomlefer

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2012

DEDICATION

This work is dedicated to my daughter, Kendalynn Marie Lyons-McKain, and my wife, Sarah Lyons. Their sacrifice while I have completed this journey has been, at times, great, and I will endeavor to show them that it was worth it. Kendalynn has never known me to be anything other than a graduate student, and I look forward to the development of our relationship as we both grow older and move forward in our lives. Sarah has been an anchor for me during this time, keeping me grounded and giving me reason to continue. Thank you both for everything. You are my greatest love and continue to be the brightest aspect of my life.

ACKNOWLEDGEMENTS

I thank James H. Leebens-Mack, my advisor, for his support in both my research and career throughout my time in his lab. Jim has been a great advocate for me and has given me more opportunity than I could have imagined in graduate school. I also thank my committee members Luis Eguiarte, Jim Hamrick, Russell Malmberg, Kathy Parker, and Wendy Zomlefer for their support, perspective, and intellectually stimulating conversations. I give special thanks to J. Chris Pires (Mizzou) and Chris Smith (Willamette University), who have served as mentors to me in a great capacity providing guidance for my career. I thank Joel McNeal, Raj Ayyampalayam, Mark Chapman, and Aaron Richardson for their mentorship, friendship, and guidance as they taught me throughout my degree. I thank Denise Domizi and Paul Quick, who helped me to further my love of teaching through the Future Faculty Program. I thank all of the friends I made through the Future Faculty Program for the unforgettable experience and long-lasting friendships they brought to me. I thank Michael Boyd, Jeff Dadisman, Melanie Smith, Kevin Tarner, and the rest of the greenhouse staff for their custodianship of the plants involved in my research. I thank my lab mates for their support and friendship throughout the years. I thank Lisa Kanizay, Stephanie Pearl, and Jeremy Rentsch for the great friendship and support they have given me over the years. Together we have achieved some great things, and I expect to see even more great things from them in the future. Lastly, I would like to acknowledge the Department of Plant Biology as a whole. I have loved my time at UGA because of the people and atmosphere that this department has provided. The department has been vital to my success, and I will forever owe it a debt of gratitude.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

Polyploidy (or whole genome duplication) has occurred throughout the eukaryotic tree of life (Otto, 2007). Angiosperms, however, are especially prone to polyploidy with estimates of modern polyploids (neopolyploids) ranging from 30% (Stebbins, 1950; Wood et al., 2009) to 80% (Goldblatt, 1980; Lewis, 1980). Ancient polyploidy (paleopolyploidy) is ubiquitous across angiosperms lineages and may have contributed to the success of the lineage (Soltis et al., 2009). Recent work by Jiao et al. (2011) has shown that all angiosperms have at least two shared duplications in their history, one prior to the divergence of gymnosperms and angiosperms (i.e. seed plants) and a second on the branch leading to the angiosperm crown group. Polyploidy has occurred in multiple lineages within the angiosperms, often leading to large species-rich clades (Soltis et al., 2009), including the eudicots (Jiao et al., 2012), Asteraceae (Barker et al., 2008), Brassicaceae (Vision et al., 2000; Blanc and Wolfe, 2004; Schranz and Mitchell-olds, 2006), Cleomaceae (Schranz and Mitchell-olds, 2006; Barker et al., 2009), Fabaceae (Blanc and Wolfe, 2004; Pfeil et al., 2005; Cannon et al., 2006, 2010), Solanaceae (Schlueter et al., 2004), and Poaceae (Blanc and Wolfe, 2004; Paterson et al., 2004; Schlueter et al., 2004; Tang et al., 2010). Additionally, the origin of key suites of novel characters correlates to the timing of polyploid events accompanied by expansion of these lineages suggesting that polyploidy is driving trait innovation and speciation (Schranz et al., 2012). Polyploidy has an important role in the evolution of angiosperms, but understanding that role requires placing these events within the context of the angiosperm phylogeny.

Detecting Paleopolyploid Events

Historically, polyploidy was inferred from analyses of chromosome counts (Stebbins, 1940). Polyploidy is defined as an organism having more than one set of chromosomes from either or both parents. Polyploidy levels in angiosperms can range from 3× (e.g. *Thalictrum* (Mooney and Johnson, 1965)) to at least 20× (e.g. *Atriplex*; (Sanderson and Stutz, 1994) and possibly more. In instances such as *Atriplex canescens*, polyploid races are found in a single species with ploidy levels ranging from 2× up to, in this case, 20× with almost every even numbered ploidy level between these extremes (Sanderson and Stutz, 1994). When chromosome numbers of closely related species or populations within a single species vary by a factor of at least 1.5 or are additive in a series by 2 (such as 2×, 4×, and 6×), a polyploid event is typically hypothesized. Cytogenetic methods are useful for identifying recent polyploidy events, but they can be problematic for detecting very ancient events.

With a polyploid event, chromosome number changes (usually doubles) and there is an increase in genome size. These can be useful traits for inferring recent polyploid events, but the genome size and chromosome number of polyploids may change over time as the genome undergoes diploidization, a series of cytological and genetic events that returns a polyploid genome to a more diploid-like state. During diploidization, duplicated gene networks undergo rewiring to produce novel networks (De Smet and Van de Peer, 2012), duplicated genes are lost in a process called fractionation (Freeling, 2009), and chromosomes are rearranged or purged from the genome (Pires et al., 2004; Chen and Ni, 2006). These restructuring events can occur within a dozen generations (Song et al., 1995; Pires et al., 2004) or over millions of years (Devos et al., 2002; Leitch and Bennett, 2004). Ultimately, traditional evidence for polyploid events becomes lost as chromosome numbers and genome sizes are reduced.

A prime example of the inability to detect paleopolyploidy comes from investigations of the model system *Arabidopsis*. *Arabidopsis thaliana* is functionally diploid (Chen et al., 2004), possesses a small number of chromosomes ($n = 5$), and a small genome size (C-value = 0.16 pg per gametic genome) (Bennett et al., 2003) with no indication of polyploidy in its history from these data. Sequencing the *Arabidopsis* genome provided evidence of a paleopolyploid event identified by a large number of segmental duplications (The Arabidopsis Genome Initiative, 2000). Further study of the *Arabidopsis* genome and other angiosperm sequence data, demonstrated three paleopolyploid events in the history of *Arabidopsis* (Bowers et al., 2003). Recent comparisons to the *Vitis* (Jaillon et al., 2007) and *Carica* (Ming et al., 2008) genomes has shown that two of these events occurred after the *Carica* and Brassicaceae lineages split, while the third can be traced to a hexaploid event shared with *Vitis* (Tang et al., 2007). The duplication history of *Arabidopsis* has been further elucidated with placement of the hexaploid event prior to the divergence of rosids and asterids (Jiao et al., 2012) in addition to the identification of the angiosperm and seed plant duplication events (Jiao et al., 2011). In all, this suggests that *Arabidopsis* is a tetracontrakaioctaploid ($48\times = 2\times2\times3\times2\times2$) relative to the common ancestor of gymnosperms and angiosperms!

Detection of paleopolyploidy using genomic sequence data has recently become more common with the increasing number of sequenced and assembled plant genomes. Further, the advent of high-throughput sequencing has allowed the generation of large amounts of genomic and transcriptomic data from a wide variety of taxa relatively cheaply (Steele et al., 2012). The extent of these projects ranges from the sequencing of a nuclear genome of a single species (e.g. *Musa acuminata*; (D'Hont et al., 2012)) to the sequencing of the transcriptomes of a few species (McKain et al., 2012). Analytical methods have been developed to identify paleopolyploid

events through comparative analyses. Transcriptome data allows for analyses of transcribed genes and limits detection of paleopolyploidy to synonymous substitution frequency distributions (or K_s plots) and phylogenetic analysis of gene families. The sequencing of whole genomes allows for the additional analyses of synteny (genes maintained in blocks) and collinearity (gene order). Multiple methods are often used to confidently identify duplication events (e.g. (Bowers et al., 2003; Paterson et al., 2004; Jiao et al., 2011, 2012)).

Synteny and collinearity

The definition of synteny within genomes has varied; it was originally used to define genes on the same chromosome regardless of genetic linkage (Renwick, 1972). Now, synteny typically refers to chromosomal segments - either between species or within a single genome – with multiple loci sharing the same ancestral history (Tang et al., 2007). The concept of collinearity refers to the shared order of genes within these syntenic blocks (Tang et al., 2007). Syntenic blocks can be the product of speciation when the blocks are drawn from different species or through some form of segmental, chromosomal, or genome duplication when the blocks are drawn from the same species or different species sharing a polyploidy common ancestor. Prior to speciation, an ancestral genome will possess a certain number of chromosomes with a particular gene order. Following speciation, each sister species genome may independently undergo chromosome fusion and fission events (Jones, 1998) and rearrangements (Rieseberg, 2001) resulting in variation of gene order, gene location on chromosomes, and chromosome number. Identification of synteny between these two genomes assumes that conserved proximity of genes to one another between species is a consequence or shared ancestry (vs. convergence). Syntenic blocks are also formed through segmental, chromosomal,

or genome duplication. When genome duplication occurs, the newly doubled genome may possess complete synteny between its two subgenomes (as in autopolyploidy) or lower amounts of synteny inherited from the differences of the parental genomes (allopolyploidy). Subsequent fusion and fission events and rearrangements will lower synteny between subgenomes leading to smaller, dispersed regions of the genome possessing ancestral gene order and proximity (e.g. (Schlueter et al., 2008)). The identification of syntenic blocks within a genome is suggestive of past segmental duplications, and when these syntenic regions occur in large number throughout the genome, they are likely the result of whole genome duplication (Bowers et al., 2003).

Synteny analyses are the most definitive method for identifying duplicated portions of genomes as they not only rely on sequence similarity to find homologous genes but also examine the proximity of homologous genes to one another. The likelihood that sequence similarity and gene proximity would randomly converge multiple times across a genome is probably very low. The major caveat to using synteny analyses is the requirement of a sequenced genome (or at least a large portion of it) assembled into chromosomes or linkage groups. Many angiosperm lineages do not include species for which a genome has been sequenced and assembled [e.g., Asparagales with the closest sequenced genomes being *Musa* (D'Hont et al., 2012) or *Phoenix* (Al-Dous et al., 2011)), so synteny analyses are limited to a few groups.

K_s *analyses of paralogs*

A polyploid event creates two copies of each gene in the genome (disregarding allelic copies in the diploid progenitors). Following duplication, these genes are thought to have three fates: i) loss, ii) neofunctionalization, or iii) subfunctionalization (Lynch and Conery, 2000; Lynch and Force, 2000). Gene loss following polyploidy often occurs in a biased fashion, with

retention of genes exhibiting dosage sensitivity being much higher (Thomas et al., 2006; Sankoff et al., 2012). Conversely, consistently single-copy homologs found in the *Arabidopsis*, *Oryza*, *Populus*, and *Vitis* genomes suggest that some genes may be maladaptive when duplicated, even when gene duplications occur as a consequence of polyploidization (Duarte et al., 2010). When compared across land plants, this trend appears to be consistent across taxa (Duarte et al., 2010) although duplicates for some of these gene families may exist in some lineages. Neofunctionalization occurs when one of the gene duplicates (paralogs) retains the ancestral function of the gene and the other is able to undergo relaxed constraint and potentially develop a novel function (Lynch and Conery, 2000). Subfunctionalization occurs when either paralogs undergo segmental silencing and each copy produces a different portion of the gene product (Ohno, 1970) or the paralogs become spatiotemporally separated in function (De Smet and Van de Peer, 2012). Regardless of their fate, retention of large numbers of duplicate genes is a hallmark for identifying polyploid events.

Paleopolyploid events are often identified using frequency distribution plots of synonymous substitutions in protein-coding genes (K_s plot analysis) (Blanc and Wolfe, 2004; Barker et al., 2008, 2009; Shi et al., 2010; Jiao et al., 2011, 2012; Vanneste et al., 2012). This method was first used by Lynch and Conery (2003) and has gained popularity, especially with the advent of high-throughput transcriptome sequencing (Wang et al., 2009). K_s plot analyses assume random gene and genome segment duplications throughout the history of the genome. When these occur, they create paralogous genes. These duplicates, if retained, are subjected to the same fates as genes duplicated in a polyploid event. Over time, the nucleotide sequences of the paralogs diverge. If only synonymous substitutions (i.e. nucleotide substitutions that do not change the encoded amino acid) are considered when examining divergence, then this is

representative of the neutral mutation rate and is thought to be constant or similar across the genome (Lynch and Conery, 2003; Blanc and Wolfe, 2004). Since randomly duplicated genes are lost relatively quickly before diverging to any great degree (Lynch and Conery, 2003), plots of K_s frequencies exhibit exponential decay with high numbers of duplicates with low K_s values and increasingly fewer duplicates with higher K_s values. In instances where a polyploid event occurs, there is a large increase in the number of duplicates retained. This creates a secondary peak on the K_s plot with a mode centered around the mean K_s value of the duplicates created in the polyploid event (Lynch and Conery, 2003; Blanc and Wolfe, 2004). K_s plots are most often used to identify an event, but they are also occasionally used to estimate relative timing of the polyploid event to speciation events (Blanc and Wolfe, 2004; Barker et al., 2008, 2009; Shi et al., 2010).

K_s frequency plot analyses are the least definitive method of describing paleopolyploidy. In general, they are useful for identifying whole genome duplication events as long as substitution rates are not too high (< 2.0) and substitution saturation has not been met (Blanc and Wolfe, 2004). As time progresses after a duplication event, the secondary peak marking the event in a K_s plot begins to gain a wider distribution of values while the height (frequency) of the peak declines because duplicate pairs are spread over the larger K_s distribution (Cui et al., 2006). When substitution rates are high or enough time has passed to diminish the secondary peak, the duplication event cannot be detected in a K_s frequency plot. Therefore, they should only be used to identify polyploid events but not to definitively prove the absence of such an event. K_s frequency plots have often been used to place paleopolyploidy events relative to speciation events (Blanc and Wolfe, 2004; Barker et al., 2008, 2009; Shi et al., 2010) but this is under an assumption that substitution rates are relatively equal (Smith and Donoghue, 2008). When

comparing the K_s plots of different species, rate variation must be taken into account (Cui et al., 2006; Vanneste et al., 2012). Additionally, the divergence of paralogs estimated in K_s plots relates to different events depending upon whether the polyploid event was autopolyploid or allopolyploid (Doyle and Egan, 2010). When autopolyploid, the secondary peak can be related to the polyploid event, since there is no inherent divergence between the two subgenomes. However, if the event was allopolyploid, then the secondary peak relates to the original divergence o the two parental subgenomes (i.e. speciation), that may have occurred long before the subgenomes were united (Doyle and Egan, 2010).

Phylogenetic analyses of gene families

Gene family histories are a reliable method for identifying duplication events, whether they are specific to that gene family or a whole genome duplication (Lott et al., 2009). Within gene family trees, duplications are indicated when a node gives rise to two clades with overlapping taxa. This phenomenon has been characterized in various gene families and can be related to polyploid events (e.g. Litt and Irish, 2003; Preston and Kellogg, 2006). A major challenge when using large genomic and transcriptomic data sets is the circumscription of gene families. The most common approach is to use a clustering method based on sequence similarity, such as OrthoMCL (Li et al., 2003), one of a few programs available allowing circumscription of gene families (often referred to as orthogroups) from multiple taxa (Chen et al., 2007). The OrthoMCL method has been proficient at splitting orthogroups along duplication lines, creating smaller and more precisely defined gene families (Chen et al., 2007). Splitting gene families at duplication events is not useful when identifying polyploidy in gene family trees. The number of gene families with evidence of polyploidy will be reduced, and those

present are likely instances with little-divergent paralogous clades. Despite this limitation, gene family phylogenies using OrthoMCL circumscription have been used to identify and phylogenetically place major angiosperm paleopolyploidy events (Jiao et al., 2011, 2012).

Gene families circumscribed using transcriptomic data sets often have missing data. The RNA-Seq method, typically used to generate transcriptomes, is a shotgun approach (Wang et al., 2009) and can result in taxa not being sampled for a gene family or genes sampled with only a small proportion of their true length. Estimated phylogenies of these gene families can have poorly resolved topologies making it difficult to reconcile gene family trees to species trees and therefore, place polyploid events on the species tree. When considering gene family trees to estimate the timing of polyploidy events, support values should be used to determine certainty of topology (Jiao et al., 2011, 2012). Strict filtering of gene family trees is often necessary to remove trees with evidence of long branch attraction artifacts that cannot be used to ascertain the timing of polyploid events. After filtering, there is usually a small fraction of trees suitable for further analyses (Jiao et al., 2011, 2012; McKain et al., 2012).

As outline above, identification of paleopolyploid events using genomic and transcriptomic data can be done in various ways. An effective method of identifying and phylogenetically placing paleopolyploidy events is a combination of these methods so that multiple lines of evidence can support or reject a given hypothesis (e.g. (Bowers et al., 2003; Paterson et al., 2004; Tang et al., 2010; Jiao et al., 2011, 2012; McKain et al., 2012).

Purpose of Study

The phylogenetic placement of paleopolyploidy events is essential for understanding the role polyploidy of in the evolution of angiosperms. In this study, a phylogenomic approach is

taken that first identifies paralogs in genomic and transcriptomic data sets using either synteny analysis or K_s frequency plots, circumscribes gene families and uses the paralog pairs to collapse the gene families so they include polyploid events, and then utilizes the paralog pairs to identify polyploid events on reconstructed gene family phylogenies. This approach is used to identify the timing of the *rho* and *sigma* events of Poaceae (Tang et al., 2010), clarifying the role of whole genome duplication in Poales. The approach is then applied to subfamily Agavoideae to determine whether paleopolyploidy coincided with the origin of the bimodal karyotype. These applications provide a proof of concept for this method and represent potential instances of paleopolyploidy influencing angiosperm evolution.

References

AL-DOUS, E.K. ET AL. 2011. De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nature Biotechnology* 29: 521–527.

BARKER, M.S. ET AL. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution* 25: 2445–2455.

BARKER, M.S., H. VOGEL, AND M.E. SCHRANZ. 2009. Paleopolyploidy in the Brassicales: analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. *Genome Biology and Evolution* 1: 391–399.

BENNETT, M.D., I.J. LEITCH, H.J. PRICE, AND J.S. JOHNSTON. 2003. Comparisons with *Caenorhabditis* (100 Mb) and *Drosophila* (175 Mb) using flow cytometry show genome size in *Arabidopsis* to be 157 Mb and thus 25 % larger than the *Arabidopsis* genome initiative estimate of 125 Mb. *Annals of Botany* 91: 547–557.

BLANC, G., AND K.H. WOLFE. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant Cell* 16: 1667–1678.

BOWERS, J.E., B.A. CHAPMAN, J. RONG, AND A.H. PATERSON. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422: 433–438.

CANNON, S.B. ET AL. 2006. Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. *Proceedings of the National Academy of Sciences* 103: 14959–14964.

CANNON, S.B. ET AL. 2010. Polyploidy did not predate the evolution of nodulation in all legumes. *PloS One* 5: e11630.

CHEN, F., A.J. MACKEY, J.K. VERMUNT, AND D.S. ROOS. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PloS One* 2: e383.

CHEN, Z., J. WANG, AND L. TIAN. 2004. The development of an *Arabidopsis* model system for genome-wide analysis of polyploidy effects. *Biological Journal of the Linnean Society* 82: 689–700.

CHEN, Z.J., AND Z. NI. 2006. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *BioEssays* 28: 240–252.

CUI, L. ET AL. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Research* 16: 738–749.

DEVOS, K.M., J.K.M. BROWN, AND J.L. BENNETZEN. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Research* 12: 1075–1079.

DOYLE, J.J., AND A.N. EGAN. 2010. Dating the origins of polyploidy events. *The New Phytologist* 186: 73–85.

DUARTE, J.M. ET AL. 2010. Identification of shared single copy nuclear genes in *Arabidopsis, Populus, Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology* 10: 61.

D'HONT, A. ET AL. 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* 488: 213–217

FREELING, M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual Review of Plant Biology* 60: 433–453.

GOLDBLATT, 1980. Polyploidy in angiosperms: Monocotyledons. In Lewis W. H. [ed.], Polyploidy: Biological relevance. Plenum Press, New York, New York.

THE ARABIDOPSIS GENOME INITIATIVE. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796–815.

JAILLON, O. ET AL. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449: 463–467.

JIAO, Y. ET AL. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biology* 13: R3.

JIAO, Y. ET AL. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.

JONES, K. 1998. Robertsonian fusion and centric fission in karyotype evolution of higher plants. *The Botanical Review* 64: 273–289.

LEITCH, I.J., AND M.D. BENNETT. 2004. Genome downsizing in polyploid plants. *Biological Journal of the Linnean Society* 82: 651–663.

LEWIS, W.H. 1980. Polyploidy in angiosperms: Dicotyledons. In W. H. Lewis [ed.], Polyploidy: Biological relevance. Plenum Press, New York, New York.

LI, L., C.J. STOECKERT, AND D.S. ROOS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research* 13: 2178–2189.

LITT, A., AND V.F. IRISH. 2003. Duplication and diversification in the *APETALA1/FRUITFULL* floral homeotic gene lineage: implications for the evolution of floral development. *Genetics* 165: 821–833.

LOTT, M., A. SPILLNER, K.T. HUBER, A. PETRI, B. OXELMAN, AND V. MOULTON. 2009. Inferring polyploid phylogenies from multiply-labeled gene trees. *BMC Evolutionary Biology* 9: 216.

LYNCH, M., AND J.S. CONERY. 2003. The evolutionary demography of duplicate genes. *Journal of Structural and Functional Genomics* 3: 35–44.

LYNCH, M., AND J.S. CONERY. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151–1155.

LYNCH, M., AND A. FORCE. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154: 459–473.

MCKAIN, M.R. ET AL. 2012. Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in Agavoideae (Asparagaceae). *American Journal of Botany* 99: 397–406.

MING, R. ET AL. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452: 991–996.

MOONEY, H., AND A. JOHNSON. 1965. Comparative physiological ecology of an Arctic and alpine population of *Thalictrum Alpinum*. *Ecology* 46: 721–727

OHNO, S. 1970. *Evolution by Gene Duplication*. Springer-Verlag.

OTTO, S. 2007. The evolutionary consequences of polyploidy. *Cell* 131: 452–462.

PATERSON, A.H., J.E. BOWERS, AND B.A. CHAPMAN. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences* 101: 9903–9908.

PFEIL, B.E., J.A. SCHLUETER, R.C. SHOEMAKER, AND J.J. DOYLE. 2005. Placing paleopolyploidy in relation to taxon divergence: A phylogenetic analysis in legumes using 39 gene families. *Systematic Biology* 54: 441–454.

PIRES, J.C. ET AL. 2004. Flowering time divergence and genomic rearrangements in resynthesized *Brassica* polyploids (Brassicaceae). *Biological Journal of the Linnean Society* 82: 675–688.

PRESTON, J.C., AND E. A KELLOGG. 2006. Reconstructing the evolutionary history of paralogous *APETALA1/FRUITFULL*-like genes in grasses (Poaceae). *Genetics* 174: 421–437.

RENWICK, J.H. 1972. The mapping of human chromosomes. *Annual Review of Genetics* 5: 81–120.

RIESEBERG, L.H. 2001. Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution* 16: 351–358.

SANDERSON, S.C., AND H.C. STUTZ. 1994. High chromosome numbers in Mojavean and Sonoran desert *Atriplex canescens* (Chenopodiaceae). *American Journal of Botany* 81: 1045–1053.

SANKOFF, D., C. ZHENG, AND B. WANG. 2012. A model for biased fractionation after whole genome duplication. *BMC Genomics* 13(Suppl. 1) : S8.

SCHLUETER, J. A, B.E. SCHEFFLER, S. JACKSON, AND R.C. SHOEMAKER. 2008. Fractionation of synteny in a genomic region containing tandemly duplicated genes across *Glycine max*, *Medicago truncatula*, and *Arabidopsis thaliana*. *The Journal of Heredity* 99: 390–395.

SCHLUETER, J.A. ET AL. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* 47: 868–876.

SCHRANZ, M.E., AND T. MITCHELL-OLDS. 2006. Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. *Plant Cell* 18: 1152–1165.

SCHRANZ, M.E., S. MOHAMMADIN, AND EDGER. 2012. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Current Opinion in Plant Biology* 15: 147–153.

SHI, T., H. HUANG, AND M.S. BARKER. 2010. Ancient genome duplications during the evolution of kiwifruit (*Actinidia*) and related Ericales. *Annals of Botany* 106: 497–504.

DE SMET, R., AND Y. VAN DE PEER. 2012. Redundancy and rewiring of genetic networks following genome-wide duplication events. *Current Opinion in Plant Biology* 15: 168–176.

SMITH, S. A, AND M.J. DONOGHUE. 2008. Rates of molecular evolution are linked to life history in flowering plants. *Science* 322: 86–89.

SOLTIS, D.E. ET AL. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* 96: 336–348.

SONG, K., LU, K. TANG, AND T.C. OSBORN. 1995. Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. *Proceedings of the National Academy of Sciences* 92: 7719–7723.

STEBBINS, G.L. 1940. The significance of polyploidy in plant evolution. *The American Naturalist* 74: 54–66.

STEBBINS, G.L. 1950. *Variation and Evolution in Plants*. Columbia University Press, New York, New York.

STEELE, R., K.L. HERTWECK, D. MAYFIELD, M.R. MCKAIN, J. LEEBENS-MACK, AND J.C. PIRES. 2012. Quality and quantity of data recovered from massively parallel sequencing: Examples in Asparagales and Poaceae. *American Journal of Botany* 99: 330–348.

TANG, H., J.E. BOWERS, X. WANG, R. MING, M. ALAM, AND A.H. PATERSON. 2007. Synteny and collinearity in plant genomes. *Science* 320: 486–488.

TANG, H., J.E. BOWERS, X. WANG, AND A.H. PATERSON. 2010. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proceedings of the National Academy of Sciences* 107: 472–477.

THOMAS, B.C., B. PEDERSEN, AND M. FREELING. 2006. Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dosage-sensitive genes. *Genome Research* 16: 934–946.

VANNESTE, K., Y. VAN DE PEER, AND S. MAERE. 2012. Inference of genome duplications from age distributions revisited. *Molecular Biology and Evolution* . Available at: http://mbe.oxfordjournals.org/content/early/2012/08/30/molbev.mss214.abstract.

VISION, T.J., D.G. BROWN, AND S.D. TANKSLEY. 2000. The origins of genomic duplications in *Arabidopsis*. *Science* 290: 2114–2117

WANG, Z., M. GERSTEIN, AND M. SNYDER. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10: 57–63.

WOOD, T.E., N. TAKEBAYASHI, M.S. BARKER, I. MAYROSE, B. GREENSPOON, AND L.H. RIESEBERG. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences* 106: 13875–13879.

CHAPTER II

DETERMINING THE TIMING OF RHO AND SIGMA POLYPLOIDY EVENTS IN THE HISTORY OF POALES

Polyploidy (or whole genome duplication) is a ubiquitous evolutionary phenomenon in the history of flowering plants. Estimates of the frequency of polyploidy in angiosperms have ranged from 30% (Stebbins, 1950) to 80% (Goldblatt, 1980; Lewis, 1980) based on analyses of chromosome counts. However, chromosome counts alone are not sufficient for inferring polyploidy because chromosome reduction often occurs following polyploid events (Blanc and Wolfe, 2004; Lysak et al., 2006) and, in some cases, within as few as 12 generations (Song et al., 1995; Xiong et al., 2011). Wood et al. (2009) used chromosome counts, allowing for among genera variation in chromosome number and for chromosome reduction, to estimate incidence of neopolyploids in extant angiosperm lineages and found cases of polyploidy ranging from ~30% in basal monocots to ~49% in within the monocots (Asparagales and above). Genomic and transcriptomic analyses of duplicate gene retention (Bowers et al., 2003; Blanc and Wolfe, 2004) and collinearity of gene order (Bowers et al., 2003; Tang et al., 2007) have been used to identify paleopolyploid events, occurring millions of years ago. A phylogenomic analysis of assembled genomes and large transcriptomes revealed a whole genome duplication (WGD) event in a common ancestor of all seed plants, and a second event that is shared by all angiosperms (Jiao et al., 2011). Whole genome duplication events have also influenced the evolution of many angiosperm lineages, including the eudicots (Jiao et al., 2012), Fabaceae (Blanc and Wolfe, 2004; Cannon et al., 2010), Brassicales (Vision et al., 2000; Blanc and Wolfe, 2004; Schranz and

Mitchell-olds, 2006; Barker et al., 2009), Poaceae (Paterson et al., 2004; Tang et al., 2010), Asteraceae (Barker et al., 2008), the commelinids (D'Hont et al., 2012), magnoliids and Nymphales (Cui et al., 2006).

Polyploidy is hypothesized to have played a major role in the evolution of key innovations in the angiosperms through the propagation of gene and gene networks through duplication (Ohno, 1970; Stebbins, 1971; Levin, 1983; Van de Peer et al., 2009). The redundancy of duplicated genes and gene networks creates an intrinsic "genomic buffer" allowing for non-deleterious gene loss or shift in function (Gu et al., 2003). Specialization or novelty in gene function may contribute to diversification as polyploids undergo diploidization through fractionation (i.e. gene loss (Freeling, 2009)). Further, reciprocal gene loss (differential gene duplicate loss from parental genomes) can create reproductive isolation and ultimately speciation through a variation of the Bateman-Dobzhansky-Mueller incompatibility (Werth and Windham, 1991; Orr, 1996; Lynch and Force, 2000; Taylor et al., 2001; Soltis et al., 2009). Although reciprocal gene loss does appear to be evident in *Sacchromyces* (Scannell et al., 2006) and teleost fishes (Postlethwait et al., 2000; Naruse et al., 2004), recent work examining syntenic gene deletion in grasses suggests that reciprocal gene loss may not be a driving force of diversification in this group since fractionation seems to be biased to one or the other subgenome (i.e. preferential retention of one parental genome over the other) (Schnable et al., 2012). Genomic redundancy provides an opportunity for shifts in gene function. Novelty can arise either through the evolution of gene sequences (Taylor and Raes, 2004) or the rewiring of gene networks (De Smet and Van de Peer, 2012) and can lead to potentially major evolutionary innovations, such as, the flower (De Bodt et al., 2005; Soltis et al., 2009; Jiao et al., 2011).

Paleopolyploid events in the grasses have been well documented with multiple more recent events (e.g. *Zea* (Gaut and Doebley, 1997) and *Sacchrum* (Grivet et al., 1996)). An older event (~70 million years ago (mya)), identified as *rho* (Tang et al., 2010) is shared by the PACCMAD and BEP clades, as defined by the Grass Phylogeny Working Group (2001) (Blanc and Wolfe, 2004; Paterson et al., 2004; Schlueter et al., 2004; Yu et al., 2005; Wang et al., 2011). An even older event based on relative K_s values, designated as *sigma*, was estimated to have occurred sometime prior to the diversification of Poaceae (Tang et al., 2010). The *rho* event has been implicated in the success of Poaceae through the role of duplicated MADS-box genes in the development of the spikelet (Preston and Kellogg, 2006; Preston et al., 2009) and the role of duplicated pathways in the development of starch-rich seeds (Wu et al., 2008; Comparot-Moss and Denyer, 2009). The placement of the *sigma* event (~ 130 mya) within the monocot phylogeny is unclear (Tang et al., 2010). The sequencing of the *Musa* genome, the first non-grass monocot genome, allowed genomic comparison of much deeper divergent lineages within the monocots and placed the *rho* and *sigma* events on the Poales lineage prior to the divergence of the PACMADD and BEP clades (D'Hont et al., 2012). A third paleopolyploid event in the history of the grass genomes was suggested by Tang et al. (Tang et al., 2010) but analyses of the banana genome did not yield evidence for this third event (D'Hont et al., 2012). Analysis of the *APETALA1/FRUITFULL* MADS-box gene family supports duplication within this gene family that occurred prior to the divergence of Poales and Commelinales (Litt and Irish, 2003), which may indicate of older WGD event in the monocots. Polyploidy has played a major role in the evolution of the grasses and other monocots, but a better understanding of when these events occurred in the phylogeny is needed to ascertain their influence on monocot evolution.

In this study, transcriptomes of major clades and families within the monocot order Poales were sequenced and compared to existing transcriptome and genomic data to determine the timing of paleopolyploid events. The phylogenetic placement of WGD events has been estimated using syntenic block analyses of genomes (Paterson et al., 2004; Jaillon et al., 2007; Tang et al., 2007, 2010), analyses of synonymous substitutions per synonymous site (K_s) and interspecies comparisons of transcriptomes (Lynch and Conery, 2000; Blanc and Wolfe, 2004; Schlueter et al., 2004; Cui et al., 2006; Barker et al., 2008, 2009; Shi et al., 2010), and phylogenomic analysis of gene families (Bowers et al., 2003; Jiao et al., 2011, 2012; McKain et al., 2012). A phylogenomic approach is taken to compare transcriptomes and genomes from 38 angiosperm species, effectively placing two paleopolyploid events (*rho* and *sigma*) on the Poales phylogeny and giving strong evidence for a third, deep monocot paleopolyploid event (*tau*) prior to the divergence of Asparagales and the commelinids. These events are then considered within the context of the role of WGD in key innovations and species diversification within the monocots.

Materials and Methods

Taxon sampling

A recent study using a large, chloroplast data set has resolved many of the relationships within Poales (Givnish et al., 2010). Taxa sampled for this study represent all major clades and families found in Poales. Transcriptome and genome data sets were combined from Genbank, Phytozome, the OneKP project, and new species sampled for this study. Table 2.1 summarizes species, data types, and sources.

RNA isolation and sequencing

RNA was isolated from fresh young leaf or apical meristematic tissue using an RNeasy Plant Mini Kit (Qiagen, Valencia, California, USA). Samples were kept on liquid nitrogen prior to isolation. RNA was eluted into a final volume of 100 µL of RNase-free water.

RNA total mass and quality were estimated using an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, California, USA). Samples were deemed acceptable if RIN scores were greater than 8.0.

The TruSeq RNA Sample Preparation Kit (Illumina, San Diego, California, USA) was used to construct pair-end libraries with an average fragment length of 300 base pairs (bp). RNAseq libraries were sequenced on Illumina HiSeq sequencers (Illumina, San Diego, California, USA), six samples per lane, at the DNA Core Facility at the University of Missouri-Columbia or Cold Spring Harbor Labs.

Transcriptome assembly

Illumina data generated for this study, generated by the OneKP project, and downloaded from the NCBI Short Read Archive (SRA) were all assembled as follows. FastQC v0.10.1 (Andrews, 2010) was used to determine if sequencing adapter contamination was present in reads. Cutadapt v1.1 (Martin, 2011) was used to clean contaminated sequences identified by FastQC from reads. Two mismatches per 10 bp were allowed and a minimum overlap of at least 10 bp was required for contaminant identification and trimming. Cleaned reads were further trimmed using a custom perl script (available from M. R. McKain upon request) that trimmed reads from the ends until there were three consecutive bases with a quality score of 20 or more.

Reads with a median quality score less than 22 and with more than three uncalled bases were filtered out. Trimmed reads less than 40 bp were removed.

Cleaned and filtered data sets were assembled using Trinity (Release 2012-06-08) (Grabherr et al., 2011) with default parameters. Reads were aligned to the Trinity assembly using bowtie v0.12.8 (Langmead et al., 2009) through the alignReads.pl script available in the Trinity distribution. Output from this script was then piped into the run_RSEM.pl script (also packaged with Trinity), which utilizes RSEM v1.2.0 (Li and Dewey, 2011) to quantify transcript abundance. Fragments per kilobase of exon per million fragments mapped (FPKM) was estimated for each component (gene), and the percentage of these mapped fragments that correspond to each isoform assembled per component was estimated using the summarize_RSEM_fpkm.pl script packaged with Trinity using an average fragment length of 300. Isoforms that had 1% or less of all fragments mapped to a component were filtered out of the Trinity assembly to create a data set representing well supported transcripts.

Transcriptome data for *Elaeis guineensis* (GenBank accession numbers: SRX059258-SRX059263) was generated on the 454 pyrosequencing platform. These reads were assembled with MIRA (Chevreux et al., 2004) using default parameters.

Twenty-two-genome data set used for gene family circumscription

A data set comprised of the coding DNA sequence (CDS) for 22 green plant genomes was compiled by collaborators in the dePamphilis lab (Penn State, University Park, PA) and used to circumscribe gene families as OrthoMCL clusters (Table 2.2). Gene family circumscription for the 22-genome set was estimated using the inferred amino acid sequence from CDS data. An all-by-all blastp (Altschul et al., 1997) search was conducted on the concatenated 22-genome

amino acid sequences used for clustering. Gene family circumscriptions were estimated by clustering amino acid sequences using OrthoMCL v2.0 (Li et al., 2003) and suggested parameter settings.

Transcriptome translation and gene family circumscription

Translation of the transcriptome assembly was conducted using a set of custom of perl scripts (available from M. R. McKain upon request). A blastx search of transcriptome assemblies against the 22-genome set amino acid sequences was conducted using a cutoff e-value of 1e-10. The output of this blast was filtered for best hits, identified by lowest e-value, identifying a single best hit to amino acid sequence from the 22-genome set for each transcriptome contig. Each transcriptome contig was then translated using GeneWise (part of the Wise2 v2.2.0 package) (Birney et al., 2004) and the best hit amino acid sequence from the 22-genome set. GeneWise predicts gene structure using homology of similar protein sequences. Cleaning scripts are implemented in perl that take the longest GeneWise translation from either direction, remove internal stop codons and fragment CDS codons, and splice together cleaned CDS and amino acid sequences into a final version.

Blastp searches of translated transcriptome assemblies against the 22-genome amino acid sets were run with an e-value cutoff of 1e-10 and then filtered for best hit per translated contig based on e-value. Translated contigs were then sorted into the 22-genome gene families using the best hit to a 22-genome set sequence. Taxa in the 22-genome set but not listed in Table 2.1 were removed from gene families prior to further analysis.

Estimation of syntelogs and gene tree estimation for gene families

To elucidate the history of WGD in Poales and the monocots, gene families were collapsed using syntelogous gene pairs identified through synteny analyses. Syntelogs are paralogous genes, duplicates created during a WGD or segmental duplication event, identified through collinearity in regions of the genome. Syntelogs were identified in the *Oryza sativa* and *Sorghum bicolor* genome using the methodology of Tang et al. (2010). Two rounds of synteny were identified in the *Oryza* and *Sorghum* genomes. The first set of syntenic regions was defined using a chaining distance of 40 genes. These syntenic blocks were assigned to the *rho* WGD event. A second set of syntenic blocks was reconstructed through comparisons of the *rho* blocks using a chaining distance of 60 genes. These older syntenic blocks were the *sigma* WGD event. Syntelogous pairs were identified to either the *rho* or *sigma* events and used to collapse gene families if pair members were found to be in separate gene families.

Peptide sequences within gene families were aligned using MUSCLE v3.8.31 (Edgar, 2004), and CDS were aligned onto the amino acid alignments using PAL2NAL v13 (Suyama et al., 2006). Alignments were filtered using two criteria: columns in alignments were removed if gaps were observed in more than 90% of the sequences (rows) and transcript translations (rows) were removed if they covered less than 30% of the total alignment length for the gene family. Maximum likelihood (ML) gene trees were estimated using RAxML v7.3.0 (Stamatakis, 2006) with the GTR + gamma evolutionary model and 500 bootstrap replicates. Gene trees were rooted to outgroup taxa (*Amborella*, *Aquilegia*, and *Vitis*) found in each gene family.

The timing of whole genome duplications within the scope of estimated syntenic blocks were assessed by querying estimated gene family trees for the last common ancestor (LCA) of all syntelog pairs and their descendant genes using custom perl scripts (available from M. R.

McKain upon request). The sister lineage to the clade containing the syntelog pairs was also assessed, and genes present were recorded. Gene trees with the LCA node of the syntelog pair with a bootstrap value (BSV) of ≥ 50 were retained for further analysis. The LCA node of a particular gene tree was discarded if the outgroups (*Amborella*, *Aquilegia*, or *Vitis*) were embedded within the syntelog pair clade. This was accessed on a node-by-node basis, as some trees had multiple nodes depicting different syntelog pair LCAs. Assessment of the timing of duplication events (either *rho* or *sigma*) was conducted by examining taxa above the LCA node to estimate a potential placement of the origin of the pair. The node was discarded when the sister lineage to the LCA node contained any taxa found within the LCA clade. The node was also discarded if the sister lineage to the LCA clade did not contain taxa from the sister lineage to the corresponding node in the species tree (Givnish et al., 2010). The remaining LCA nodes and gene trees were ranked by assigned WGD event (*rho* or *sigma*), BSV, and the relative placement of the event within the species phylogeny. Gene tree topologies were also inspected manually to verify the results of automated analyses.

Results

Assemblies and translations

Table 2.3 summarizes contig counts and lengths for post-translation assemblies for all species assembled. All translations were used for gene tree analyses.

Phylogenetic analysis of gene trees

To determine the timing of *rho* and *sigma* duplication events, phylogenies of gene families containing syntelogous pairs identified to either event from *Sorghum bicolor* and *Oryza*

sativa genomes were estimated. A total of 53,136 gene families were circumscribed using OrthoMCL clustering of the 22-genome data set. The synteny analyses identified 56 *Oryza sativa rho* blocks, 58 *Orzya sativa sigma* blocks, 39 *Sorghum bicolor rho* blocks, and 63 *Sorghum bicolor sigma* blocks. Within these blocks, there were 4296 syntelog pairs from *Oryza rho* blocks, 1782 syntelog pairs from *Oryza sigma* blocks, 3971 syntelog pairs from *Sorghum rho* blocks, and 1898 syntelog pairs from *Sorghum sigma* blocks. A total of 1692 gene families were collapsed into 1186 larger gene families in order to join syntelog pairs that were split across gene families. The instances of collapsing were split among the species and block types as follows: 835 from *Oryza sativa rho* blocks, 249 from *Oryza sativa sigma* blocks, 376 from *Sorghum bicolor rho* blocks, and 232 from *Sorghum bicolor sigma* blocks. After gene families were collapsed, those that did not contain outgroup species (*Amborella*, *Aquilegia*, and *Vitis*) or syntelog pairs were discarded. A total of 6612 syntelog pairs were found in 2116 gene families with outgroup species sampled. All 2116 alignments and ML gene family trees will be deposited in the DRYAD database in the coming months when this work is submitted for publication.

Syntelog pairs were used to identify their corresponding WGD within the gene trees, which corresponds to the LCA of the pairs. These LCAs were filtered based on bootstrap values (BSV), and all LCA nodes with BSV less than 50% were discarded. There were 3396 syntelogs pairs found in 1511 gene family trees with LCA node BSV greater than or equal to 50%.

Relationships of genes within ML gene family trees relative to identified syntelog pair LCAs were analyzed in the context of species relationships (Givnish et al., 2010). Gene family trees were queried to determine taxa that were found within the syntelog pair LCA node. Taxa in the lineage sister to the syntelog pair LCA node were also queried. An acceptable LCA node had a sister lineage with at least one taxon present in the sister lineage to the equivalent node in the

species tree. There also could not be any taxa present in the sister lineage that were found within the LCA clade. This approach allowed a conservative estimate of the timing of duplication events by only using LCA nodes that give unequivocal interpretations of when events occurred on the species phylogeny. These filtering methods were necessary due to the data sampling methodology that utilizes a shotgun transcriptome sequencing approach. By filtering based on presence/absence of taxa within a gene family tree in relationship to a species tree, gene family trees that have broad taxon sampling were used to identify timing of these events despite the high likelihood of missing data. Moreover, these trees may include artifacts due to poor and typically incomplete sequence alignment. Therefore, only the filtered set of trees was used to infer the timing of WGD events. The only assumption about the timing of WGD events imposed by the filtering steps was that all WGDs occurred after the divergence of monocots and eudicots.

Of the 1511 gene family trees that passed these filtering steps, 872 were informative for estimating the timing of WGD events. Within these 872 trees, 912 unique syntelog pair LCAs were identified as informative. These 912 LCAs represent 1529 different syntelog pairs. Informative gene families included 762 trees with signal for the placement of *rho* block syntelog pair duplications and 110 trees with signal for the placement of *sigma* block syntelog pair duplications. The *rho* block trees included 800 unique LCA nodes (i.e. duplication events), defining the ancestors of 1258 syntelog pairs, while the *sigma* block gene tree had 112 unique LCA nodes defining the ancestors of 271 syntelog pairs. Syntelog pair LCA nodes were evaluated based on BSV and classified into two groups, $50\% \leq BSV < 80\%$ and $BSV \geq 80\%$.

Gene family trees containing LCAs of syntelog pairs in *rho*-identified syntenic blocks showed evidence for the placement of the *rho* WGD event as prior to the origin of Poaceae but after the divergence of Ecdeiocoleaceae. Of 800 informative LCA nodes, 373 were supported by

a BSV of at least 80% for a Poaceae-specific *rho* event, and another 122, a BSV of at least 50%, representing a total of 61.9% of informative LCA nodes for *rho* syntelog pairs (Fig. 2.1 A, Table 2.4). A second potential placement of the *rho* event within Poaceae, after the divergence of *Streptochaeta* was suggested by 33% of the informative LCA nodes, with 113 LCAs of 80% or greater BSV and 151 with BSV of at least 50%. The number of pairs represented by these LCAs showed similar proportions with 60.3% of the pairs suggesting a pre-diversification Poaceae event while 32.4% suggest a post-*Streptochaeta* divergence event (Fig. 2.1 B, Table 2.4). The phylogenetic position of 26 LCA nodes suggested placement *of rho* at other points within Poales (6 LCAs BSV of at least 80%), representing 44 syntelog pairs. Twelve LCAs suggested a placement prior to the divergence of Asparagales and the commelinids (10 LCAs BSV of at least 80%), representing 22 syntelog pairs.

Analysis of gene family trees that contained LCAs of syntelog pairs assigned to *sigma* syntenic blocks show evidence of two events, a younger event (*sigma*) and an older event (*tau*). Of 112 informative LCA nodes, 20 exhibit a BSV of at least 80% in support of a WGD event occurring after the divergence of the lineage leading to Poales but prior to the diversification of Poales. Additionally, 12 LCA nodes had a BSV of at least 50%, representing 28.6% of the LCA nodes initially assigned to the *sigma* WGD (Fig. 2.2 A, Table 2.5). Strong evidence of an event that occurred prior to the divergence of Asparagales is found in the originally hypothesized *sigma* block LCA nodes, with 53, with a BSV of at least 80%, and 15, with a BSV of at least 50%, representing 60.7% of all informative *sigma* LCA nodes (Fig. 2.2 A). Of the 271 syntelog pairs represented by the 112 *sigma* LCA nodes, 61 (48 of BSV 80% or greater) supported the Poales event, and 174 (140 of BSV 80% or greater) supported the pre-Asparagales-commelinid divergence event, representing 22.5% and 64.2% of informative *sigma* syntelog pairs,

respectively (Fig. 2.2 B, Table 2.5). Other potential placements of *sigma* blocks were negligible with 3 LCA representing within Poales (BSV 50% or greater), 1 LCA representing Zingiberales (BSV 80% or greater), and 8 LCAs representing Arecales (6 BSV 50% or greater and 2 BSV 80% or greater).

The analyses of *rho* and *sigma* block syntelog pairs in gene family trees suggest at least three WGD events in the history of all Poaceae genomes. The earliest, *rho*, identified by Paterson et al. (2004) to be common to all grasses occurs on the lineage after the divergence of Poaceae from the graminids but prior to diversification of the family. The second event, *sigma*, first identified by Tang et al. (2010) occurred on the lineage after the divergence of Poales from the commelinids but prior to diversification. A third event, *tau*, suggested by Tang et al. (2010), occurred prior to the divergence of Asparagales from the commelinids, but its exact placement is unclear.

Discussion

Understanding the effect of polyploidy on the evolution of angiosperms requires examining modern and ancient events for both short term and long term effects. Detection and phylogenetic mapping of paleopolyploidy events is becoming a more prominent endeavor as large amounts of genomic data are becoming available for a growing number of taxa. Methods for detecting and characterizing WGD events include looking at genomic structure (i.e. syntenic blocks of genes), distribution of divergence between genes based on synonymous substitutions (K_s), and phylogenetic analyses with gene families considered separately and duplications are inferred from gene family topology compared to species tree topology (Jaillon et al., 2009). Each of these methods has its caveats (Bowers et al., 2003; Vanneste et al., 2012) though

combinations can be used to give multiple lines of evidence to support events (Bowers et al., 2003; Paterson et al., 2004; Tang et al., 2010; Jiao et al., 2011, 2012; D'Hont et al., 2012; McKain et al., 2012).

The youngest event identified in this study is *rho*, previously shown as a WGD event shared by all members of the PACCMAD+BEP clade occurring ~ 70 mya (Paterson et al., 2004; Tang et al., 2010). A goal of this study was to determine if this event occurred prior to the diversification of Poaceae or within the family prior to the diversification of the PACCMAD+BEP clade (Soltis et al., 2009). Syntelog pairs from *rho* blocks indicated two possible positions for the timing of *rho*, prior to the splitting of *Streptochaeta* from the rest of the Poaceae and after the *Streptochaeta* split. *Streptochaeta* is part of the Anomochlooideae, the basal subfamily of Poaceae (Grass Phylogeny Working Group, 2001) representing the earliest possible branch within the family for comparison in this study. With 61.9% of the acceptable *rho* syntelog pair LCA nodes (Fig. 2.1), analysis of gene family trees suggests that the *rho* event occurred prior to the Poaceae diversification. Not only are the counts higher for this phylogenetic placement compared to post-Poaceae diversification placement (495 vs. 264, Fig. 2.1), further dissection of these counts shows that the number of highly supported (≥80% BSV) LCA nodes depicting the pre-Poaceae event is over three times that of the highly supported nodes supporting the post-Poaceae event. Ultimately, despite the fairly large number of LCA nodes in support of the post-Poaceae event, they are not well supported and could be artifacts of long-branch attraction caused by increased rates of molecular evolution in the core grasses (Eyre-Walker and Gaut, 1997; Muse, 2000) or potential differences in GC content between basal and core grass species (Haberer et al., 2005; Wang and Hickey, 2007; Muyle et al., 2011). These

potential issues might be through increased taxon sampling within the PACCMAD+BEP clade, the Anomochlooideae or the two unsampled basal subfamilies, Pharoideae and Puelioideae.

Placement of *rho* as prior to Poaceae diversification could elucidate the evolution of the gene networks involved in the development of grass features. The starch biosynthesis pathway duplicated genes in both *Oryza* and *Zea*, placing the origin prior to the diversification of the PACCMAD+BEP clade (Wu et al., 2008). This duplication has been hypothesized to be associated with *rho*, but the gene families involved did not sample basal Poaceae species or the other graminid families Ecdeiocoleaceae, Joinvilleaceae, and Flagellariaceae.

The spikelet is a potential synapomorphy of the grasses, and the evolution of this inflorescence type has been difficult to discern due to the similarity of inflorescence in closely related groups (Rudall and Stuppy, 2005; Sajo and Rudall, 2012) and the difficulty in determining homology in inflorescence structure between early diverging Poaceae species and the core Poaceae (Sajo et al., 2008, 2012; Preston et al., 2009). The influence of the *rho* WGD event on the development of the spikelet is suggested through analyses of the MADS-box transcription factor gene family, *AP1/FUL*, which demonstrated that the *FUL* gene was duplicated prior to the diversification of Poaceae (including *Streptochaeta*) but after the divergence of Joinvilleaceae (Preston and Kellogg, 2006). The paralogs were not maintained in duplicate in either *Streptochaeta* or *Pharus*, both early diverging lineages with variable inflorescence structures (Preston and Kellogg, 2006). These findings, in concordance with our placement of *rho*, may indicate that the retention of duplicate genes led to the development of the spikelet in grasses, although more work examining paralog retention and expression in the inflorescence in Poaceae is required.

The second event identified in this study corresponds to the *sigma* event first described by Tang et al. (2010). Synteny analyses and phylogenomic analyses that included commelinids (D'Hont et al., 2012) suggest that *sigma* occurred sometime prior to the diversification of Poaceae but after the divergence of Poales from commelinids as this event was not shared by taxa of Zingiberales or Arecales. The timing of this event was estimated as ~130 mya (Tang et al., 2010), prior to the diversification of the Poales crown group at ~108.95 mya (Magallón and Castillo, 2009), which is older than the estimated Poales stem group (~123.03 mya). Tang et al. (2010) cautioned that the rough age estimate was due to saturation of *sigma* paralogs in K_s analyses. Tang et al. (2010) did not include standard errors on their duplicate divergence time estimates. In this work, focus is on the timing of WGD events relative to speciation events rather than assigning divergence time estimates.

In this study, syntelog pairs identified to *sigma* blocks support two events within the history of all Poales taxa sampled. The younger event, placed prior to the diversification of Poales but after the divergence of Poales, is supported by 28.6% of informative syntelog pair LCAs found in *sigma* blocks (Fig. 2.2 A, Table 2.5). This is a relatively low percentage of the LCAs found in *sigma* blocks. Due to the lack of evidence of this event occurring within the Poales or at nodes relatively close to the Poales divergence node, these results suggest that this is an event with lower LCA support. Rate variation in Poales varies with chloroplast genome data (Givnish et al., 2010), which are correlated to rates of nuclear genomes (Eyre-Walker and Gaut, 1997). This variation could cause uncertainty in the estimation of gene family trees. When all syntelog pair LCA nodes are considered for *sigma* events regardless of BSV, there is a disproportionate number supportive of this placement of the *sigma* event with BSV values less than 50% (Table 2.5). This suggests that added support for placement of *sigma* may be obtained

with further taxon sampling to cut long branches within the Poales phylogeny. Another possible explanation is that duplicated genes were purged en mass instead of retained in duplicate. Transcriptomes in this study were sampled from young leaf and apical meristematic tissue, so retained duplicate genes could be expressed in unsampled tissue, at varying ages, or under other environmental circumstances.

The impact of the *sigma* event on the evolution of Poales is reflected in the diversity of the group. The order comprises ca. 21,000 species, representative of ~33% of all monocot species and is ecologically dominant in a number of habitats (Linder and Rudall, 2005; Givnish et al., 2010). Three Poales families contain most of the species: Poaceae (~11,300 spp.), Cyperaceae (~5,700 spp.), and Bromeliaceae (~3,100 spp.) (The Plant List, 2010). Though the number of species in Poaceae may be attributable to the *rho* event due to diversification in the PACCMAD+BEP clade, other diversifications could be linked to innovations spurred by gene duplications from the *sigma* event, such as the evolution of the epiphytic habit in Bromeliaceae (Givnish et al., 2011). Investigation into the evolutionary effects of the *sigma* event will require further identification of duplicated genes and gene networks derived from the event and deep genomic sampling across Poales.

The third event, *tau*, identified by this study occurred prior to the divergence of Asparagales from the commelinids and was suggested previously (Tang et al., 2010), though there was no indication of when it occurred. Syntelog pair LCA nodes identified from *sigma* blocks showed overwhelming support for this event (60.7% of informative *sigma* LCAs). A similarly timed duplication of MADS-box genes has been reported as shared by Commelinales and Poales, although sampling and support was not high enough to further estimate the timing of the duplication (Litt and Irish, 2003). The lack of resolution between *sigma* and *tau* in the

synteny analysis suggests that there may be conservation of these blocks in monocot genomes, which may indicate important or vital function (Lee et al., 2006). This event was not identified in the analysis of the banana genome (D'Hont et al., 2012). This omission in the banana genome analysis is attributable to the lower number of syntelog pairs analyzed in that study relative to the present study. Another possible explanation for the lack of evidence for *tau* in the *Musa* genome study may be the taxon sampling. Only members of the PACCMAD+BEP clade were used to represent Poales and may be a case of long-branch attraction due to the high rates of evolution within the grasses. Increased taxon sampling can provide a better estimate of phylogenetic relationships by breaking long branches (Leebens-Mack et al., 2005). Further sampling of the monocots is needed to identify the placement of *tau* and to understand possible implications of this event for the history of the monocots.

Phylogenomic analyses are becoming much more common as high-throughput sequencing becomes the standard. These types of analyses provide insight into the evolution of large portions of the genome of many taxa previously not sampled in genomic studies. Combining phylogenomic analyses with other methods for identifying WGD events, such as synteny analysis, allows for the detection of paleopolyploid events and their phylogenetic placement. Understanding when these events occurred in the history of angiosperms, will help elucidate the long-term evolutionary patterns associated with polyploidy and may further the understanding of how this group of organisms has become so widespread and diverse.

References

ALTSCHUL, S.F. ET AL. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25: 3389–3402.

ANDREWS, S. 2010. FastQC. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

The Plant List. 2010. The Plant List. Version 1. Available at: http://www.theplantlist.org/ [Accessed September 8, 2012].

BARKER, M.S. ET AL. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution* 25: 2445–2455.

BARKER, M.S., H. VOGEL, AND M.E. SCHRANZ. 2009a. Paleopolyploidy in the Brassicales: Analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. *Genome Biology and Evolution* 1: 391–399.

BIRNEY, E., M. CLAMP, AND R. DURBIN. 2004. GeneWise and Genomewise. *Genome Research* 14: 988–95.

BLANC, G., AND K.H. WOLFE. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant Cell* 16: 1667–1678.

DE BODT, S., S. MAERE, AND Y. VAN DE PEER. 2005. Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution* 20: 591–597.

BOWERS, J.E., B.A. CHAPMAN, J. RONG, AND A.H. PATERSON. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422: 433–438.

CANNON, S.B. ET AL. 2010. Polyploidy did not predate the evolution of nodulation in all legumes. *PloS One* 5: e11630.

CHEVREUX, B. ET AL. 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research* 14: 1147–59.

COMPAROT-MOSS, S., AND K. DENYER. 2009. The evolution of the starch biosynthetic pathway in cereals and other grasses. *Journal of Experimental Botany* 60: 2481–2492.

CUI, L. ET AL. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Research* 16: 738–749.

D'HONT, A. ET AL. 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* 488: 213–217.

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

EYRE-WALKER, A., AND B.S. GAUT. 1997. Correlated rates of synonymous site evolution across plant genomes. *Molecular Biology and Evolution* 14: 455–460.

FREELING, M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual Review of Plant Biology* 60: 433–453.

GAUT, B.S., AND J.F. DOEBLEY. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences* 94: 6809–6814.

GIVNISH, T.J. ET AL. 2010. Assembling the tree of the Monocotyledons: plastome sequence phylogeny and evolution of Poales. *Annals of the Missouri Botanical Garden* 97: 584–616.

GIVNISH, T.J. ET AL. 2011. Phylogeny, adaptive radiation, and historical biogeography in Bromeliaceae: insights from an eight-locus plastid phylogeny. *American Journal of Botany* 98: 872–895.

GOLDBLATT, 1980. Polyploidy in angiosperms: Monocotyledons. In Lewis W. H. [ed.], Polyploidy: Biological relevance. Plenum Press, New York, New York.

GRABHERR, M.G. ET AL. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29: 644–652.

GRIVET, L., A. D'HONT, AND D. ROQUES. 1996. RFLP mapping in cultivated sugarcane (*Saccharum* sp): genome organization in a highly polyploid and aneuploid interspecific hybrid. *Genetics* 142: 987–1000.

GRASS PHYLOGENY WORKING GROUP. 2001. Phylogeny and subfamilial classification of the grasses (Poaceae). *Annals of the Missouri Botanical Garden* 88: 373–457.

GU, Z., L.M. STEINMETZ, X. GU, C. SCHARFE, R.W. DAVIS, AND W.-H. LI. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature* 421: 63–66.

HABERER, G. ET AL. 2005. Structure and architecture of the Maize genome. *Plant Physiology* 139: 1612–1624.

JAILLON, O. ET AL. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449: 463–467.

JAILLON, O., J.-M. AURY, AND WINCKER. 2009. "Changing by doubling", the impact of whole genome duplications in the evolution of eukaryotes. *Comptes Rendus Biologies* 332: 241–253.

JIAO, Y. ET AL. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biology* 13: R3.

JIAO, Y. ET AL. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.

LANGMEAD, B., C. TRAPNELL, M. POP, AND S.L. SALZBERG. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10: R25.

LEE, A., E.G.L. KOH, A. TAY, S. BRENNER, AND B. VENKATESH. 2006. Highly conserved syntenic blocks at the vertebrate Hox loci and conserved regulatory elements within and outside Hox gene clusters. *Proceedings of the National Academy of Sciences* 103: 6994–6999.

LEEBENS-MACK, J. ET AL. 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.

LEVIN, D. A.. 1983. Polyploidy and novelty in flowering plants. *The American Naturalist* 122(1): 1-25.

LEWIS, W.H. 1980. Polyploidy in angiosperms: Dicotyledons. In W. H. Lewis [ed.], Polyploidy: Biological relevance. Plenum Press, New York, New York.

LI, B., AND C.N. DEWEY. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12: 323. .

LI, L., C.J. STOECKERT, AND D.S. ROOS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research* 13: 2178–2189.

LINDER, H., AND J. RUDALL. 2005. Evolutionary history of Poales. *Annual Review of Ecology, Evolution, and Systematics* 36: 107–124.

LITT, A., AND V.F. IRISH. 2003. Duplication and diversification in the *APETALA1/FRUITFULL* floral homeotic gene lineage: implications for the evolution of floral development. *Genetics* 165: 821–833.

LYNCH, M., AND J.S. CONERY. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151–1155.

LYNCH, M., AND A. FORCE. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154: 459–473.

LYSAK, M. A, A. BERR, A. PECINKA, R. SCHMIDT, K. MCBREEN, AND I. SCHUBERT. 2006. Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. *Proceedings of the National Academy of Sciences* 103: 5224–5229.

MAGALLÓN, S., AND A. CASTILLO. 2009. Angiosperm diversification through time. *American Journal of Botany* 96: 349–65.

MARTIN, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17: 10–12.

MCKAIN, M.R. ET AL. 2012. Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in Agavoideae (Asparagaceae). *American Journal of Botany* 99: 397–406.

MUSE, S.V. 2000. Examining rates and patterns of nucleotide substitution in plants. *Plant Molecular Biology* 42: 25–43.

MUYLE, A., L. SERRES-GIARDI, A. RESSAYRE, J. ESCOBAR, AND S. GLÉMIN. 2011. GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). *Molecular Biology and Evolution* 28: 2695–2706.

NARUSE, K., M. TANAKA, K. MITA, A. SHIMA, J. POSTLETHWAIT, AND H. MITANI. 2004. A medaka gene map: the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. *Genome Research* 14: 820–828.

OHNO, S. 1970. *Evolution by Gene Duplication*. Springer-Verlag.

ORR, H.A. 1996. Dobzhansky, Bateson, and the genetics of speciation. *Genetics* 144: 1331–1335.

PATERSON, A.H., J.E. BOWERS, AND B.A. CHAPMAN. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences* 101: 9903–9908.

VAN DE PEER, Y., S. MAERE, AND A. MEYER. 2009. The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics* 10: 557–564.

PLANT LIST. 2010. The plant list, version 1 [online]. Available at: http://www.theplantlist.org/ [Accessed September 8, 2012].

POSTLETHWAIT, J.H. ET AL. 2000. Zebrafish Comparative genomics and the origins of vertebrate chromosomes. *Genome Research* 10: 1890–1902.

PRESTON, J.C., A. CHRISTENSEN, S.T. MALCOMBER, AND E.A. KELLOGG. 2009. MADS-box gene expression and implications for developmental origins of the grass spikelet. *American Journal of Botany* 96: 1419–1429.

PRESTON, J.C., AND E.A. KELLOGG. 2006. Reconstructing the evolutionary history of paralogous *APETALA1/FRUITFULL*-like genes in grasses (Poaceae). *Genetics* 174: 421–437.

RUDALL, , AND W. STUPPY. 2005. Evolution of reproductive structures in grasses (Poaceae) inferred by sister-group comparison with their putative closest living relatives. *American Journal of Botany* 92: 1432–1443.

SAJO, M.G., H.M. LONGHI-WAGNER, AND J. RUDALL. 2008. Reproductive morphology of the early-divergent grass *Streptochaeta* and its bearing on the homologies of the grass spikelet. *Plant Systematics and Evolution* 275: 245–255.

SAJO, M.G., N. PABÓN-MORA, J. JARDIM, D.W. STEVENSON, AND J. RUDALL. 2012. Homologies of the flower and inflorescence in the early-divergent grass *Anomochloa* (Poaceae). *American Journal of Botany* 99: 614–28.

SAJO, M.G., AND J. RUDALL. 2012. Morphological evolution in the graminid clade: comparative floral anatomy of the grass relatives Flagellariaceae and Joinvilleaceae. *Botanical Journal of the Linnean Society* 170: 393–404. .

SCANNELL, D.R., K. BYRNE, J.L. GORDON, S. WONG, AND K.H. WOLFE. 2006. Multiple rounds of speciation associated wtih reciprocal gene loss in polyploid yeasts. *Nature* 440: 341–345.

SCHLUETER, J.A. ET AL. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* 47: 868–876.

SCHNABLE, J.C., M. FREELING, AND E. LYONS. 2012. Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biology and Evolution* 4: 265–77.

SCHRANZ, M.E., AND T. MITCHELL-OLDS. 2006. Independent Ancient Polyploidy Events in the Sister Families Brassicaceae and Cleomaceae. *Plant Cell* 18: 1152–1165.

SHI, T., H. HUANG, AND M.S. BARKER. 2010. Ancient genome duplications during the evolution of kiwifruit (*Actinidia*) and related Ericales. *Annals of Botany* 106: 497–504.

DE SMET, R., AND Y. VAN DE PEER. 2012. Redundancy and rewiring of genetic networks following genome-wide duplication events. *Current Opinion in Plant Biology* 15: 168–176.

SOLTIS, D.E. ET AL. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* 96: 336–348.

SONG, K., LU, K. TANG, AND T.C. OSBORN. 1995. Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. *Proceedings of the National Academy of Sciences* 92: 7719–7723.

STAMATAKIS, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Nucleic Acids Research* 22: 2688–2690.

STEBBINS, G.L. 1950. *Variation and Evolution in Plants*. Columbia University Press, New York, New York.

STEBBINS, G.L. 1971. *Chomosomal evolution in higher plants*. Addison-Wesley, Reading, Massachusetts.

SUYAMA, M., D. TORRENTS, AND BORK. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* 34: W609–W612.

TANG, H., J.E. BOWERS, X. WANG, R. MING, M. ALAM, AND A.H. PATERSON. 2007. Synteny and collinearity in plant genomes. *Science* 320: 486–488.

TANG, H., J.E. BOWERS, X. WANG, AND A.H. PATERSON. 2010. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proceedings of the National Academy of Sciences* 107: 472–477.

TAYLOR, J.S., Y. VAN DE PEER, AND A. MEYER. 2001. Genome duplication, divergent resolution and speciation. *Trends in Genetics* 17: 459–473.

TAYLOR, J.S., AND J. RAES. 2004. Duplication and divergence: the evolution of new genes and old ideas. *Annual Review of Genetics* 38: 615–643.

VANNESTE, K., Y. VAN DE PEER, AND S. MAERE. 2012. Inference of genome duplications from age distributions revisited. *Molecular Biology and Evolution* . Available at: http://mbe.oxfordjournals.org/content/early/2012/08/30/molbev.mss214.abstract.

VISION, T.J., D.G. BROWN, AND S.D. TANKSLEY. 2000. The origins of genomic duplications in *Arabidopsis*. *Science* 290: 2114–2117.

WANG, H.-C., AND D. A HICKEY. 2007. Rapid divergence of codon usage patterns within the rice genome. *BMC Evolutionary Biology* 7: 241–253.

WANG, X., H. TANG, AND A.H. PATERSON. 2011. Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major Poaceae lineages. *The Plant Cell* 23: 27–37.

WERTH, C.R., AND M.D. WINDHAM. 1991. A model for divergent, allopatric speciation of polyploid Pteridophytes resulting from silencing of duplicate-gene expression. *The American Naturalist* 137: 515–526.

WOOD, T.E., N. TAKEBAYASHI, M.S. BARKER, I. MAYROSE, B. GREENSPOON, AND L.H. RIESEBERG. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences* 106: 13875–13879.

WU, Y., Z. ZHU, L. MA, AND M. CHEN. 2008. The preferential retention of starch synthesis genes reveals the impact of whole-genome duplication on grass evolution. *Molecular Biology and Evolution* 25: 1003–1006.

XIONG, Z., R.T. GAETA, AND J.C. PIRES. 2011. Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*. *Proceedings of the National Academy of Sciences* 108: 7908–7913.

YU, J. ET AL. 2005. The genomes of *Oryza sativa*: a history of duplications. *PLoS Biology* 3: 1003–1006.

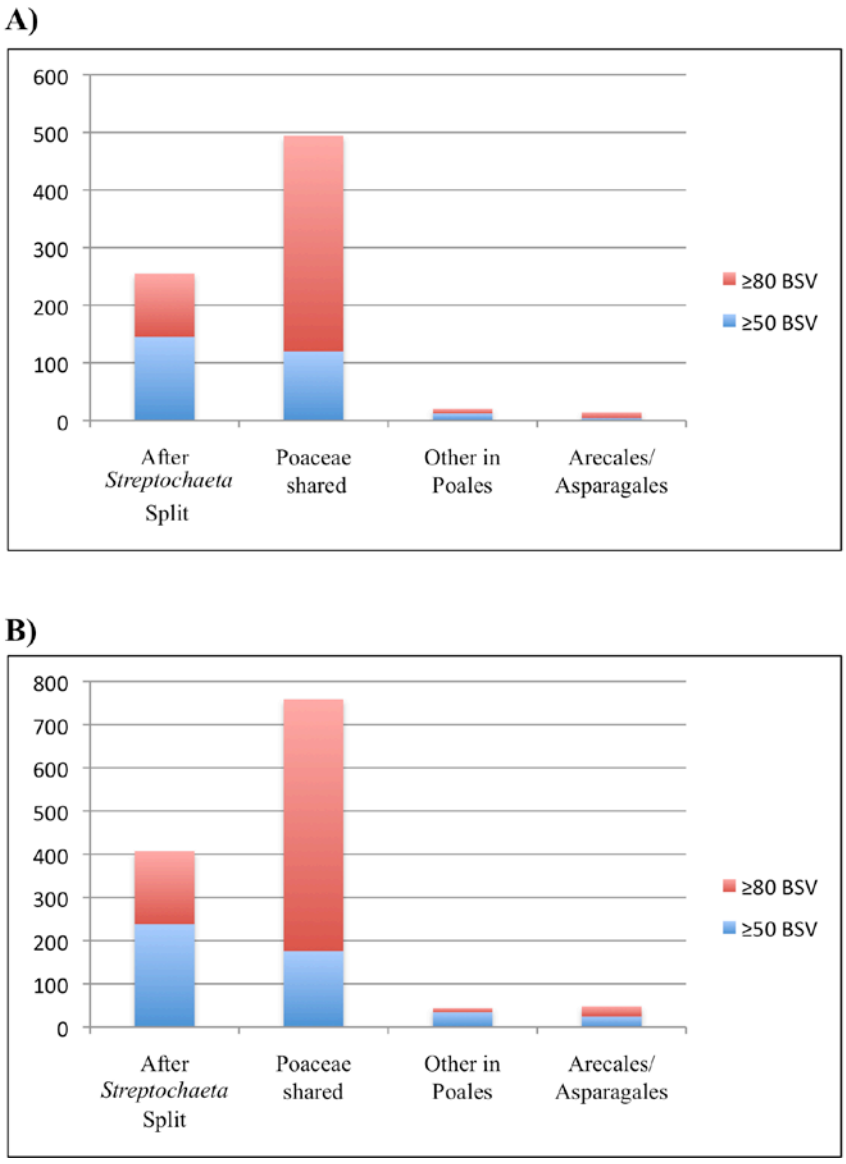

Fig. 2.1. Total counts for phylogenetic placement of duplication inferred from *rho* syntelog pairs for least common ancestor (LCA) nodes and syntelog pairs.

Total counts for LCA nodes (A) and syntelog pairs (B) support the duplication event implicated by *rho* syntenic blocks as shared by all members of Poaceae.

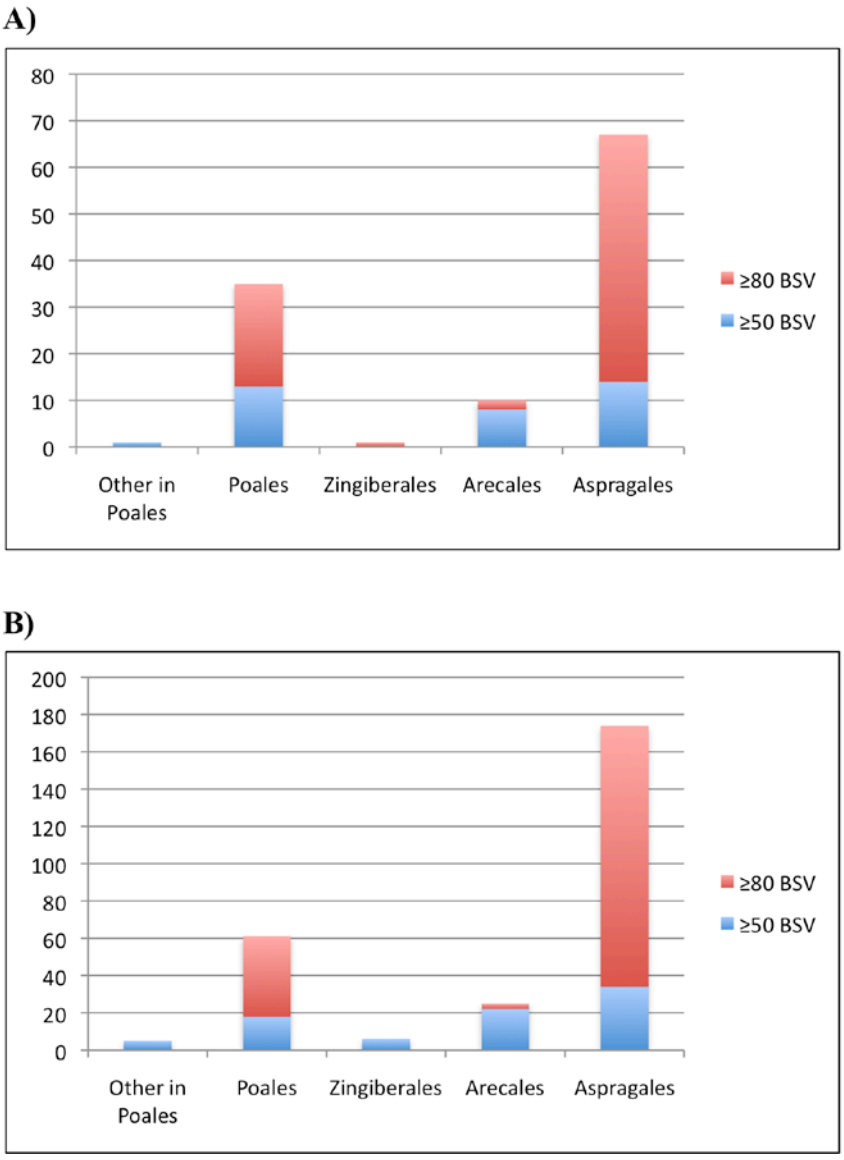

Fig. 2.2. Total counts for phylogenetic placement of duplication inferred from *sigma* syntelog pairs for least common ancestor (LCA) nodes and syntelog pairs.

Total counts for LCA nodes (A) and syntelog pairs (B) support two duplication events implicated by *sigma* syntenic blocks, one prior to the diversification of Poales and one prior to the divergence of Asparagales and the commelinids.

Table 2.1. Taxon sampling.

Taxa sampled including the data type (transcriptome or genome), sequence type (Illumina, 454, or previously assembled), and data source. *Joinvillea ascendens* and *Typha latifolia* were sequenced for this study and the OneKP project and were combined into a single assembly for each species.

Order	Clade/Grade	Family	Species	Data Type	Sequence Type	Source
Amborellales		Amborellaceae	*Amborella trichopoda*	Genome	Assembled	Phytozome v6.0
Vitales		Vitaceae	*Vitis vinifera*	Genome	Assembled	Phytozome v6.0
Ranunculales		Ranunculaceae	*Aquilegia coerulea*	Genome	Assembled	Phytozome v6.0
Asparagales		Asparagaceae	*Hosta venusta*	Transcriptome	Illumina	GenBank: SRX116252
Asparagales		Asparagaceae	*Yucca filamentosa*	Transcriptome	Illumina	OneKP
Arecales		Arecaceae	*Phoenix dactylifera*	Genome	Assembled	
Arecales		Arecaceae	*Elaeis guineensis*	Transcriptome	454	GenBank: SRX059258-SRX059263
Zingiberales		Zingiberaceae	*Zingiber officinale*	Transcriptome	Illumina	OneKP
Zingiberales		Musacaceae	*Musa acuminata*	Genome	Assembled	
Poales		Bromeliaceae	*Neoregelia carolinae* cv *argentea*	Transcriptome	Illumina	OneKP
Poales		Bromeliaceae	*Brocchinia reducta*	Transcriptome	Illumina	OneKP
Poales		Typhaceae	*Typha angustifolia*	Transcriptome	Illumina	OneKP
Poales		Typhaceae	*Typha latifolia*	Transcriptome	Illumina	OneKP/This study
Poales		Rapateaceae	*Stegolepis ferruginea*	Transcriptome	Illumina	This study
Poales	Cyperids	Cyperaceae	*Cyperus alternifolius*	Transcriptome	Illumina	This study
Poales	Cyperids	Cyperaceae	*Cyperus papyrus*	Transcriptome	Illumina	OneKP
Poales	Cyperids	Cyperaceae	*Mapania palustris*	Transcriptome	Illumina	OneKP
Poales	Cyperids	Cyperaceae	*Lepidosperma gibsonii*	Transcriptome	Illumina	OneKP
Poales	Cyperids	Juncaceae	*Juncus inflexus*	Transcriptome	Illumina	OneKP
Poales	Cyperids	Juncaceae	*Juncus effusus*	Transcriptome	Illumina	This study
Poales	Xyrids	Eriocaulaceae	*Lachnocaulon anceps*	Transcriptome	Illumina	This study
Poales	Xyrids	Mayacaceae	*Mayaca fluviatilis*	Transcriptome	Illumina	This study

Poales	Xyrids	Xyridaceae	*Xyris jupicai*	Transcriptome	Illumina	This study
Poales	Restids	Centrolepidaceae	*Centrolepis monogyna*	Transcriptome	Illumina	This study
Poales	Restids	Centrolepidaceae	*Aphelia*	Transcriptome	Illumina	This study
Poales	Restids	Restionaceae	*Chondropetalum tectorum*	Transcriptome	Illumina	OneKP
Poales	Restids	Restionaceae	*Elegia fenestrata*	Transcriptome	Illumina	This study
Poales	Graminids	Ecdeiocoleaceae	*Ecdeiocolea monostachya*	Transcriptome	Illumina	This study
Poales	Graminids	Flagellariaceae	*Flagellaria indica*	Transcriptome	Illumina	This study
Poales	Graminids	Joinvilleaceae	*Joinvillea ascendens*	Transcriptome	Illumina	OneKP/This study
Poales	Graminids	Poaceae	*Sorghum bicolor*	Genome	Assembled	Phytozome v6.0
Poales	Graminids	Poaceae	*Zea mays*	Genome	Assembled	Phytozome v8.0
Poales	Graminids	Poaceae	*Setaria italica*	Genome	Assembled	Phytozome v8.0
Poales	Graminids	Poaceae	*Oryza sativa*	Genome	Assembled	Phytozome v6.0
Poales	Graminids	Poaceae	*Brachypodium distachyon*	Genome	Assembled	Phytozome v6.0
Poales	Graminids	Poaceae	*Streptochaeta angustifolia*	Transcriptome	Illumina	This study
Poales	Graminids	Poaceae	*Aristida stricta*	Transcriptome	Illumina	OneKP
Poales	Graminids	Poaceae	*Dendrocalamus latiflorus*	Transcriptome	Illumina	GenBank: SRX156240

Table 2.2. Taxa and sources for genomes sampled for the 22-genome data set.

Order	Clade / Grade	Family	Species	Source
Funariales	Bryophyta	Funariaceae	*Physcomitrella patens*	Phytozome v.6.0
Selaginellales	Lycopod	Selaginellaceae	*Selaginella moellendorffii*	Phytozome v.6.0
Amborellales	Basal Angiosperms	Amborellaceae	*Amborella trichopoda*	EVM27
Zingiberales	Monocots	Muscaceae	*Musa acuminata*	Release 1
Arecales	Monocots	Arecaceae	*Phoenix dactylifera*	PDK30 v.3
Poales	Monocots	Poaceae	*Sorghum bicolor*	Phytozome v.6.0
Poales	Monocots	Poaceae	*Brachypodium distachyon*	Phytozome v.6.0
Poales	Monocots	Poaceae	*Oryza sativa*	Phytozome v.6.0
Ranunculales	Basal Eudicots	Ranunculaceae	*Aquilegia coerulea*	Phytozome v.6.0
Proteales	Basal Eudicots	Nelumbonaceae	*Nelumbo nucifera*	Ray Ming, unpublished
Lamiales	Asterids	Phrymaceae	*Mimulus guttatus*	Phytozome v.6.0
Solanales	Asterids	Solanaceae	*Solanum tuberosum*	PGSC_DM_v3.4
Solanales	Asterids	Solanaceae	*Solanum lycopersicum*	ITAG v.2.3
Vitales	Rosids	Vitaceae	*Vitis vinifera*	Phytozome v.6.0
Malvales	Rosids	Malvaceae	*Theobroma cacao*	CocoaGen DB
Malpighiales	Rosids	Salicaceae	*Populus trichocarpa*	Phytozome v.6.0
Fabales	Rosids	Fabaceae	*Medicago truncatula*	Phytozome v.6.0
Fabales	Rosids	Fabaceae	*Glycine max*	Phytozome v.6.0
Rosales	Rosids	Rosaceae	*Fragaria vesca*	Vescagenemodels2 v.1
Brassicales	Rosids	Caricaceae	*Carica papaya*	Phytozome v.6.0
Brassicales	Rosids	Brassicaceae	*Arabidopsis thaliana*	Phytozome v.6.0
Brassicales	Rosids	Brassicaceae	*Thellungiella parvula*	Phytozome v.6.0

Table 2.3. Contig statistics for transcriptome assembly translations for all non-genomic study species.

Species	Contig Total	Mean Contig Size (bp)	90% Contig Size (bp)
Hosta venusta	35545	885	276
Yucca filamentosa	42599	1011	327
Elaeis guineensis	30815	1455	597
Zingiber officinale	42459	855	291
Neoregelia carolinae cv *argentea*	35214	1314	429
Brocchinia reducta	37151	621	261
Typha angustifolia	30330	885	315
Typha latifolia	36886	1302	426
Stegolepis ferruginea	75055	1281	366
Cyperus alternifolius	27490	1296	435
Cyperus papyrus	27427	981	339
Mapania palustris	31905	756	279
Lepidosperma gibsonii	29872	1092	384
Juncus inflexus	27468	924	312
Juncus effusus	33059	1197	393
Lachnocaulon anceps	42801	1392	498
Mayaca fluviatilis	42133	468	219
Mayaca 9/10	38638	1401	525
Xyris jupicai	24970	1023	324
Centrolepis monogyna	26849	735	249
Aphelia	46229	438	213
Chondropetalum tectorum	36187	1098	342
Elegia fenestrata	44361	1056	300
Ecdeiocolea monostachya	33476	1194	363
Flagellaria indica	30205	810	276
Joinvillea ascendens	29855	1053	336
Streptochaeta angustifolia	33302	1035	306
Aristida stricta	43827	750	270
Dendrocalamus latiflorus	49871	825	264
Average	36758	1005	342

Table 2.4. Counts for unique syntelog pair least common ancestor (LCA) nodes and syntelog pairs for *rho* identified syntelogs with bootstrap values (BSV).

	After Streptochaeta Split		Poaceae shared		Other in Poales		Zingiberales		Arecales		Asparagales	
	LCA	Syntelog Pairs	LCA	Syntelog Pairs	LCA	Syntelog Pairs	LCA	Syntelog Pairs	LCA	Syntelog Pairs	LCA	Syntelog Pairs
≥80 BSV	113	169	373	583	6	9	0	0	1	4	10	19
≥50 BSV	151	238	122	176	20	35	0	0	2	22	2	3
<50 BSV	120	204	60	96	21	44	2	3	3	2	1	2

Table 2.5. Counts for unique syntelog pair least common ancestor (LCA) nodes and syntelog pairs for *sigma* identified syntelogs with bootstrap values (BSV).

	Other in Poales		Poales		Zingiberales		Arecales		Asparagales	
	LCA	Syntelog Pairs	LCA	Syntelog Pairs	LCA	Syntelog Pairs	LCA	Syntelog Pairs	LCA	Syntelog Pairs
≥80 BSV	0	0	20	43	0	0	2	3	53	140
≥50 BSV	3	5	12	18	1	6	6	22	15	34
<50 BSV	9	10	19	47	1	4	6	13	6	33

CHAPTER III

PLASTID PHYLOGENOMICS OF AGAVOIDEAE (ASPARAGACEAE) WITH EMPHASIS ON THE AGAVOIDEAE BIMODAL KARYOTYPE CLADE

The Asparagaceae subfamily Agavoideae (APG III, 2009; Chase et al., 2009) is a primarily rosette-forming group of petaloid monocots found throughout tropical to temperate regions. The clade has several economically important taxa including *Agave tequilana* F. A. C. Weber (tequila), *Hosta* Tratt (horticulture), *Chlorophytum* Ker. Gawl. (horticulture), and *Anemarrhena asphodeloides* Bunge (Eastern traditional medicine). The subfamily comprises ca.30 genera and over 600 species. The few species rich genera include *Agave* L. (~196 species) (The Plant List, 2010), *Echeandia* Ort. (~76 species) (Cruden, 1999), and *Yucca* L. (~50 species) (The Plant List, 2010), which have diversified in xeric habitats of Mexico and the American Southwest (Cruden, 1999; Rocha et al., 2006). Subfamily Agavoideae includes the segregated families Anemarrhenaceae, Behniaceae, Herreriaceae, Anthericaceae and Agavaceae (APG II, 2003; APG III, 2009; Chase et al., 2009). Agavoideae is divided into three major clades: i) the monotypic genus *Anemarrhena* Bunge (Bogler et al., 2006; Pires et al., 2006; Kim et al., 2010; Steele et al., 2012) is sister to the rest of the subfamily; ii) a clade including the former Behniaceae, Herreriaceae and Anthericaceae *s.s.* (Bogler et al., 2006; Chase et al., 2006; Graham et al., 2006; Pires et al., 2006; Kim et al., 2010; Seberg et al., 2012; Steele et al., 2012); and iii) the former Agavaceae *s.l.*

Agavaceae *s. l.*, as previously circumscribed by Bogler et al. (2006), is now well-defined clade within Agavoideae comprising 15 genera (over 370 spp.) (The Plant List, 2010), all possessing karyotypes with two non-overlapping chromosome size classes or bimodal karyotypes. Ten genera are characterized by the "*Yucca-Agave*" karyotype of 25S (small) + 5L

(large) chromosomes (McKelvey and Sax, 1933; Whitaker, 1934; Sato, 1935; Granick, 1944): *Hosta*, *Hesperaloe*, *Hesperoyucca*, *Yucca*, *Furcraea*, *Beschorneria*, *Polianthes*, *Manfreda*, *Prochnyanthes*, and *Agave* (Granick, 1944; Tamura, 1995). The other five genera (*Schoenolirion*, *Hesperocallis*, *Hastingsia*, *Chlorogalum*, and *Camassia*) display a variation of the 25S+5L karyotype, ranging from 12S+3L in *Chlorogalum* and *Camassia* (Gould, 1942; Sen, 1975; Fernández and Daviña, 1991; Tamura, 1995) to 18S+6L in *Hesperocallis* (Cave, 1948). The bimodal karyotype of these genera has long been considered a potential synapomorphy for the former Agavaceae clade (Gould, 1942; Granick, 1944), which recent molecular studies have confirmed (Pires et al., 2004; Bogler et al., 2006). The current APG III (2009) classification does not name this clade, but we have referred to it as the Agavoideae Bimodal Karyotype clade (ABK clade) (McKain et al., 2012).

Evolutionary relationships of the genera within the ABK clade have been difficult to discern. Similarities in karyotype (Sato, 1935; Gould, 1942), sporogenesis and embryo sac development (Cave, 1948), and serological affinities (Chupov and Kutiavina, 1981) have been used to independently group genera in the ABK clade, but other similarities, such as pollen morphology (Chung and Jones, 1989), have led phylogenies incongruent with molecular phylogenetic data (Pires et al., 2004; Bogler et al., 2006; Steele et al., 2012). Early molecular phylogenetic studies provided strong support for the ABK clade including *Hosta* and *Camassia* (Eguiarte et al., 1994; Bogler and Simpson, 1995, 1996; Eguiarte, 1995). *Camassia* has been historically allied with a few other genera: *Chlorogalum*, *Hastingsia*, *Schoenolirion*, and *Hesperocallis* (Gould, 1942; Sherman and Beckling, 1991; Fishbein et al., 2010). The placement of *Hesperocallis* has long been a point of contention, ranging from within Funkiaceae with *Hosta* (Chung and Jones, 1989) to affiliation with *Camassia*, *Chlorogalum*, and *Schoenolirion* (Gould,

1942). Pires et al. (2004) placed *Hesperocallis* in the ABK clade with high support based on plastid sequences. This finding was further supported by analyses of plastid regions, ITS, and the mitochondrial gene *atpI* (Bogler et al., 2006; Kim et al., 2010). The placement of *Hesperocallis* within the ABK clade has been inferred as basal in some studies (Bogler et al., 2006; Seberg et al., 2012) and nested within the clade with *Hosta* as the basal lineage in others (Kim et al., 2010). When sampled with *Hesperocallis, Hosta* is nested within the clade (Seberg et al., 2012), though never with high support. In studies that did not include *Hesperocallis*, *Hosta* has been placed basal to all other sampled taxa in the ABK clade with high support (Bogler and Simpson, 1996; Steele et al., 2012).

Camassia and *Chlorogalum* are sister taxa with high support in analyses of plastid loci (Bogler et al., 2006; Smith et al., 2008), and, recently, *Hastingsia* has been shown to be a part of this clade sister to *Camassia* (Fishbein et al., 2010; Halpin, 2011). Recent work by Halpin (2011), however, did not place *Schoenolirion* in the *Camassia*/*Chlorogalum*/*Hastingsia* clade, but as sister to *Hesperaloe* in a well-supported clade with *Hesperoyucca* (Halpin, 2011). Another well-defined group within the ABK clade is a subclade comprising *Beschorneria*, *Furcraea*, *Polianthes*, *Manfreda*, *Prochnyanthes*, and *Agave* (Bogler and Simpson, 1996; Bogler et al., 2006; Good-Avila et al., 2006; Seberg et al., 2012; Steele et al., 2012) although the relationships of these genera within this clade are not highly supported. Relationships among these three clades and the remaining genera, *Hosta*, *Hesperocallis*, and *Yucca*, have remained unclear (Bogler et al., 2006; Good-Avila et al., 2006; Seberg et al., 2012).

Understanding the intergeneric relationships within the ABK clade is a key step in understanding the evolution of karyotypic variation, pollination systems, and other ecological characters within this clade. A recent transcriptome-based phylogenomic study detected a whole

genome duplication (WGD) event prior to diversification of the ABK clade (McKain et al., 2012). This WGD potentially led to the formation of the "*Yucca-Agave*" bimodal karyotype, though the study did not test a causative link between these two events. Variation in the karyotype of *Camassia*, *Chlorogalum*, *Hastingsia*, *Hesperocallis*, and *Schoenolirion* is of interest, as it varies from the otherwise stable of the "*Yucca-Agave*" karyotype. The yucca-yucca moth pollination mutualism is an intriguing aspect of the reproductive biology of *Hesperoyucca* and *Yucca* species. This relationship, a textbook example of coevolution, is unique due to the specialized morphology and behavior of the moths and plants and shared diversification of each group (Riley, 1872; Pellmyr, 2003).

This study utilizes high-throughput sequencing technology to sequence entire chloroplast genomes for representative species across the ABK clade and Asparagaceae to resolve relationships among the genera of the ABK clade. Large datasets of chloroplast protein coding genes have been used to resolve difficult angiosperm phylogenies in very old lineages (Leebens-Mack et al., 2005; Moore et al., 2010) and in younger, highly diversified lineages (Givnish et al., 2010; Xi et al., 2012). Here, we extend these large-scale approaches to include non-coding regions, effectively utilizing the entire chloroplast genome to provide a fully resolved set of intergeneric relationships for the ABK clade. Additionally, we combined data from this study with that of Steele et al. (2012) and Givnish et al. (2010) to estimate the timing of diversification in the ABK clade. Character evolution within the ABK clade is considered within the context of the estimated phylogeny and divergence times.

Materials and Methods

Taxon sampling

Species samples were selected based on intergeneric relationships found in previous analyses; sixteen species from the ABK clade, two species from the former Anthericaceae *s. l.*, *Behnia reticulata*, *Anemarrhena asphodeloides*, two species from the subfamily Scilloideae (Asparagaceae) and *Nolina atopocarpa*, serving as an outgroup. Table 3.1 is a complete list of taxa including collection information.

Chloroplast isolation

Chloroplast isolation was performed on a subset of the species in this study (Table 3.1). Leaf material was collected from either wild or cultivated specimens and kept at 4°C until isolations were made. Chloroplasts were isolated using the protocol of Jansen et al. (2005) and stored at -20°C.

Chloroplast isolations were lysed and prepared for whole genome amplification using the protocol for cells provided with the REPLI-g Midi Kit (QIAGEN, Germantown, Maryland, USA) with some modification. A total of 1.0 µL of chloroplast isolation was used for each lysis reaction instead of the recommended 0.5 µL. Each amplification reaction was 25.0 µL with 2.5 µL of lysis reaction, 7.75 µL of nuclease-free water, 14.5 µL reaction buffer and 0.25 µL DNA polymerase. Reactions were run overnight at 30°C (~16-18 hours) and heat inactivated at 65°C for 3 minutes.

The resulting amplification was precipitated by mixing with 0.10 volumes of 3M sodium acetate and 2.0 volumes of 100% ethanol and spinning at 14000 rpm at 4°C for 10 minutes. The supernatant was poured off and the remaining pellet was cleaned with 70% ethanol and spun

again for 5 minutes. The supernatant was removed, the pellet dried, and then suspended in 100 µL nuclease-free water.

Assessment of plastid DNA abundance

The percentage plastid DNA in REPLI-g rolling circle amplifications was assessed using quantitative PCR (qPCR). A standard curve for critical threshold value (Ct) and percent chloroplast DNA (%cp) was estimated using a standard *Asparagus officinalis* plastome identified in a bacterial artificial chromosome (BAC) library constructed at the Arizona Genomics Institute (BAC library AO_Ba). DNA was isolated from a single BAC clone that included the complete 150 kilobase (kb) *Asparagus* plastid genome. Known concentrations (2.5 – 0.0025 ng/ul) of the isolated DNA were used in rtPCR amplifications of a 150 base pair portion of the *rbcL* gene in order to construct a standard curve for the number of cycles required for amplification above rtPCR detection threshold levels (i.e. Ct values). The standard curve was used to interpret Ct values for rtPCR experiments performed on REPLI-g rolling circle amplicons. Amplicons with at least 5% plastome DNA were sequenced.

DNA isolation

DNA was isolated from all samples for either PCR reactions for gap sequencing of finished chloroplast genomes or for direct whole genome shotgun sequencing. Samples for PCR were isolated using a modified CTAB DNA extraction protocol (Doyle and Doyle, 1987), starting with 2-3 grams of fresh or 0.25-0.5 grams of dried leaf tissue. Material was crushed in liquid nitrogen and remained frozen until CTAB was added. Samples for direct sequencing were isolated using the DNeasy Plant DNA Extraction Kit (QIAGEN, Germantown, Maryland, USA).

Sequencing

Sequencing for plastid isolation samples was conducted using an Illumina GAIIx Genome Analyzer (Illumina, San Diego, California, USA), six samples per lane, at Cold Spring Harbor Laboratory.

For samples sequenced from total DNA extractions, sequencing libraries were made using NEB Prep kit E6000L (New England Biolabs, Inc., Ipswich, Massachusetts, USA) and the following the protocol of Steele et al. (Steele et al., 2012). Sample libraries were sequenced on a GAIIx Genome Analzyer (Illumina, San Diego, California, USA), six samples per lane, or a HiSeq sequencer using v1 TruSeq chemistry, 8-12 samples per lane, at the DNA Core Facility at the University of Missouri-Columbia.

Chloroplast genome assembly

Sequence data from Illumina reactions was processed for assembly by removing identification tags. Table 3.1 summarizes total reads, number of chloroplast reads and read length.

The programs YASRA v.1.01.00 (Ratan, 2009) and Velvet v.1.0.09 (Zerbino and Birney, 2008) were used to assemble chloroplast genomes. YASRA is a referenced-based assembler that builds an assembly by layering reads on an existing sequence. The chloroplast genome sequence of the most closely related species available to the species being assembled was used as the reference. As the assemblies progressed, the recently assembled chloroplast genomes from this project were used as a reference instead of those publicly available. Velvet was used to make *de novo* assemblies and optimized with VelvetOptimiser v.2.1.7 (Gladman 2009).

Sequencher (Genecodes, Ann Arbor, MI, USA) was used to assemble contigs from both YASRA and Velvet. For a contig to be considered in the final assembly, we required both programs to have a contig covering a region. If the contigs were incongruent, the reads were searched for the region and assembled with the contigs. The base or arrangement that was best represented in the reads was incorporated into the assembly. For regions where there was only one contig, the reads were searched starting where more than one contig was found and assembled across the single contig until complete coverage. Contigs were identified for a particular region of the chloroplast genome using DOGMA (Wyman et al., 2004). If two or more contigs from the same region could not be assembled, the reads were searched for the area at the two ends of the closest contigs. These were assembled onto the contigs and repeated until the gap between the contigs was closed with more than 100 bases overlapping. This was repeated until an entire chloroplast genome was assembled. For gaps that could not be filled, N's were used as spacers to join the assembled contigs.

Final plastid genome sequences were analyzed using a Perl script (available from M. R. McKain upon request) that used the chloroplast sequence to search reads for 20 bp sequences using a sliding window and returned the total coverage (for an exact match) of the 20 bp sequence. Coverage thresholds, usually half the total coverage, were set and values less than the threshold were identified along with the position of the start of the 20 bp sequence in the genome. Each identified position was further investigated by searching and assembling reads matching the focal region within the chloroplast genome. Changes to chloroplast genome assemblies using this technique better represented reads, including base changes and the resolution of indels.

Filling gaps using PCR

Chloroplast genomes were used to design primers to complete chloroplast assembly for those species with gaps. Table 3.2 is a list of primers, regions and species.. Each PCR reaction used 20ng of DNA, 0.5 µL of 10mM primers, 1.0 µL of 25mM DNTPs, 1.0 µL of 10mM $MgCl_2$, and 0.2 µL of Taq. Reactions were run for 3 minutes at 94.0°C, then 35 cycles of 94.0°C for 30 seconds, 50.0°C for 45 seconds, 72.0°C for 2 minutes and a final extension at 72.0°C for 10 minutes. An aliquot of 8 µL of PCR product was added to 0.75 µL of Antarctic phosphatase (New England BioLabs, Ipswich, MA, USA) and 0.25 µL exonuclease I (New England BioLabs, Ipswich, MA, USA) and incubated for 60 minutes at 37.0°C followed by inactivation for 25 minutes at 80.0°C. A sequencing reaction for each sample used 2.0 µL of digested PCR product, 0.5 µL BigDye (BigDye Terminator v3.1, Applied Biosystems, Carlsbad, CA, USA), 2.0 µL 5x sequencing buffer, 0.6 µL primers (10mM) and 6.9 µL nuclease-free water. Reactions were run for 2 minutes at 94.0°C, then 35 cycles of 94.0°C for 20 seconds, 50.0°C for 10 seconds, 60.0°C for 90 seconds and a final extension at 60.0°C for 8 minutes. 5.0 µL of nuclease-free water was added to each sample after the reaction was complete.

Sequencing products were cleaned using a Sephadex protocol. Five µL of ddH_2O was added for each 2 µL of each PCR product. Sephadex G-50 was added to a Multiscreen HV Plate (Millipore, Billerica, MA, USA) and 300 µL of ddH_2O was added to each well. After 40 minutes, the Multiscreen HV Plate was spun at 910 relative centrifugal force (rfc) for three minutes, rotated 180°, spun again, and repeated. Sequencing reaction samples were added to the Sephadex column and spun into a clean 96-well plate. Samples were sequenced on a 3730xl 96-capillary DNA Analyzer (Applied Biosystems, Carlsbad, CA, USA) at the University of

Georgia's Genomics Facility at the University of Georgia. Sequences were added to existing plastome assemblies using Sequencher.

Alignment of plastomes

Plastome assemblies were annotated using DOGMA (Wyman et al., 2004). Protein coding regions, rRNAs, tRNAs, introns, and intergenic regions were annotated and extracted from DOGMA. Each region was added to the MonAToL Plastid Gene Database (http://jlmwiki.plantbio.uga.edu/PlastidDB/). The partitions were individually aligned for all 23 species using MUSCLE v3.7 (Edgar, 2004). Protein coding region translations were extracted from DOGMA, and these were aligned using MUSCLE v3.7. Nucleotide sequences were aligned to the amino acids alignments using PAL2NAL v13 (Suyama et al., 2006).

When chloroplast genomes had missing genes, the regions were aligned with those from the closest relative that was not missing the gene. Intergenic regions were aligned, annotated, and added to the database.

Phylogenetic analyses

Alignments were concatenated into three data sets based on the position of the region in the plastome: single copy, inverted repeat and full plastome. These concatenations resulted in 115669, 26817, and 142486 column supermatrixes for the single copy, inverted repeat, and full plastome regions, respectively. Maximum likelihood analyses were conducted using RAxML v7.3.0 [8] using the GTR + gamma evolutionary model, 500 bootstrap replicates, and partitioned into protein coding regions, tRNAs, rRNAs, introns, and intergenic regions.

Bayesian analyses were coducted using Mr.Bayes v3.2.1 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) in parallel (Altekar et al., 2004). Data was partioned by region as with the RAxML analyses. Among-site rate variation was modeled using a gamma distribution with 8 rate categories. For each data set, two independent runs of 5 million generations were performed with four chains, three hot and one cold, and sampled every 500 generations. A burnin fraction of 25% was used on the sampled trees. AWTY (Nylander et al., 2008) was used to confirm that the independent runs reached convergence.

Divergence time estimation

A concatentated data set of protein coding genes was assembled from new data, Steele et al. (2012), Givnish et al. (2010), and a selected set of whole plastomes downloaded from GenBank. Genes and species were chosen to optimize the largest portion of the genome with the largest amount of taxonomic variation. Ultimately, a set of 69 protein coding genes (*atpA*, *atpB*, *atpE*, *atpF*, *atpH*, *atpI*, *ccsA*, *cemA*, *clpP*, *matK*, *ndhA*, *ndhB*, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhI*, *ndhJ*, *petA*, *petB*, *petD*, *petG*, *petL*, *petN*, *psaA*, *psaB*, *psaI*, *psaJ*, *psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*, *rbcL*, *rpl14*, *rpl16*, *rpl12*, *rpl20*, *rpl22*, *rpl23*, *rpl36*, *rpoA*, *rpoB*, *rpoC1*, *rpoC2*, *rps11*, *rps12*, *rps14*, *rps15*, *rps18*, *rps2*, *rps3*, *rps4*, *rps7*, *rps8*, *ycf1*, *ycf2*, *ycf3*, *ycf4*) from 67 taxa was selected. Table 3.3 summarizes taxa, source, and missing genes.

Individual protein coding genes were translated and aligned using MUSCLE v3.7. Nucleotide sequences were then aligned on the amino acids using PAL2NAL. These sequences were concatentated into a single supermatrix with 68730 columns. A maximum likelihood

phylogeny of this data set was estimated using RAxML v7.3.0 using the GTR + gamma evolutionary model and 500 bootstrap replicates.

Divergence times were estimated using BEAST v1.7.2 (Drummond et al., 2012). BEAUti v1.6.2 was used to create the BEAST input XML file. Taxa were grouped according to the maximum likelihood tree estimated from the data (Figure 3.1). Groups were defined as nodes with bootstrap values of 100. Monophyly was enforced on these groups during the BEAST runs. A GTR + gamma evolutionary model was used with 10 gamma rate categories. The data was partitioned into the three codon positions with substition rate, rate heterogeneity, and base frequencies unlinked across the partitions. An uncorrelated lognormal, relaxed clock model (Drummond et al., 2006) was estimated for the data set. The maximum likelihood tree estimated from the data was used as the starting tree with speciation modeled by the yule process. Age calibrations were imposed for Alismatales (112 million years; Friis et al., 2010), Arecaceae (93 million years; Harley, 2006), and *Yucca* (14 million years; Tidwell and Parker, 1990) using lognormal age distributions and an estimated age calibration for Asparagales (133.11 million years; Magallón and Castillo, 2009) using a normal age distribution. Nineteen separate analyses for up to 50 million generations were run, sampling every 1000 generations. Convergence was assessed using Tracer v1.5, and runs were continued until the effective sample sizes (ESS) were greater than 200. Runs were combined using LogCombiner v1.6.2 (part of the BEAST package) with a 10% sample burn in. Maximum clade credibility trees with mean node heights were created using TreeAnnotator v1.6.2 (part of the BEAST package).

Results

Chloroplast genomes

Full chloroplast genomes were assembled for all species. *Albuca* cf. *kirkii*, *Behnia reticulata* and *Chlorophytum rhizopendulum* had some regions that could not be fully assembled and the missing likely less than 100 base pairs (bp). *Behnia* has a more substantial set of missing regions since plant material was not available for PCR-based sequencing. These missing regions did not include genes, tRNAs, or rRNAs.

For species in the ABK clade, plastome size ranged from 155,453 bp in *Camassia scilloides* to 158,028 bp in *Yucca brevifolia* (Table 3.1). Plastome size for other species were within this range except for *Chlorophytum* and *Echeandia*, which had plastome sizes of 153,594 and 154,356, respectively.

There were no major rearrangements in the sequenced plastomes relative to the most common angiosperm chloroplast genome gene order (e.g. Guisinger et al., 2010) although there are some differences. Gene losses were evident in five species. *Beschorneria septentrionalis* and *Hosta ventricosa* were missing exon 2 of the *rps16* gene. *Camassia*, *Chlorophytum* and *Echeandia* were missing *rps19* completely. *rps19* was a pseudogene in *Behnia reticulata, Chlorogalum pomeridianum*, *Hesperaloe campanulata*, *Hesperaloe parviflora*, *Schoenolirion croceum*, and *Hosta ventricosa*. For many taxa in the Asparagales, *rps19* is found in the inverted repeat (IR) regions of the chloroplast genome (Wang et al., 2008) as found in most sample taxa. The gene is not found in the IR of *Chlorogalum pomderidianum* and the two *Hesperaloe* species. Other pseudogenes identified included *infA* for *Albuca*, *Behnia*, *Camassia*, *Chlorophytum*, *Echeandia*, *Hesperocallis, Yucca brevifolia*, *Y. filamentosa*, *Y. queretaroensis* and *Y. schidigera* and *ndhK* for *Albuca, Nolina atopocarpa*, *Polianthes* sp., *Yucca*

queretaroensis, and Y. schidigera. The *cemA* gene did not contain a start codon either through the traditional ATG codon or the alternative AGG codon with RNA editing in *Chlorogalum, Chlorophytum, Echeandia, Yucca filamentosa, Y. queretaroensis* and *Y. schidigera* (Table 3.1).

Chloroplast phylogenies

Alignments were made on each annotated region of the chloroplast genome separately and then concatenated into three groups: "Full Plastome" (FP), comprising the full chloroplast genome but only one copy of the inverted repeat; "Inverted Repeat" (IR), comprising the inverted repeat region; and "Single Copy Region" (SCR), comprising the large single copy and small single copy regions. The maximum likelihood (Fig 3.2) and Bayesian (Fig 3.3) phylogenetic reconstructions for all three regions depict variation in relative evolutionary rates of these regions. The IR is shown on a scale that is a quarter that of the scale shown for the FP and SCR. The relative branch lengths of the IR region compared to the SCP is ~20%. These shorter branch lengths, and corresponding lower evolutionary rates, support questionable relationships depicted in the IR trees with lower support values and polytomies.

The FP and SCR trees fully resolved relationships within the ABK clade with high bootstrap and posterior probability support for all nodes except those within *Agave s.l.* (*Agave*, *Manfreda*, and *Polianthes*) and *Yucca*. Based on this phylogeny, the basal member of the ABK clade is *Hosta* [bootstrap value (ML BSV) of 100 and posterior probability (BI PP) of 1.0] with the rest of the family split into two clades. This relationship is supported by the IR trees (ML BSV 99; BI PP 0.97). One clade includes a *Hesperaloe-Hesperoyucca-Schoenolirion* subclade, previously recovered with strong support (Halpin, 2011), and a *Camassia-Chlorogalum-Hesperocallis* subclade (Fig. 3.2). *Schoenolirion* is sister to *Hesperaloe* (ML BSV 100; BI PP

1.0 in FP and SCR trees). *Hesperocallis* is highly supported (ML BSV 100; BI PP 1.0 in FP and SCR trees) as sister to a clade containing *Camassia* and *Chlorogalum*. The other major clade in Agavaceae includes *Yucca* as sister to the rest of the clade including *Beschorneria* and *Agave* s. l. The relationships among *Agave, Manfreda* and *Polianthes* are tentative, with *Agave* sister to *Manfreda* in the FP and SCR trees (support values 86 ML BSV; 1.0 BI PP). The IR tree, however, suggests that *Polianthes* and *Agave* are sister taxa (ML BSV 91) and is represented as a polytomy in the Bayesian tree. Conflicts among these alternative topologies are most likely attributable to the very short branches leading to these three species depicted in all trees. *Agave*, *Manfreda*, and *Polianthes* form a well-defined group (ML BSV 100; BI PP 1.0) in all trees. Within *Yucca*, *Y. queretaroensis* is sister to *Y. schidigera* with high support (ML BSV 100; BI PP 1.0 in the FP and SCR trees; ML BSV 91; BI PP 1.0 in the IR tree), which differs from Pellmyr et al.'s (2007) placement of *Y. queretaroensis* as the basal lineage for *Yucca*. Smith et al. (2008) included the *Y. quertaroensis* sample from Pellymyr et al. (2007) and added a new individual. They found that the sample from Pellymyr et al. remained basal in analyses but the new sample was nested within the fleshy-fruited yuccas, a result that agrees with the current study. Whether these results are due to introgression (a known phenomenon in yuccas; Leebens-Mack et al., 1998; Rentsch and Leebens-Mack, 2012) or retained ancestral polymorphism is unknown. The relative placement of *Yucca brevifolia*, *Y. filamentosa*, and the *Y. schidigera*-*Y. queretaroensis* clade is not clear. Low ML BSV (e.g. 46) for these relationships demonstrate a lack of variation in the plasid genomes that could be attributed to rapid diversification of the clade.

Outside of the ABK clade, the FP tree places *Anemarrhena* as sister to the rest of Agavoideae (ML BSV 100 and BI PP 1.0 in all trees). *Behnia* is sister to a clade including

Chlorophytum rhizopendulum and *Echeandia* sp. (ML BSV 100 and BI PP 1.0 in FP and SCR trees). These relationships have both been previously recovered with high support (Bogler et al., 2006; Kim et al., 2010; Seberg et al., 2012).

The topology of the FP tree is congruent with that of the SCR tree (Fig. 3.2 and Fig. 3.3). The ML BSV and BI PP are equivalent between the two trees and the branch lengths are relatively similar (Fig. 3.2 and Fig. 3.3). The topology of the IR tree differs in the lack of resolution of *Hosta*, the *Yucca-Beschorneria-Agave s.l.* clade (YBA clade), and the *Hesperoyucca-Hesperaloe-Schoenolirion-Hesperocallis-Camassia-Chlorogalum* clade (HHS-HCC clade). There is also a difference in the relationship of *Agave s. l.* where *Polianthes* is more closely related to *Agave*. Other relationships in the IR tree are the same in the FP and SCR trees but with much lower support values at a number of nodes subtended by very short branches.

Divergence time estimates

Due to a lack of calibration points for the ABK clade, the sampling was extended to include a large sampling of Asparagaceae, Asparagales, and other monocot lineages. The resulting chronogram from the BEAST analysis places all events within Agavoideae ranging from 79.25 to 1.10 mya. Table 3.4 is an overview of the timing of events within Agavoideae. Many of the mean ages are slightly older than previous estimates for Agavoideae (Good-Avila et al., 2006; Smith et al., 2008), but the distributions of these estimates are quite large. Mean ages for older nodes, such as the origin of monocots are also much older than previous estimates but age distributions overlap with previous studies (Magallón and Castillo, 2009). The Agavoideae crown group was estimated to have diversified 63.00 mya within the Paleocene (95% highest

probability density [HPD]: 79.25-47.27 mya [Upper Cretaceous to middle Eocene]). The divergence of the ABK clade from *Behnia*, *Chlorophytum*, and *Echeandia* is estimated at 43.06 mya (middle Eocene). Within the *Behnia*-*Chlorophyum*-*Echeandia* clade, *Behnia* first branched off 35.28 mya (late Eocene), and the split between *Chlorophytum* and *Echeandia* occurred 13.84 mya (middle Miocene).

Hosta, as the basal lineage of the ABK clade, diverged from the rest of the clade 29.39 mya (middle Oligocene; 95% HPD: 40.12-19.97 mya; middle Eocene to early Miocene). This distribution overlaps with previously estimated ages, although they were relatively younger (25.8 ± 3.4 mya; Good-Avila et al., 2006 and 23.96 ± 9.66 mya; Smith et al., 2008). The *Yucca*/*Beschorneria*/*Agave s.l.* clade (YBA clade) and *Hesperoyucca*/*Hesperaloe*/*Schoenolirion* (HHS) - *Hesperocallis*/*Chlorogalum*/*Camassia* (HCC) clades diverged 27.03 mya (late Oligocene). Within the HHS-HCC clade, the HHS and HCC clades split 22.57 mya (early Miocene). Further diversification of the HHS clade occurred 15.98 mya (middle Miocene) with the splitting of *Hesperoyucca* and the divergence of *Schoenolirion* and *Hesperaloe* 7.53 mya (late Miocene). Major events within the HCC clade occurred with the divergence of *Hesperocallis* 178.31 mya (early Miocene) and the divergence of *Camassia* and *Chlorogalum* 11.01 mya (late Miocene).

The YBA clade is the most species-rich clade within the ABK clade, containing *Yucca* (~50 species) and *Agave* (~196 species). *Yucca* split from the rest of the clade 20.22 mya (early Miocene; 95% HPD: 28.17-13.84 mya; middle Oligocene to middle Miocene). The *Yucca* crown group began to diversify 12.59 mya (middle Oligocene) that is within the estimated range from Smith et al. (2008; 10.19 +/- 6.49 mya). *Beschorneria* split from *Agave s. l.* 11.77 mya (middle Miocene). *Manfreda*, *Polianthes*, and *Agave* diversified 3.30 mya (Pliocene). This date is much

younger than previously estimated dates (8.3 ± 2.4 mya; Good-Avila et al., 2006), although sampling for the Good-Avila et al. (2006) study was much more thorough for *Agave s. l.* and may better represent the evolutionary diversity of the group.

Discussion

Chloroplast genome evolution

The structure of angiosperm chloroplast genomes has remained relatively stable since the origin of the lineage. Most angiosperm chloroplast genomes contain a large single copy region, a small single copy region and two inverted repeat regions (IR_A and IR_B) that separate the large and small single copy regions (Palmer, 1985). Exceptions to this are in subfamily Faboideae (Fabacaceae), where one copy of the IR is lost (Lavin et al., 1990). This phenomenon has also been seen in gymnosperms with the differential loss of IR repeats in Pinaceae and Cupressaceae (Wu et al., 2011). Smaller changes in the structure and gene content of chloroplast genomes is more common, for example, Poaceae which have undergone multiple inversions and loss or pseudogenization of genes (Doyle et al., 1992; Maier et al., 1995). Steele et al. (2012) found that five genes (*clpP*, *ndhF*, *rpl32*, *rps16*, and *rps19*) were missing from various taxa throughout Asparagales. Additionally, they showed that the loss of *rpl32* was evident in all sampled members of subfamily Asphodeloideae (Xanthorrhoeaceae), demonstrating that these shared losses are the result of common ancestry rather than homoplasy.

Plastid genomes of the ABK clade exhibit gene loss in two genes (*rps19* and *rps16*, exon 2). The *rps16* exon was independently lost in *Hosta* and *Beschoernia*. Multiple independent losses for this gene have occurred in Fabaceae (Doyle et al., 1995), *Dioscorea* (Hansen et al., 2007), *Populus* (Okumura et al., 2006), and a number of Asparagaceae taxa (Steele et al., 2012).

Ueda et al. (2008) have shown that in *rps16* loss in *Medicago* and *Populus*, an *rps16* homolog from the mitochondrion targets the chloroplast, acting in its stead. This dual targeting of the mitochondrial *rps16* occurs in eudicots and monocots, even when a functional chloroplast *rps16* was present. As more angiosperm chloroplast genomes are sequenced, it may become evident that *rps16* is lost frequently without consequence due to the redundancy of function in the mitochondrial *rps16*.

Pseudogenization of *infA* and *ndhK* was evident in the ABK plastomes. The *infA* gene was independently pseudogenized in *Hesperocallis*, *Camassia*, and in all sampled *Yucca* species. There was an independent pseudogenization of *ndhK* in *Polianthes* and in the *Yucca schidigera*/*Y. queretaroensis* clade (Fig 3.2). The occurrence of shared pseudogenizations in lineages, as well as gene losses and stuctural changes, could be used to support relationships.

The *rps19* gene is found in different forms across Agavoideae plastomes. Prior to the divergence of Asparagles, the IR boundary expanded to include *rps19* (Wang et al., 2008), which apparently is ancestral state for Agavoideae. A trend within Agavoideae is pseudogenization leading to eventual loss of the gene in the lineage (Table 3.2), found in *Behnia* (pseudogenization) and *Chlorophytum* and *Echeandia* (loss). A potential step-wise scenario occurs in the HCC clade where *Hesperocallis* has a pseudogenized version of the gene that is found in the IR. *Chlorogalum*, sister to *Camassia* + *Hastingsia* (Halpin, 2011) in a clade with those genera sister to *Hesperocallis* (Fig. 3.2), maintains a pseudogenized version of *rps19* although it has been removed from the IR. As the last step, *Camassia* completely lacks *rps19*. Further taxon sampling along the lineage leading to *Chlorophytum* + *Echeandia* (after the divergence of *Behnia*), could show examples of intermediary steps, such as the removal of the pseudogenized *rps19* from the IR. Another example of this evolutionary loss may occur in the

HHS clade, where *Schoenolirion*, sister to *Hesperaloe* (Fig. 3.2), shares a pseudogenized *rps19* with *Hesperaloe*. *Schoenolirion* has *rps19* in the IR, while *Hesperaloe* (the two sampled species) does not. This pattern suggests a maintainance of *rps19* within the IR while it is a functioning gene, possibly due to dosage compensation from the two copies, but when the gene is pseudogenized [and function taken over by a paralog as in *rps16* (Ueda et al., 2008)], the IR boundary becomes more fluid. Further investigation into this phenomenon through deeper sampling of the *Behnia* + *Chlorophytum* + *Echeandia* clade and *Hesperaloe* may lead to more examples of this step-wise loss. Sampling within *Hesperaloe* populations may show variation in the presence and loss of *rps19*. Additionally, expression studies of functioning IR *rps19* and comparison to potential *rps19* paralogs from the nuclear genome would help to elucidate if the IR boundary undergoes a release from purifying selection as maintaining two copies of *rps19* is not required.

Implications of chloroplast phylogeny and divergence time estimates

Interpreting evolutionary history within the ABK clade has been difficult due to lack of resolution of intergeneric relationships (Eguiarte et al., 1994; Bogler and Simpson, 1995; Eguiarte, 1995; Bogler et al., 2006; Good-Avila et al., 2006; Pires et al., 2006). The chloroplast phylogeny presented here suggests that the difficulty in resolution is due rapid diversification of the major clades or a slow rate of molecular evolution during diversification, as depicted by short branch lengths of the internodes (Fig. 3.2 and Fig. 3.3). McKain et al. (2012) showed a whole genome duplication event prior to the diversificaiton of the ABK clade but after divergence from *Behnia* + *Chlorophtyum* + *Echeandia*. The rapid diversification may be a consequence of this event, either through a special case of the Bateson-Dobzhansky-Mueller speciation model (Orr,

1996) related to differential loss of paralogs or genomic plasticity (e.g. wheat; Dubcovsky and Dvorak, 2007) allowing for rapid adaptation to arid habitats and novel pollinator interactions (Good-Avila et al., 2006).

The placement of *Hosta* as basal to the rest of the ABK clade suggests that the ancestral state of the group was mesophytic, lacking the characteristic xeric adaptations found thoughout the clade. *Hosta* is the only member of the ABK clade not native to the Western hemisphere (Rocha et al., 2006), sharing an eastern Asian range with that of *Anemarrhena*. This supports the hypothesis that arid adaptation led to diversificaiton of the ABK clade as the ancestral state of the two major clades (YBA and HHS-HCC) is parsimoniously a xerophyte. This diversification occurred over a relatively short amount of time (~6 my; Table 3.4, Fig. 3.4). *Hosta*'s basal position also suggests that the "*Yucca-Agave*" bimodal karyotype of 25S + 5L chromosomes is the ancestral karyotype for the ABK clade.

Variation from the "*Yucca-Agave*" karyotype occurs in two clades nested within the HHS-HCC clade. In both instances, there is a transition from a large, semi-succulent to succulent habit to a much smaller, herbaceous and mesically adapted habit. In both instances, these transitions are accompanied by a shift in range from the American Southwest to either the Pacific Northwest (HCC clade) or to the Eastern coastal plain (HHS clade). *Schoenolirion croceum* has a karyotype of n = 12, with 5L and 7S chromsomes (personal observation,; Sherman, 1969). Its sister genus, *Hesperaloe*, exhibits the "*Yucca-Agave*" karyotype (Whitaker, 1934) as does *Hesperoyucca*, suggesting a reduction from the ancestral karyotype in *Schoenolirion*. *Schoenolirion croecum* and *S. wrightii* grow on granite outcrops throughout the Alabama plateau, lack leaf thickening or succulence other than fleshy leaf bases, and are spring ephemerals, avoiding the hot, dry summers of the granite outcrop (eFloras, 2008). *Schoenolirion*

albiflorum, the remaining species in *Schoenolirion*, is found in the marshy pinelands, cypress bogs, and wet savannahs of Georgia and Florida and lacks a fleshy leaf base. The sister clade to SHH, comprises *Hesperocallis*, *Chlorogalum*, *Camassia* (all sampled in this study), and *Hastingsia* (Fig. 3.2, (Halpin, 2011)), all have bimodal karyotypes but with lower chromosome numbers than the ancestral ABK karyotype. *Hesperocalllis* maintains arid adaptations such as coriaceous leaves and a fleshy bulb and roots, is the basal lineage of this clade, has a karyotype of n = 18S + 6L (Cave, 1948), and occurs in the Sonoran and Mojave deserts of Arizona, California, and Nevada. The next genus in the lineage is *Chlorogalum*, which has karyotypes ranging from n = 25S + 5L to n = 12S + 3L (Cave, 1970). *Chlorogalum* species are dispersed in mostly dry, open areas of California and Oregon and have coriaceous leaves and fleshy leaf bases (eFloras, 2008). *Camassia* and *Hastingsia* are sister genera (Halpin, 2011) with bimodal karytoypes of n = 12S + 3L and n = 5-6L + 21-20S, respectively. *Hastingsia* is found in wet seepage areas of serpentine outcrops in northern California and southern Oregon and lacks the coriaceous leaves of *Chlorogalum* and *Hesperocallis* (eFloras, 2008). *Camassia* is found throughout the northwest and eastern United States in wet meadows, forest understories, and seepage areas of granite outcrops. Like *Hastingsia*, *Camassia* species lack coriaceous leaves. In both the SHH and HCC clades, the trend is towards smaller herbaceous plants lacking the desert adaptations of their predecessors. This change in lifestyle follows a movement out of the center of diversity for *Yucca* (Rocha et al., 2006) and to more mesic habitats. The karyotypic variation that also follows this trend could be the result of higher rates of molecular evolution associated with the change to a herbaceous habit (Smith and Donoghue, 2008).

Yucca and *Hesperoyucca* species are actively pollinated by moths in the Prodoxidae genera *Tegeticula* and *Parategiticula*. Female moths have a unique prehensile tentacle-like

structure used to actively collect and distribute pollen from *Yucca* and *Hesperoyucca* (Pellmyr, 2003). Pellmyr and Leebens-Mack (1999) estimated that the colonization of yuccas by ancestral members of Prodoxidae occurred 41.5 ± 9.8 mya, and the origin of active pollination in the ancestor of *Tegeticula* and *Parategeticula* occurred 35.6 ± 9.0 mya. *Yucca* and *Hesperoyucca* are separated by 27.03 my (95% HPD: 37.10-18.31 my) (Table 3.4), which overlaps with the estimated timing of *Yucca* colonization and pollinator origin in the moths. There are two scenarios for the evolution of the yucca-yucca moth relationship. The first is that the moths independently colonized *Yucca* and *Hesperoyucca*, and that morphological similarity between these two genera is an example of convergent evolution. The origin of the *Yucca* stem group is 20.22 mya (95% HPD: 28.17-13.84 mya), and the origin *Hesperoyucca* is 15.98 my (95% HPD: 25.04-7.43 mya). The range of *Yucca* origin overlaps with that of yucca moth pollination origin, but the origin of *Hesperoyucca* does not. This suggests that yucca moths may have originated on an ancestor of extant *Yucca* species and later colonized *Hesperoyucca*. A second possibility is that active yucca moth pollination evolved on a common ancestor of *Yucca* and *Hesperoyucca*, and active pollination by yucca moths was lost on multiple lineages within the ABK clade. This would require a mnimum of three independent losses of yucca moth pollination: i) in the clade formed by *Agave s. l.* and *Beschorneria* after the divergence of *Yucca* , ii) in the HCC clade after the divergence of the SHH clade, and iii) in *Schoenorlirion* + *Hesperaloe* after the divergence *Hesperoyucca*. The timing of the *Yucca-Hesperoyucca* split overlaps with both the origin of pollinators and the suspected evolution of active polination within the *Yucca* lineage. Yucca moths have demonstrated that it is possible to escape the “obligate” mutualistic relationship of yucca-yucca moth pollination (Pellmyr et al., 1996; Pellmyr and Leebens-Mack, 2000), and recent work by Rentsch and Leebens-Mack (unpublished data) has shown that this has occurred

in *Yucca aloifolia*. A more detailed estimate of the timings of diversification in the yucca moths is needed to further investigate these two hypotheses.

The utility of whole chloroplast genome phylogenomics is demonstrated by the reconstruction of the relationships within Agavoideae and the ABK clade. Rapid diversification is an obstruction to well-resolved relationships that can be remedied with increased taxon sampling and increased sequence data per taxon (Leebens-Mack et al., 2005; Xi et al., 2012). New methodologies in sequencing make these increased sampling schemes affordable. Relationships within the ABK clade have long been obscured by a rapid diversification and now that the phylogeny is becoming more resolved, more evolutionary studies regarding deep clades can be conducted in the proper phylogenetic framework. Future analyses will include coalescence-based analyses including the plastid genome and many nuclear loci.

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.

ALTEKAR, G., S. DWARKADAS, J. HUELSENBECK, AND F. RONQUIST. 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* 20: 407–415.

BOGLER, D., AND B. SIMPSON. 1995. A chloroplast DNA study of the Agavaceae. *Systematic Botany* 20: 191–205. Available at: http://www.jstor.org/stable/10.2307/2419449 [Accessed October 22, 2012].

BOGLER, D.J., J.C. PIRES, AND J. FRANCISCO-ORTEGA. 2006. Phylogeny of Agavaceae based on *ndhF*, *rbcL*, and ITS sequences : implications of molecular data for classification. *Aliso* 22: 313–328.

BOGLER, D.J., AND B.B. SIMPSON. 1996. Phylogeny of Agavaceae Based on ITS rDNA Sequnce Variation. *American Journal of Botany* 83: 1225–1235.

CAVE, M. 1948. Sporogenesis and embryo sac development of *Hesperocallis* and *Leucocrinum* in relation to their systematic position. *American Journal of Botany* 35: 343–349.

CAVE, M.S. 1970. *Chromosomes of the California Liliaceae*. University of California Press.

CHASE, M.W. ET AL. 2006. Multigene analyses of monocot relationships: a summary. *Aliso* 22: 63–75.

CHASE, M.W., J.L. REVEAL, AND M.F. FAY. 2009. A subfamilial classification for the expanded asparagalean families Amaryllidaceae , Asparagaceae and Xanthorrhoeaceae. *Botanical Journal of the Linnean Society* 161: 132–136.

CHUNG, M., AND S. JONES JR. 1989. Pollen morphology of *Hosta* Tratt. (Funkiaceae) and related genera. *Bulletin of the Torrey Botanical Club* 116: 31–44.

CHUPOV, V., AND N. KUTIAVINA. 1981. Serological studies in the order Liliales. 1. *Botanicheskii Zhurnal* 66: 75–81.

CRUDEN, R.W. 1999. A new subgenus and fifteen new species of *Echeandia* (Anthericaceae) from Mexico and the United States. *Novon* 9: 325–338.

DOYLE, J., J. DOYLE, AND J. PALMER. 1995. Multiple independent losses of two genes and one intron from legume chloroplast genomes. *Systematic Botany* 20: 272–294.

DOYLE, J.J., J.I. DAVIS, R.J. SORENG, D. GARVIN, AND M.J. ANDERSON. 1992. Chloroplast DNA inversions and the origin of the grass family (Poaceae). *Proceedings of the National Academy of Sciences* 89: 7722–7726.

DOYLE, J.J., AND J.L. DOYLE. 1987. CTAB DNA extraction in plants. *Phytochemical Bulletin* 19: 11–15.

DRUMMOND, A. J., M. A. SUCHARD, D. XIE, AND A. RAMBAUT. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 29: 1969–1973.

DRUMMOND, A.J., M.A. SUCHARD, D. XIE, AND A. RAMBAUT. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology* 4: 699–710.

DUBCOVSKY, J., AND J. DVORAK. 2007. Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* 316: 1862–1865.

EFLORAS. 2008. eFloras [online]. Available at http://www.efloras.org [Accessed November 1, 2012].

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

EGUIARTE, L.E. 1995. Hutchinson (Agavales) Vs. Huber y Dahlgren (Asparagales): análisis moleculares sobre' la filogenia y evolución de la familia Agavaceae sensu Hutchinson dentro de las monocotiledóneas. *Boletín de la Sociedad Botánica de México* 56: 45–56.

EGUIARTE, L.E., M.R. DUVALL, G.H. LEARN, AND M. CLEGG. 1994. The systematic status of the Agavaceae and Nolinaceae and related Asparagales in the monocotyledons: an analysis based on the rbcL gene sequence. *Boletín de la Sociedad Botánica de México* 54: 35–56.

FERNÁNDEZ, A., AND J. DAVIÑA. 1991. Heterochromatin and genome size in *Fortunatia* and *Camassia* (Hyacinthaceae). *Kew bulletin* 46: 255–305.

FISHBEIN, M. ET AL. 2010. Phylogeny of *Camassia* (Agavaceae) inferred from plastid *rpl16* Intron and *trnD – trnY – trnE – trnT* intergenic spacer DNA sequences : implications for species delimitation *Systematic Botany* 35: 77–85.

FRIIS, E.M., K.R. PEDERSEN, AND R. CRANE. 2010. Diversity in obscurity: fossil flowers and the early history of angiosperms. *Philosophical Transactions of the Royal Society of London, B, Biological Sciences* 365: 369–82.

GIVNISH, T.J. ET AL. 2010. Assembling the tree of the Monocotyledons: plastome sequence phylogeny and evolution of Poales. *Annals of the Missouri Botanical Garden* 97: 584–616.

GOOD-AVILA, S.V., V. SOUZA, B.S. GAUT, AND L.E. EGUIARTE. 2006. Timing and rate of speciation in *Agave* (Agavaceae). *Proceedings of the National Academy of Sciences* 103: 9124–9.

GOULD, F. 1942. A systematic treatment of the genus *Camassia* Lindl . *American Midland Naturalist* 28: 712–742.

GRAHAM, S.W. ET AL. 2006. Robust inference of monocot deep phylogeny using an expanded multigene plastid data set. *Aliso* 22: 3–20.

GRANICK, E.B. 1944. A karyosystematic study of the genus *Agave*. *American Journal of Botany* 31: 283–298.

GUISINGER, M.M., T.W. CHUMLEY, J.V. KUEHL, J.L. BOORE, AND R.K. JANSEN. 2010. Implications of the plastid genome sequence of *Typha* (Typhaceae, Poales) for understanding genome evolution in Poaceae. *Journal of Molecular Evolution* 70: 149–166.

HALPIN, K.M. 2011. A chloroplast phylogeny of Agavaceae subfamily Chlorogaloideae with a focus on speciesrRelationships in *Hastingsia*. Oklahoma State University.

HANSEN, D.R. ET AL. 2007. Phylogenetics and evolutionary implications of complete chloroplast genome sequences of four early-diverging angiosperms: *Buxus* (Buxaceae), *Chloranthus* (Chloranthaceae), *Dioscorea* (Dioscoreaceae), and *Illicium* (Schisandraceae). *Molecular Phylogenetics and Evolution* 45: 547–563.

HARLEY, M.M. 2006. A summary of fossil records for Arecaceae. *Botanical Journal of the Linnean Society* 151: 39–67.

HUELSENBECK, J., AND F. RONQUIST. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Nucleic Acids Research* 17: 754–755.

JANSEN, R.K. ET AL. 2005. Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods in Enzymology* 395: 348–384.

KIM, J.-H., D.-K. KIM, F. FOREST, M.F. FAY, AND M.W. CHASE. 2010. Molecular phylogenetics of Ruscaceae sensu lato and related families (Asparagales) based on plastid and nuclear DNA sequences. *Annals of Botany* 106: 775–90.

LAVIN, M., J.J. DOYLE, AND J.D. PALMER. 1990. Evolutionary significance of the loss of the chloroplast-DNA inverted repeat in the Leguminosae subfamily Papilionoideae. *Evolution* 44: 390–402.

LEEBENS-MACK, J. ET AL. 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.

LEEBENS-MACK, J., O. PELLMYR, AND M. BROCK. 1998. Host specificity and the genetic structure of two yucca moth species in a yucca hybrid zone. *Evolution* 52: 1376–1382.

MAGALLÓN, S., AND A. CASTILLO. 2009. Angiosperm diversification through time. *American Journal of Botany* 96: 349–365.

MAIER, R.M., K. NECKERMANN, G.L. IGLOI, AND H. KÖSSEL. 1995. Chloroplast DNA inversions and the origin of the grass family (Poaceae). *Journal of Molecular Biology* 251: 614–628.

MCKAIN, M.R. ET AL. 2012. Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in Agavoideae (Asparagaceae). *American Journal of Botany* 99: 397–406.

MCKELVEY, S.D., AND K. SAX. 1933. Taxonomic and cytological relationships of *Yucca* and *Agave*. *Journal of the Arnold Arboretum* 14: 76–81.

MOORE, M.J., S. SOLTIS, C.D. BELL, J.G. BURLEIGH, AND D.E. SOLTIS. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Academy of Sciences* 107: 4623–4628.

NYLANDER, J. A A, J.C. WILGENBUSCH, D.L. WARREN, AND D.L. SWOFFORD. 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics* 24: 581–583.

OKUMURA, S. ET AL. 2006. Transformation of poplar (*Populus alba*) plastids and expression of foreign proteins in tree chloroplasts. *Transgenic Research* 15: 637–646.

ORR, H.A. 1996. Dobzhansky, Bateson, and the genetics of speciation. *Genetics* 144: 1331–1335.

PALMER, J. 1985. Comparative organization of chloroplast genomes. *Annual Review of Genetics* 19: 325–354.

PELLMYR, O. 2003. Yuccas, yucca moths, and coevolution : A Review. *Annals of the Missouri Botanical Garden* 90: 35–55.

PELLMYR, O., AND J. LEEBENS-MACK. 1999. Forty million years of mutualism: evidence for Eocene origin of the yucca-yucca moth association. *Proceedings of the National Academy of Sciences* 96: 9178–9183.

PELLMYR, O., K. A SEGRAVES, D.M. ALTHOFF, M. BALCÁZAR-LARA, AND J. LEEBENS-MACK. 2007. The phylogeny of yuccas. *Molecular Phylogenetics and Evolution* 43: 493–501.

PIRES, J.C. ET AL. 2004. Molecular data confirm the phylogenetic placement of the enigmatic *Hesperocallis* (Hesperocallidaceae) with Agave. *Madroño* 51: 307–311.

PIRES, J.C. ET AL. 2006. Phylogeny, genome size, and chromosome evolution of Asparagales. *Aliso* 22: 287–304.

PLANT LIST. 2010. The plant list, version 1 [online]. Available at: http://www.theplantlist.org/ [Accessed September 8, 2012].

RATAN, A. 2009. Assembly algorithms for next generation sequence data. Pennsylvania State University.

RENTSCH, J.D., AND J. LEEBENS-MACK. 2012. Homoploid hybrid origin of *Yucca gloriosa*: intersectional hybrid speciation in *Yucca* (Agavoideae, Asparagaceae). *Ecology and Evolution* 2: 2213–2222.

RILEY, C.V. 1872. The fertilization of the yucca plant by *Pronuba yuccasella*. *Canadian Entomologist* 4: 182.

ROCHA, M. ET AL. 2006. Pollination biology and adaptive radiation of Agavaceae, with special emphasis on the genus *Agave*. *Aliso* 22: 329–344.

RONQUIST, F., AND J. HUELSENBECK. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572–1574.

SATO, D. 1935. Analysis of the karyotypes in *Yucca*, *Agave* and the related genera with special reference to the phylogenetic significance. *The Japanese Journal of Genetics* 11: 272–278.

SEBERG, O. ET AL. 2012. Phylogeny of the Asparagales based on three plastid and two mitochondrial genes. *American Journal of Botany* 99: 875–889.

SEN, S. 1975. Cytotaxonomy of Liliales. *Feddes Repertorium* 86: 255–305.

SHERMAN, H., AND R. BECKLING. 1991. The generic distinctness of *Schoenolirion* and *Hastingsia*. *Madroño* 38: 130–138.

SHERMAN, H.L. 1969. A systematic study of the genus *Schoenolirion* (Liliaceae). Vanderbilt University.

SMITH, C.I., O. PELLMYR, D.M. ALTHOFF, M. BALCÁZAR-LARA, J. LEEBENS-MACK, AND K. A SEGRAVES. 2008. Pattern and timing of diversification in *Yucca* (Agavaceae): specialized pollination does not escalate rates of diversification. *Proceedings of the Royal Society, B, Biological Sciences* 275: 249–58.

SMITH, S. A, AND M.J. DONOGHUE. 2008. Rates of Molecular Evolution are Linked to Life History in Flowering Plants. *Science* 322: 86–89.

STEELE, R., K.L. HERTWECK, D. MAYFIELD, M.R. MCKAIN, J. LEEBENS-MACK, AND J.C. PIRES. 2012. Quality and quantity of data recovered from massively parallel sequencing: Examples in Asparagales and Poaceae. *American Journal of Botany* 99: 330–348.

SUYAMA, M., D. TORRENTS, AND BORK. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* 34: W609–W612.

TAMURA, M. 1995. A karyological review of the orders Asparagales and Liliales (Monocotyledonae). *Feddes Repertorium* 106: 83–111.

TIDWELL, W., AND L. PARKER. 1990. *Protoyucca shadishii* gen. et sp . nov ., An Arborescent Monocotyledon with Secondary Growth from the Middle Miocene of Northwestern Nevada, U.S.A. *Review of Palaeobotany and Palynology* 62: 79–95.

UEDA, M. ET AL. 2008. Substitution of the gene for chloroplast RPS16 was assisted by generation of a dual targeting signal. *Molecular Biology and Evolution* 25: 1566–1575.

WANG, R.-J., C.-L. CHENG, C.-C. CHANG, C.-L. WU, T.-M. SU, AND S.-M. CHAW. 2008. Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. *BMC Evolutionary Biology* 8: 36.

WHITAKER, T.W. 1934. Chromosome constitution in certain monocotyledons. *Journal of the Arnold Arboretum* 15: 135–143.

WU, C.-S., Y.-N. WANG, C.-Y. HSU, C.- LIN, AND S.-M. CHAW. 2011. Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and cupressophytes and influence

of heterotachy on the evaluation of gymnosperm phylogeny. *Genome Biology and Evolution* 3: 1284–1295.

WYMAN, S.K., R.K. JANSEN, AND J.L. BOORE. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20: 3252–3255.

XI, Z. ET AL. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proceedings of the National Academy of Sciences*109: 17519–17524.

ZERBINO, D.R., AND E. BIRNEY. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 18: 821–829.

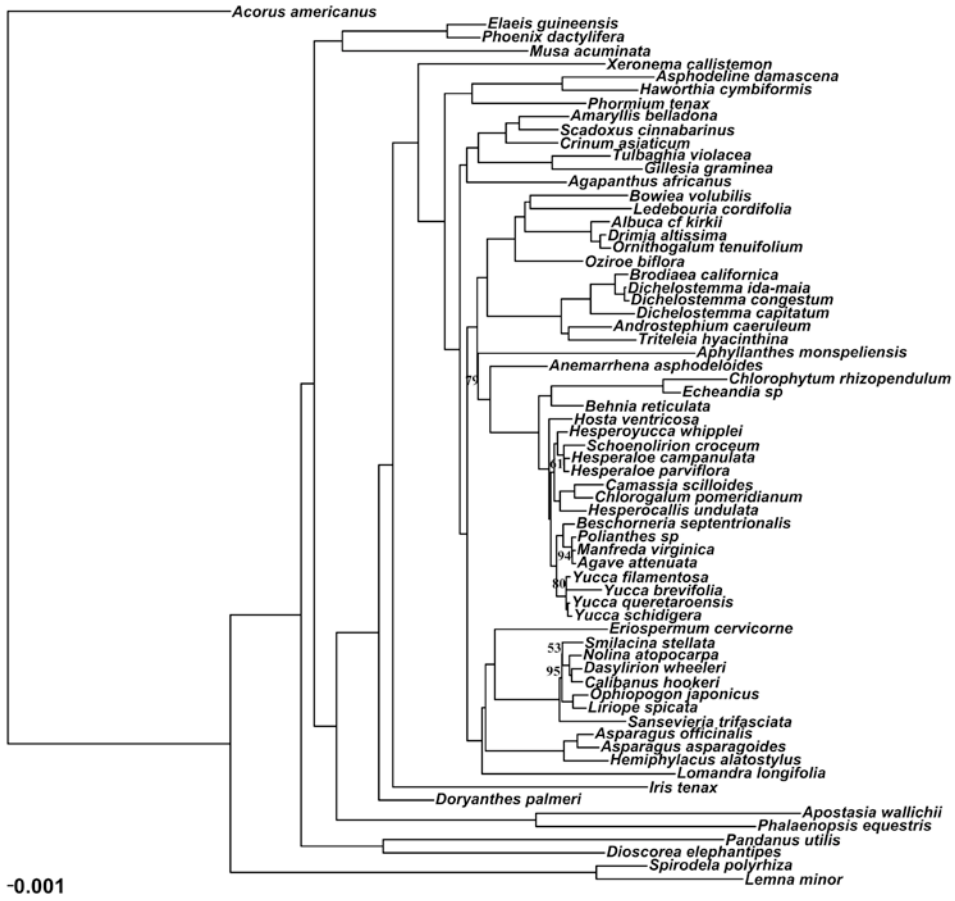

Figure 3.1. Maximum likelihood phylogeny of taxa used in the divergence time estimation analysis.

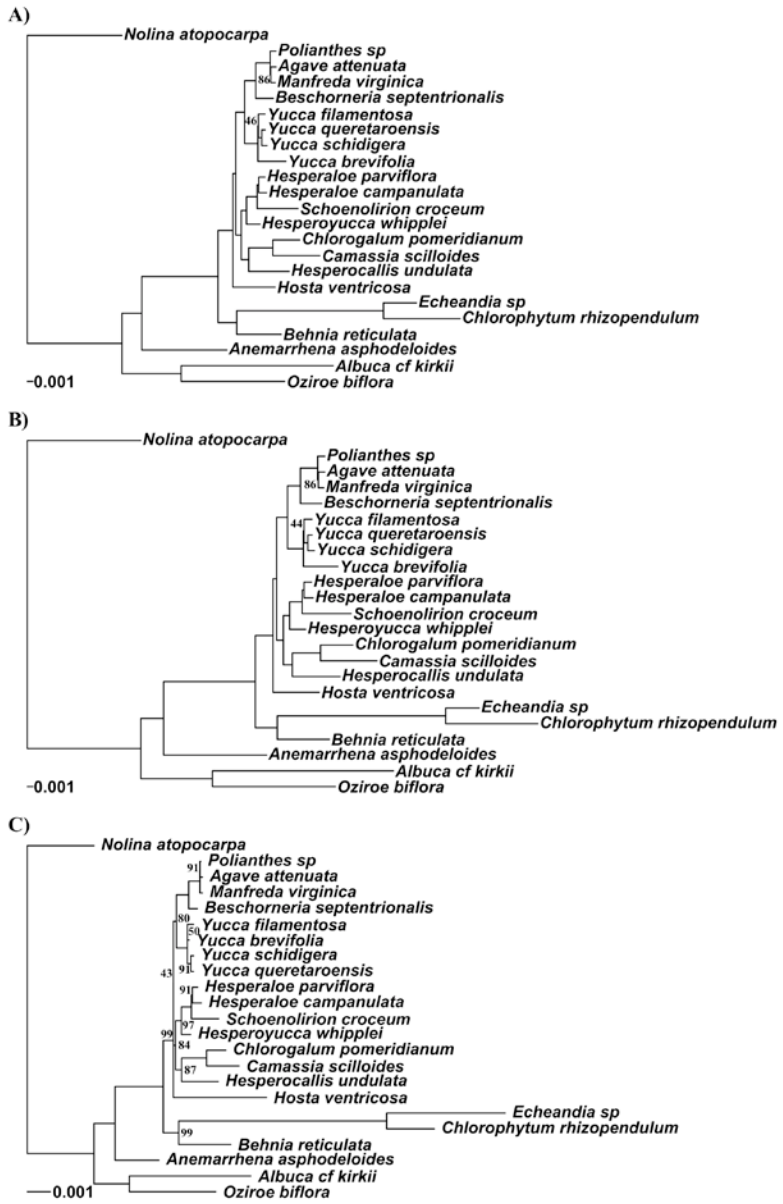

Figure 3.2. Maximum likelihood phylogenies of chloroplast regions for Agavoideae.

Maximum likelihood phylogenies using full plastome (A), single copy (B) and inverted repeat (C) regions. The full plastome and single copy regions are plotted on the same scale, while the inverted repeat region is on a quarter scale compared to A and B. All nodes are bootstrap values (BSV) of 100 unless otherwise noted.

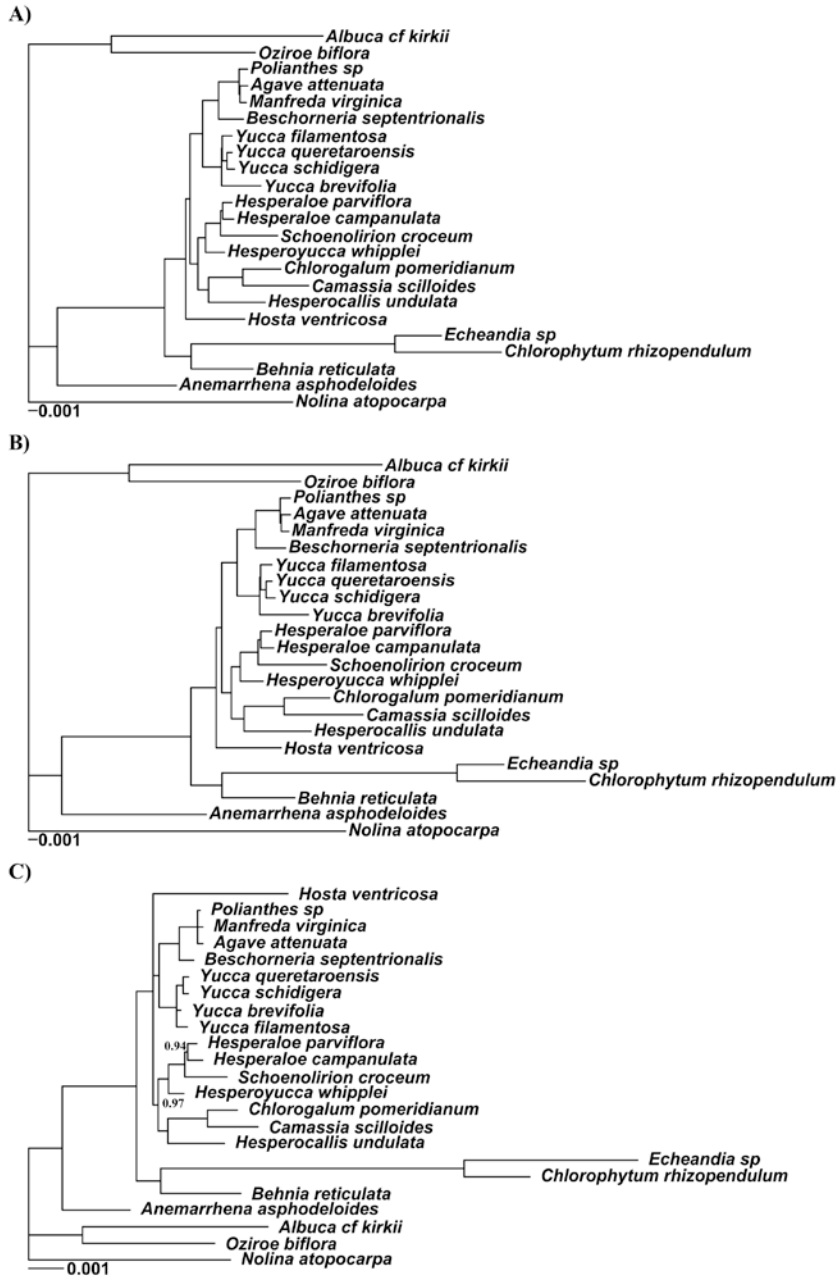

Figure 3.3. Bayesian phylogenies of chloroplast regions for Agavoideae.

Bayesian phylogenies using full plastome (A), single copy (B) and inverted repeat (C) regions. The full plastome and single copy regions are plotted on the same scale, while the inverted repeat region is on a quarter scale compared to A and B. All nodes are posterior probabilities (PP) of 1.0 unless otherwise noted.

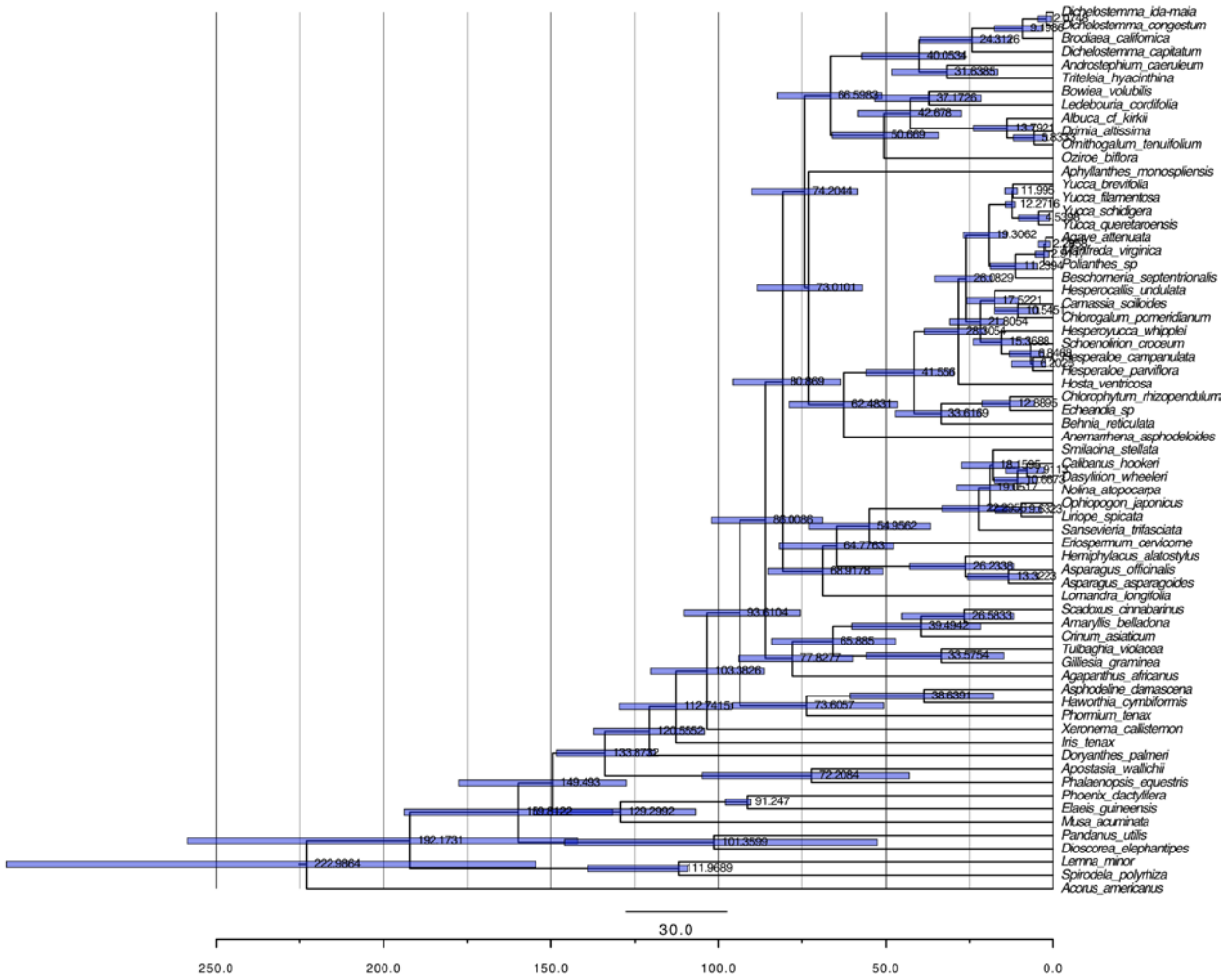

Figure 3.4. Chronogram of Agavoideae with extended monocot sampling.

Mean ages of nodes are given with bars representing 95 % high density probability (HPD) estimations for node ages. Ages estimated here suggest that the ABK clade originated ~ 28 million years ago.

Table 3.1. Chloroplast genome taxa and assembly statitistics.

rpl32 in *Yucca schidigera* is missing due to lack of data, not necessarily the gene being absent from the plastid genome.

Species	Collector and collection number (herbarium)	Isolation	Final Size	Gaps	Genes Missing	Pseudogenes	No Start Codon	*rps19* in IR
Agave attenuata Salm.	McKain 109 (GA)	Plastid	157451	No				yes
Albuca cf. *kirkii* (Baker) Brenan	McKain 111 (GA)	DNA	156401	Yes		*infA, ndhK*		yes
Anemarrhena asphodeloides Bunge	Steele 1089 (UMO)	DNA	156844	No				yes
Behnia reticulata (Thunb.) Didr.		DNA	158066	Yes		*infA, rps19*		yes
Beschorneria septentrionalis A. Garcîa-Mendoza	McKain 108 (GA)	Plastid	157043	No	*rps16*, exon 2			yes
Camassia scilloides (Raf.) Cory.	McKain 107 (GA)	Plastid	155453	No	*rps19*	*infA*		no
Chlorogalum pomeridianum (DC.) Kunth	McKain 104 (GA)	Plastid	157288	No		*rps19*	*cemA*	no
Chlorophytum rhizopendulum Bjorå and Hemp	McKain 110 (GA)	Plastid	153594	Yes	*rps19*	*infA*	*cemA*	no
Echeandia sp. Ortega	Steele 1101 (UMO)	DNA	154356	No	*rps19*	*infA*	*cemA*	no
Hesperaloe campanulata G. Starr	NYBG	???	157446	No		*rps19*		no
Hesperaloe parviflora (Torr.) J. M. Coult.	McKain 102 (GA)	Plastid	157393	No		*rps19*		no
Hesperocallis undulata A. Gray	Prince, Columbus & Schmidt (RSA)	Plastid	157143	No		*infA, rps19*		yes
Hesperoyucca whipplei (Torr.) Trel.	McKain 119 (GA)	Plastid	157832	No				yes

Hosta ventricosa (Salisb.) Stearn	McKain 106 (GA)	Plastid	156577	No	*rps16*, exon 2	*rps19*		yes
Manfreda virginica (L.) Salisb.	Missouri Native Plants	DNA	157308	No				yes
Nolina atopocarpa Bartlett	McKain 114 (GA)	Plastid	156792	No		*ndhK*		yes
Oziroe biflora (Ruiz & Pav.) Speta	Arroyo 28509 (xx)	DNA	155633	No				yes
Polianthes sp. L.	JC Pires 2011 (UMO)	DNA	157337	No		*ndhK*		yes
Schoenolirion croceum	McKain 102 (GA)	DNA	156632	No		*rps19*		yes
Yucca brevifolia L.	Smith	DNA	158028	No		*infA*	*cemA*	yes
Yucca filamentosa L.	McKain 101 (GA)	Plastid	157785	No		*infA*	*cemA*	yes
Yucca queretaroensis Piña	Eguiarte	Plastid	157814	No		*infA, ndhK*	*cemA*	yes
Yucca schidigera Roezi ex Ortgies	NCBI		156158	Yes	*rpl32*	*infA, ndhK*	*cemA*	yes

Table 3.2. Primers used to fill gaps for plastid genome assemblies.

Name	Sequence	Region	Mate	Size	Species Sequenced
accD_F1M	GGTTCACAAGCGGCTGAGTAT	accD-psaI	R22_psaI		*Anemarrhena, Oziroe, Polianthes*
atpB_F2M	GTACTGGGCCAATAATTTGAGC	atpB-rbcL	rbcL_R2M		
rbcL_R2M	CTCTGTTTGTGGTGACATAAGTC	atpB-rbcL	atpB_F2M	919	*Oziroe*
atpF_F3M	CTCTTCCCGAACCAAACATG	atpF intron	atpF_R3M		*Albuca, Anemarrhena,*
atpF_R3M	ATGAGGAATTAGTGGATTCGCTC	atpF intron	atpF_F3M	1327	*Hesperoyucca, Polianthes*
matK_F4M	GTGCAATATGGTCAGAACAGAG	matK-trnK	trnK_R4M		
trnK_R4M	CTCAACGGTAGAGTATTCGGC	matK-trnK	matK_F4M	748	*Anemarrhena, Chlorogalum*
ndhA_F5M	CTATAGGYTGACGCCACAG	ndhA intron	ndhAin_R5M		
ndhAin_R5M	CGATTCCGATCTAGAGTATGCTC	ndhA intron	ndhA_F5M	719	*Oziroe, Chlorogalumalum*
ndhC_F6M	GGATTTGGTCTGTCGAATTGTTC	ndhC-trnV	trnV_R6M		
trnV_R6M	GAGAGCTTCTCTGGTCCTTC	ndhC-trnV	ndhC_F6M	1286	*Echeandia, Polianthes*
ndhF_F7M	GCAMTYGGTCGTGTGAACC	ndhF-rpl32	rpl32_R7M		*Albuca, Anemarrhena,*
rpl32_R7M	GCAGCTAAATWWCCYTTTTTCTTCC	ndhF-rpl32	ndhF_F7M	1719	*Chlorophytum, Manfeda, Oziroe, Polianthes*
rpl32_F8M	GAGCARTACATGTCTTTCACATAC	rpl32-ccsA	ccsA_R8M		*Albuca, Anemarrhena,*
ccsA_R8M	CACCATAGCGGCTTASTTGAA	rpl32-ccsA	rpl32_F8M	1388	*Hesperoyucca, Oziroe, H. campanulata*
psaC_F9M	GATAGACCCATGCTGCGAGTTG	psaC-ndhI	ndhG_R9M		*Anemarrhena, Chlorogalum,*
ndhG_R9M	GATTTACCTGGACCAATACACGA	psaC-ndhG	psaC_F9M	1739	*Oziroe, Chlorophytumytum, Albuca*
ndhH_F10M	GTRACGATYAGTCGAAGAACACC	ndhH_rps14	rps15_R10M		*Albuca, Anemarrhena,*
rps15_R10M	ATCTCCCAGACGTCGTMG	ndhH_rps15	ndhH_F10M	1246	*Oziroe*
ycf1_F11M	GACAAAGATAGCCCAGTTTACG	ycf1_ycf1	ycf1_R11M		
ycf1_R11M	CCACACGYTTGCCTTTTC	ycf1_ycf1	ycf1_F11M	1041	*Chlorogalum*
ycf1_F12M	ATTGCAAYYCCCCGAGTG	ycf1_ycf1	ycf1_R12M		*Oziroe, Albuca,*
ycf1_R12M	GGMCCACTTGGTMTGAGAT	ycf1_ycf1	ycf1_F12M	1336	*Polianthes, Chlorogalum*
ycf1_F13M	ATCTCAKACCAAGTGGKCC	ycf1_ycf1	ycf1_R13M		*Oziroe, Albuca,*
ycf1_R13M	CCTTGGCYTTGTTGTTCRTT	ycf1_ycf1	ycf1_F13M	918	*Chlorogalum*
ycf1_F14M	CACCTGTGTCTACTATTTAGGC	ycf1_ndhF	ndhF_R14M	958	*Oziroe, Hesperoyucca*

ndhF_R14M	CCTATCCTCACGAGTCGGAC	ycf1_ndhF	ycf1_F14M		
ndhK_F15M	GTCTGCTTGCCTAGGACTC	ndhK_ndhC	ndhC_R15M		*Polianthes, Manfreda,*
ndhC_R15M	CCTAAGATGGGTGGTGGATG	ndhK_ndhC	ndhK_F15M	726	*Y. queretaroensis*
petM_R16M	GCTACTGCACTGTTCATTCTAG	petN-petM	F11_petN_30364	1191	*Albuca, Chlorophytum, Polianthes, Hesperoyucca*
psaA_F17M	CCTAGTAATCCTGCTAAGTGGTG	psaA-ycf3	ycf3_R17M		
ycf3_R17M	GAATCGATTGCTGAGCCGTATG	psaA-ycf3	psaA_F17M	1463	*Chlorogalum*
psaJ_F18M	TTGAACTRCAGCATCTGACC	psaJ-rpl33	rpl33_R18M		
rpl33_R18M	CTCTTCTATTTTCGACCCGAAC	psaJ-rpl33	psaJ_F18M	693	*Albuca, Chlorogalum*
trnH_F19M	TTGATCCACTTGGCTACATCC	trnH_psbA	psbA_R19M		
psbA_R19M	CGTCCTTGGATTGCTGTTG	trnH_psbA	trnH_F19M	1222	*Hesperoyucca, Chlorogalum*
trnGsp_R20M	GAATGGAYCYTTTGTCGAACA	psbI-trnG	F4_psbI_8158	1662	*Anemarrhena, Chlorophytum, Echeandia, Oziroe, Hesperoyucca*
psbIsp_F21M	TCCMATCGTRGATGTTATGCC	psbI-trnS	R3_trnS_8455	820	*Anemarrhena, Albuca*
rpoC2_F22M	GTACCTAAGGGACCCAACAAATC	rpoC2_rpoC2sp	rpoC2sp_R22M		
rpoC2sp_R22M	GTGGAGTATTCACAGGCGG	rpoC2_rpoC2sp	rpoC2_F22M	1909	*Chlorogalum, Polianthes*
rpoC1_F23M	GACYCGTTTMCCAAGCAGAG	rpoC1_rpoC1	rpoC1_R23M		
rpoC1_R23M	GTGAGTAGGGACCTAAAAGATCG	rpoC1_rpoC1	rpoC1_F23M	1116	*Chlorogalum*
rpoB_F24M	GCCAAGTATGGCTCGTAATAATC	rpoB_rpoB	rpoB_R24M		
rpoB_R24M	GGVTTAATTTGGAAAACCGGYAG	rpoB_rpoB	rpoB_F24M	2046	*Chlorogalum, Hesperoyucca*
rpoBsp_F25M	CTGAYTAAATCCAGGTATTGYGG	rpoBsp_trnC	trnCsp_R25M		*Chlorogalum, Oziroe, Manfreda,*
trnCsp_R25M	GTTGATCAGGCGACACCC	rpoBsp_trnC	rpoBsp_F25M	1487	*Albuca*
rps16_R26M	CGATGTGGTAGAAAGCAACG	rps16_rps16in	F3_rps16_4917	1024	*Polianthes, Oziroe*
rps16in_F27M	CGTTGCTTTCTACCACATCG	rps16in_rps16	rps16_R27M		*Hesperoyucca, Chlorogalum, Anemarrhena, Oziroe,*
rps16_R27M	GTAAGGCDKCGGGTTTTG	rps16in_rps16	rps16in_F27M	1747	*H. parviflora*
rps2_F28M	GCATCAAAAACTAAATCACAGGC	rps2_rpoC2	rps2sp_R28M		
rps2sp _R28M	GTTCGCTCGATTGGTTCAA	rps2_rpoC2	rps2_F28M	1089	*Hesperoyucca, Chlorogalum*
rps23_F29M	CACCTCATACGGCTCCTCG	rps23_rps23sp	rps23sp_R29M		
rps23sp_R29M	GTAGGACTAGTGCCAACAG	rps23_rps23sp	rps23_F29M	787	*Chlorogalum*
trnDYET_F30M	CYCTGGGTCATACACAGATCC	trnDYET	trnDYET_R30M		
trnDYET_R30M	GGTTAATGGGGACGGACTG	trnDYET	trnDYET_F30M	1193	*Hesperoyucca, H. campanulata*
trnT_R31M	CATGATTCAAKGGGTCAGGTC	trnT-psbD	F12_trnT_33414	1285	*Albuca, Chlorogalum*
trnL_R32M	CAATTTCGCCATATCCCCT	trnL-trnT	F18_trnT	1024	*Albuca, Chlorophytum, Polianthes, Oziroe, Manfeda*

trnL_F33M	GACGAGAATAAAGAGAGAGTCCC	trnL-trnF	trnF_R33M		
trnF_R33M	CTACCAACTGAGCTATCCCG	trnL-trnF	trnL_F33M	538	*Oziroe*
ycf4_F34M	GAAGAATGAATGTTTTCTCCGC	ycf4_cemA	cemA_R34M		*Anemarrhena, Chlorogalum, Polianthes*
cemA_R34M	GCAAGATATGGGAGGGAAGTC	ycf4_cemA	ycf4_F34M	1074	

Table 3.3. Taxa and data sources for divergence timing estimation analysis.

Species	Source	Missing Genes
Acorus americanus	GenBank: NC_010093	
Agapanthus africanus	Steele et al.	
Agave attenuata	This study	
Albuca cf. *kirkii*	This study	
Amaryllis belladona	Steele et al.	
Androstephium caruleum	Steele et al.	
Anemarrhena asphodeloides	This study	
Aphyllanthes monspliensis	Steele et al.	
Apostasia wallichii	Givnish et al.	*matK*-portion, *ndhE*-portion
Asparagus asparagoides	Steele et al.	
Asparagus officinalis	Steele et al.	
Asphodeline damascena	Steele et al.	
Behnia reticulata	This study	
Beschoneria septentrionalis	This study	
Bowiea volubilis	Steele et al.	
Brodieae californica	Steele et al.	
Calibanus hookeri	Steele et al.	
Camassia scilloides	This study	
Chlorogalum pomeridianum	This study	
Chlorophyutm rhizopendulum	This study	
Crinum asiaticum	Steele et al.	
Dasylirion wheelri	Steele et al.	
Dichelostemma capitatum	Steele et al.	
Dichelostemma congestum	Steele et al.	
Dichelostemma ida-maia	Steele et al.	
Dioscorea elephantipes	GenBank: NC_009601	
Doryanthes palmeri	Steele et al.	
Drimia altissima	Steele et al.	
Echeandia sp.	This study	
Elaeis guineensis	GenBank: NC_017602	
Eriospermum cervicorne	Steele et al.	
Gillesia graminea	Steele et al.	
Haworthia cymbiformis	Steele et al.	
Hemiphylacus alatostylus	Steele et al.	
Hesperaloe campanulata	This study	
Hesperaloe parviflora	This study	
Hesperocallis undulata	This study	
Hesperoyucca whipplei	This study	
Hosta ventricosa	This study	
Iris tenax	Steele et al.	
Ledebouria cordifolia	Steele et al.	
Lemna minor	GenBank: NC_010109	
Liriope spicata	Steele et al.	
Lomandra longifolia	Steele et al.	

Mandreda virginica	This study	
Musa acuminata	Givnish et al.	
Nolina atopocarpa	This study	
Ophiopogon japonicus	Steele et al.	
Ornithogalum tenuifolium	Steele et al.	
Oziroe biflora	This study	
Pandanus utilis	Givnish et al.	*matK*-portion, *ndhE*-portion, *rpoC2*-portion
Phalaenopis equestris	GenBank: NC_0176091	
Phoenix dactylifera	GenBank: NC_013991	
Phormium tenax	Steele et al.	
Polianthes sp.	This study	
Sansevieria trifasciata	Steele et al.	
Scadoxus cinnabarinus	Steele et al.	
Schoenolirion croceum	This study	
Smilacina stellata	Steele et al.	
Spirodela polyrhiza	GenBank: NC_015891	
Triteleia hyacinthina	Steele et al.	
Tulbaghia violacea	Steele et al.	
Xeronema callistemon	Steele et al.	
Yucca brevifolia	This study	
Yucca filamentosa	This study	
Yucca queretaroensis	This study	
Yucca schidigera	This study	

Table 3.4. Age estimations of major divergence events in the ABK clade and Agavoideae compared to those found in other studies.

Split	Age (My)	Lower 95% HPD (My)	Upper 95% HPD (My)	Good-Avila et al. (2006) (My)	Smith et al. (2008) (My)
Anemarrhena	63.00	47.27	79.25	-	-
ABK-*Behnia et al.*	43.06	29.43	56.94	34.2-31.7	-
Hosta	29.39	19.97	40.12	29.2-20.7	23.69 +/- 9.66
YBA + HHS-HCC	27.03	18.31	37.10	-	-
Yucca	20.22	13.84	28.17	19.5-11.1	15.81 +/- 7.35
Yucca crown	12.59	11.36	14.55	-	10.19 +/- 6.49
Beschorneria	11.77	4.72	19.82	-	-
Agave s. l. crown	3.30	1.10	5.95	13.1-5.9	12.63 +/- 5.11
HHS-HCC	22.57	14.35	32.08	-	-
Hesperocallis	18.31	10.40	27.35	-	-
Camassia + *Chlorogalum*	11.01	4.18	18.44	-	-
Hesperoyucca	15.98	7.44	25.04	-	-
Schoenolirion + *Hesperaloe*	7.53	2.77	13.47	-	-

CHAPTER IV

PHYLOGENOMIC ANALYSIS OF TRANSCRIPTOME DATA ELUCIDATES CO-OCCURRENCE OF A PALEOPOLYPLOID EVENT AND THE ORIGIN OF BIMODAL KARYOTYPES IN AGAVOIDEAE (ASPARAGACEAE)[1]

[1] [1] 2012, *American Journal of Botany*, 99: 397-406. Reprinted here with permission from the publisher

Abstract

- *Premise of the study*: The stability of the bimodal karyotype found in *Agave* and closely related species has long interested botanists. The origin of the bimodal karyotype has been attributed to allopolyploidy, but this hypothesis has not been tested. Next-generation transcriptome sequence data were used to test whether a paleopolyploid event occurred on the same branch of the Agavoideae phylogenetic tree as the origin of the *Yucca-Agave* bimodal karyotype.
- *Methods*: Illumina RNA-seq data were generated for phylogenetically strategic species in Agavoideae. Paleopolyploidy was inferred in analyses of frequency plots for synonymous substitutions per synonymous site (K_s) between *Hosta*, *Agave* and *Chlorophytum* paralogous and orthologous gene pairs. Phylogenies of gene families including paralogous genes for these species and outgroup species were estimated in order to place inferred paleopolyploid events on a species tree.
- *Key results*: *Ks* frequency plots suggested paleopolyploid events in the history of the genera *Agave*, *Hosta* and *Chlorophytum.* Phylogenetic analyses of gene families estimated from transcriptome data revealed two polyploid events: one predating the last common ancestor of *Agave* and *Hosta* and one within the lineage leading to *Chlorophytum.*
- *Conclusions*: We found that allopolyoidy and the origin of the *Yucca-Agave* bimodal karyotype co-occur on the same lineage consistent with the hypothesis that the bimodal karyotype is a consequence of allopolyploidy. We discuss this and alternative mechanisms for the formation of the *Yucca-Agave* bimodal karyotype. More generally, we illustrate how the use of next generation sequencing technology is a cost-efficient means for assessing genome evolution in non-model species.

Introduction

Karyotypes with bimodal chromosome size distributions have been described for taxa throughout monocotyledons (e.g. Bennett et al., 1992; Talavera et al., 1993; Gitaí et al., 2005). Whereas most karyotypes exhibit a continuous range of chromosome sizes, karyotypes of some taxa are bimodal with chromosomes falling into two distinct size classes often described as S for small and L for large (n = S+L). Chromosome size bimodality has arisen multiple times within Asparagales (Pires et al., 2006) including independent origins within Orchidaceae (Martínez, 1985; Giuseppina et al., 2010), Iridaceae (Goldblatt and Takei, 1997), Xanthorrhoeaceae (Taylor, 1925; Brandham and Doherty, 1998), Amaryllidaceae (Crosa, 2004) and Asparagaceae (McKelvey and Sax, 1933; Granick, 1944; Stedje, 1989). Bimodal karyotypes are most often limited to single genera (Goldblatt and Takei, 1997), small groups of closely related species (Stedje, 1996), or single species (Jones and Smith, 1967). However, bimodal chromosome size distributions are shared among multiple genera in Asphodeloideae (Xanthorrhoeaceae) and Agavoideae (Asparagaceae) (APG III, 2009; Chase et al., 2009). Within these subfamilies, bimodal karyotypes are synapomorphies for species-rich clades that may be millions of years old (Brandham and Doherty, 1998). For example within the Asphodeloideae, *Aloe*, *Astroloba*, *Gasteria* and *Haworthia* comprise ca. 689 species (The Plant List, 2010) and all exhibit a karyotype of n = 4S + 3L (Brandham, 1971). A clade composed solely of these genera is highly supported (Treutlein et al., 2003), indicating that chromosome bimodality is a synapomorphy and the ancestral karyotype for the group is n = 4S + 3L. The mechanism for persistence of bimodal karyotypes over millions of generations is unknown (Brandham and Doherty, 1998).

The clade defined as Agavaceae s.l. by Bogler et al. (2006) is composed of 15 genera and 377 species (The Plant List, 2010) sharing bimodal karyotypes (Fig 4.1). The APG III classification treats the group as an unnamed clade within Agavoideae, a subfamily of Asparagaceae (APG III, 2009; Chase et al., 2009). Here, we refer to this group as the Agavoideae bimodal karyotype clade or the ABK clade (Fig 4.1). Within the ABK clade, 10

genera (~358 species) have a karyotype of n = 25S + 5L (Akemine, 1935; Tamura, 1995) (or some multiple of this based on ploidy) and five genera, *Camassia* Lindl., *Chlorogalum* (Lindl.) Kunth, *Hastingsia, Schoenolirion* Torr. Ex Durand and *Hesperocallis* A. Gray, form a subclade of 19 species with bimodal karyotypes exhibiting varying numbers of large and small chromosomes (Gould, 1942; Cave, 1948; Sen, 1975; Fernández and Daviña, 1991; Tamura, 1995). A number of recent molecular phylogenetic analyses have placed *Camassia* Lindl., *Chlorogalum* (Lindl.) Kunth, and *Hastingsia* S.Watson, in a single clade (Bogler et al., 2006; Smith et al., 2008; Fishbein et al., 2010), and *Schoenolirion* Torr. ex Durand has long been associated with *Camassia* and *Chlorogalum* (e.g. Cronquist, 1981; Sherman and Beckling, 1991), although Halpin (2011) found support for placement of *Schoenolirion* outside of the *Camassia-Chlorogalum* clade. *Hesperocallis* has recently been placed within the ABK clade (Pires et al., 2004; Bogler et al., 2006), and whole plastome analyses place this monotypic genus in a clade with *Camassia* and *Chlorogalum* (McKain et al., in prep.).

Current analyses support *Hosta* Tratt., with a karyotype of n = 25S + 5L (Zonneveld and Iren, 2001), as sister to the rest of the ABK clade (Bogler and Simpson, 1996; Smith et al., 2008; Steele et al., submitted; McKain et al., unpublished manuscript), suggesting that the "*Yucca-Agave*" karyotype (Whitaker, 1934; Sato, 1935), was ancestral for the ABK clade. Divergence from the ancestral 25S + 5L karyotype appears to have occurred only within the clade including *Camassia*, *Chlorogalum*, *Hastingsia*, *Hesperocallis* and *Schoenolirion.* Otherwise, the "*Yucca-Agave*" karyotype has persisted throughout the ABK clade including within polyploid series where increases in ploidy coincide with proportionate gains in the number of small and large chromosomes (Robert et al., 2008).

Chromosome sizes are uniformly distributed in karyotypes for Agavoideae species outside of the ABK clade. Relationships within the sister clade to the ABK clade within Agavoideae remain elusive. Some studies place the genera *Behnia* Didr., *Herreria* Ruiz & Pav. and *Herreriopsis* H. Perrier within a clade sister to the former Anthericaceae sensu stricto. (Fig 4.1; Chase et al., 1996, Chase et al., 2006; Wurdack and Dorr, 2009), whereas others place

Behnia as sister to a clade including *Herreria*, *Herreriopsis* and the former Anthericaceae s.s. (Bogler et al., 2006; Pires et al., 2006; Kim et al., 2010). Some of the species within the former Anthericaceae s.s form a well-supported group with base chromosome numbers of x = 7,8 (Cave, 1948; Baldwin and Speese, 1951; Palomino and Romo, 1988; Bjorå et al., 2008). Karyotype information for many of the species in *Behnia*, *Herreria* and *Herreriopsis* is unavailable; however, one study did show that the chromosomes of *Herreria salsaparilha* Mart., n = 29, exhibit a uniform but broad size distribution (1.30-10.51µm) suggestive of fusion-fission events and potentially polyploidy (Gonçalves et al., 2007). *Anemarrhena asphodeloides* Bunge is sister to all other members of Agavoideae (Bogler et al., 2006; Pires et al., 2006; Kim et al., 2010, Steele et al., submitted, McKain et al., in prep.), and its karyotype (n = 11) exhibits a continuous range of chromosomes sizes (Rudall et al., 1998).

Two processes have been hypothesized to give rise to bimodal chromosome size distributions. The first is chromosome rearrangement involving fusion-fission events (Schudbert and Lysak, 2011), which has been hypothesized for bimodal karyotypes within Asparagales (Pires et al., 2006) and specifically for some members of the genus *Ornithogalum* (Asparagaceae, Scilloideae; Vosa, 1983, 2005). In these instances, large chromosomes could be the products of fusion between two smaller chromosomes, or the small chromosomes could be the result of the fission of large chromosomes. Such fusion-fission events have been attributed to genomic shock associated with an allopolyploid event (Wendel, 2000; Comai et al., 2003; Chen and Ni, 2006) followed by chromosomal rearrangements (e.g. Song et al., 1995; Pires, Zhao, et al., 2004).

A second hypothesized mechanism for bimodal karyotype formation is allopolyploidy involving parental species of different chromosome sizes. In this case, chromosomes have remained distinct following hybridization, segregating independently in the allopolyploid. Genomic in situ hybridization has been used to elucidate this mechanism in the grass (Poaceae) species *Milium montianum* Parl. (Bennett et al., 1992). This study identified *M. vernale* M. Bieb., or a closely related species, and a second, unknown species as progenitors of *M.*

montianum with the large chromosomes of the bimodal karyotype being identical in number and size to those of *M. vernale*. Chromosome bimodality in Asphodeloideae and Agavoideae has been suggested to have originated through this mechanism, although this hypothesis has not been tested in either case and parental species have not been identified (Brandham, 1983; Brandham and Doherty, 1998; Vosa, 2005; Pires et al., 2006).

In this study, we utilize next generation sequencing technology to sequence transcriptomes of strategically placed members of Agavoideae and test the hypothesis that the origin of chromosomal bimodality in this group coincides with a polyploid event. Numerous studies have evaluated divergence between duplicate genes as measured by the number of nonsynonymous substitutions per synonymous site (K_s) to identify whole genome duplications (i.e. polyploidy events) (Lynch and Conery, 2000; Blanc and Wolfe, 2004; Schlueter et al., 2004; Cui et al., 2006; Barker et al., 2008, 2009; Jiao et al., 2011) and concluded that genome duplications have been common throughout angiosperm evolution (Blanc and Wolfe, 2004; Schlueter et al., 2004; Cui et al., 2006; Soltis et al., 2009; Jiao et al., 2011; Van de Peer, 2011). Building on the approach of Jiao et al. (2011), we combined analyses of K_s plots and gene family phylogenies to test whether the origin of the bimodal karyotype on the lineage leading to the Agavoideae bimodal karyotype (ABK) clade is associated with a whole-genome duplication. In addition, the influence of polyploidy on the early diversification of the ABK clade is considered.

Materials and Methods

***Taxon sampling*—**The taxonomy of families in "core Asparagales" including the clade investigated in this study has changed in recent years with the publication of APG II (2003) and APG III (2009). APG II included Agavaceae, Anemarrhenaceae, Anthericaceae, Behniaceae and Herreriaceae (see Dahlgren et al., 1985) within Agavaceae, which was identified as an optional "bracketed" family within a broadly defined Asparagaceae. APG III (2009) abandoned the concept of bracketed families and applied the subfamilial classification Agavoideae (APG III,

2009; Chase et al., 2009). In order to simplify description of our sampling strategy, here we refer to the Agavoideae bimodal karyotype clade (ABK clade) (Fig 4.1) to describe the clade that Bogler et al. (2006) named Agavaceae s.l. The ABK clade includes all descendants of the last common ancestor of *Hosta* and *Agave* L.: *Agave*, *Beschorneria* Kunth, *Camassia*, *Chlorogalum*, *Furcraea* Vent., *Hastingsia*, *Hesperaloe* Engelm. in S.Watson, *Hesperoyucca* (Engelm.) Trel., *Hesperocallis, Hosta*, *Manfreda* Salisb., *Polianthes* L. , *Prochnyanthes* S.Watson, *Schoenolirion* and *Yucca* L. Based on current phylogenetic analyses, *Hosta* is sister to the rest of Agavaceae (Givnish et al., 2006; Smith et al., 2008, Steele et al., submitted, McKain et al., in prep.) (Fig 4.1). Therefore, within the ABK clade we generated transcriptome data for the diploid *Hosta venusta* F. Maek. (n=25S + 5L) (Zonneveld and Iren, 2001) and analyzed available EST data for *Agave tequilana* F.A.C Weber (Simpson et al., 2011), also a diploid. Transcriptome data were also generated for *Chlorophytum rhizopendulum* Bjorå & Hemp (Agavoideae, classically placed in polyphyletic Anthericaceae) as an exemplar for the sister clade to the ABK clade (Fig 4.1). Outside Agavoideae, transcriptome data were compiled for *Asparagus officinalis* L. (Asparagaceae) and *Leochilus labiatus* (Sw.) Kuntze (Orchidaceae) and combined with publically available EST data for *Allium cepa* L. (Amaryllidaceae) to root and identify ABK + *Chlorophytum* clades within gene trees.

RNA isolation and sequencing—RNA was isolated from fresh apical meristematic tissue or very young leaves using RNeasy Plant Mini Kit (Qiagen, Valencia, California, USA). All samples were kept on liquid nitrogen prior to RNA isolation. The optional step of heating the lysis solution to 65°C was used to maximize RNA yield. RNA was eluted into a final volume of 100 µL RNase-free water.

Total mass of RNA and quality was estimated using an Agilent 2100 Bioanalyzer (Agilent,
Santa Clara, California, USA). Samples were deemed acceptable if RIN scores were greater than 8.0. A minimum of 20 µg of total RNA was required for library building and sequencing.

RNA-Seq (Wang et al., 2009) paired-end libraries with average fragment lengths of 250 base pairs (bp) were constructed, and each library was sequenced on a single lane of an Illumina GAIIX sequencer flow cell (Illumina, San Diego, California, USA) at Cold Spring Harbor Laboratory to generate a minimum of 3 gigabases of 75 bp, paired-end sequences. Fastq sequence files for each taxon have been deposited in the Sequence Read Archive (SRA) database at NCBI (SRA study SRP009920).

***Transcriptome assembly*—**Illumina sequences were assembled using the CLC Genomics Workbench (CLC bio, Aarhus, Denmark). Prior to assembly, reads were trimmed to remove low quality ends, and trimmed reads shorter than 15 bp were discarded. Reads that passed the post-trimming length filter were then assembled using the default settings for *de novo* assembly, keeping only those contigs greater than 200 bp. Assemblies have been placed in a web-based searchable database (http://asparagalesdb.uga.edu).

K_s *plot estimation*—K_s frequency plots were used to initially detect potential whole genome duplication events and then to compare their origin to divergence of lineages within the ABK + *Chlorophytum* clade. Whereas studies attempting to identify more ancient genome duplications have assessed neutral sequence divergence based on transversions at four-fold degenerate sites (4DTv; e.g. Tuskan et al. 2006), we were focused on divergence following gene duplication (paralogs) or speciation (orthologs) events within Agavoideae over the last 50 million years (Good-Avila et al. 2006). Therefore, in this study, K_s was used as a measure of neutral sequence divergence (e.g. Lynch and Conery 2000, Blanc and Wolfe, 2004; Schlueter et al., 2004; Cui et al., 2006; Barker et al., 2008, 2009; Jiao et al., 2011). All-by-all BLASTN searches were performed and paralogous and orthologous pairs were identified as best matches within and between species, respectively. Paralog and putative ortholog matches with minimum alignment lengths of 300 bp and at least a 40% identity were analyzed further. These cutoffs were used to provide a minimum of 100 codons for alignments used in the estimation of the number synonymous substitutions per synonymous sites (K_s). Amino acid sequences were estimated for these homologs using ESTscan (Iseli et al., 1999; Lottaz et al., 2003), and paired

peptide sequences (orthologs or paralogs) were aligned using MUSCLE v3.7 (Edgar, 2004). Nucleotide sequences were then forced onto the amino acid alignments by codons (Cui et al., 2006, Suyama et al., 2006). Pair-wise K_s values were then calculated for each homolog pair using codeml within the PAML 4 package (Yang, 2007) paired sequence settings (yn00; Yang and Nielsen, 2000) and the F3x4 model (Goldman and Yang, 1994) for estimating codon frequencies.

Ks values were normalized for among-species differences in synonymous substitution rates for *Agave*, *Hosta*, and *Chlorophytum* genes in order to compare K_s plots for putative paralog pairs within each species and for putative orthologs between species. Putatively single copy genes were analyzed in order to estimate among-species variation in nuclear gene substitution rates. Ortholog sets for 49 single-copy genes were identified in the *Agave*, *Hosta*, *Chlorophytum*, *Asparagus*, *Allium*, and *Leochilus* transcriptome assemblies using blastx searches against a database of 970 genes inferred to be single-copy in sequenced angiosperm genomes (Wall et al. 2008, Duarte et al. 2010). Nucleotide sequences for transcripts were translated, and the amino acid sequences for each ortholog set were aligned using MUSCLE v3.7 (Edgar, 2004). Nucleotide coding sequences were then aligned to amino acid sequences using PAL2NAL v13 (Suyama et al., 2006). The resulting codon alignments were combined in a 56,372 column supermatrix, and a species tree was estimated using the GTR-Gamma model implemented in RAxML v7.0.4 (Stamatakis, 2006). The estimated tree matched previously inferred relationships (Steele et al., submitted) (Fig 4.2). K_s was estimated for each branch on the species tree using codeml (Yang, 1998; Yang and Nielsen, 1998). The cumulative K_s value for branches leading from the last common ancestor (LCA) of *Chlorophytum* and the ABK clade to the tips was lowest for *Hosta* (Fig 2). These LCA to tip K_s values estimated on the single-copy gene supermatrix, were used to make relative rate corrections of K_s values for *Agave* and *Chlorophytum* paralog-pairs. Corrections were calculated by multiplying uncorrected *Agave* and *Chlorophytum* putative paralog pair K_s values by the ratio of *Hosta/Agave* or *Hosta/Chlorophytum* LCA to tip K_s values derived from the single-copy gene analysis (Cui et al.,

2006). Similarly, *Ks* values for putative orthologs for each species pair were also normalized for differences in species-specific rates.

After normalization of the K_s values, frequency distributions for K_s values between 0.0 and 2.0 were plotted for putative paralogs within species and putative orthologs between each species pair. Only K_s values less than 2.0 were included in the plots because K_s estimates for more divergent gene pairs may be affected by saturation of substitution at synonymous sites.

Multivariate normal components were fit to the resulting K_s frequency distributions using the mixture model test implemented in EMMIX (McLachlan et al., 1999; http://www.maths.uq.edu.au/~gjm/emmix/emmix.html). The optimal number of components in the mixture model was identified using the Bayesian information criterion (BIC), and components were interpreted in terms of genome-scale duplication events and background single-gene duplications.

***Gene family circumscription and phylogeny estimation*—**Timing of WGD events relative to speciation events was deduced by comparing gene tree topologies to a species tree for taxa represented in each gene tree. Transcriptome assemblies for *H. venusta*, *A. tequilana*, *C. rhizopendulum*, *A. officinalis*, *A. cepa* and *L. labiatus* were filtered using a cut-off of 300 bp and then translated using TransPipe (Barker et al., 2010). Gene family circumscriptions were estimated by clustering of inferred amino acid sequences using OrthoMCL v2.0 (Li et al., 2003) with suggested parameter settings. Gene family clusters were selected for gene tree estimation based on the results of K_s plot analyses of paralogs for *Hosta*, *Agave*, and *Chlorophytum*. In order to test our interpretation of the K_s plots, gene tree analyses focused on gene families with gene pairs having K_s values corresponding to paralog peaks in estimated frequency plots (Fig 4.3). These gene family clusters included paralog pairs with K_s values between 0.1 and 0.3 for *H. venusta* and *A. tequilana,* and between 0.15 and 0.4 for *C. rhizopendulum*. When genes in paralog pairs were found in separate OrthoMCL clusters, the two clusters were combined for alignment and gene tree estimation.

Peptide sequences within each family were aligned using MUSCLE v3.7, and nucleotide sequences were aligned onto the amino acid alignments using PAL2NAL v13. Gappy alignments were filtered using two criteria. Columns in each alignment were removed if gaps were observed in more than 90% of the sequences (rows). Secondly, transcript assemblies (rows) were deleted if they covered less than 30% of the multiple sequence alignment's total length. Gene trees were estimated using RAxML v7.0.4 with the GTR + gamma evolutionary model and 500 bootstrap replicates. Gene trees were rooted using unigenes from the sampled outgroup taxa (*Leochilus*, *Allium* or *Asparagus*) found in each orthogroup.

Timing of gene duplication events relative to the origin of the ABK clade and divergence of lineages leading to ABK clade and *Chlorophytum rhizopendulum* were assessed by querying estimated gene trees for the last common ancestor of focal paralog pairs and their descendant genes using in-house perl scripts (available upon request). As described above, focus was placed on paralog pairs with K_s values corresponding to hypothesized genome duplication events (Fig 4.3). Gene tree topologies were inspected to determine whether individual duplication events occurred before divergence of the ABK clade and *Chlorophytum rhizopendulum* (H1), on the branch leading to the ABK clade (H2), on the branch leading to *Chlorophytum* (H3) or elsewhere on the species tree. For example, the last common ancestor of each *Hosta* and *Agave* paralog pair with K_s values between 0.1 and 0.3 were identified in rooted gene trees. If descendant genes included one or more *Chlorophytum* unigenes and the LCA of one of the putative paralogs and one of the *Chlorophytum* unigenes was sister to a clade with the other putative paralog, the duplication event was inferred to have occurred before divergence of *Chlorophytum* and the ABK clade. Alternatively, if *Chlorophytum* unigenes were placed as sister to a clade defined by the last common ancestor of focal paralog pair, the duplication event was inferred as having occurred after divergence of *Chlorophytum* and the ABK clade. Bootstrap percentages for the clade defined by the last common ancestor of a focal paralog pair were used to assess confidence in the inferred timing of duplication events relative to speciation events. The same approach was used to estimate the timing of *Chlorophytum* gene duplications relative to divergence of

Chlorophytum and the ABK clade. Gene tree topologies were also inspected manually to check the results of automated gene tree queries.

Results

***Assemblies*—**Assembly contig counts and lengths are shown in Table 4.1 for all contigs and successfully translated contigs (as estimated with TransPipe (Barker et al., 2010)). All successful translations were used for gene family estimation.

K_s *analyses of duplicate pairs*—Two paleopolyploid events were inferred from analyses of rate-normalized K_s values for comparisons of *H. venusta*, *A. tequilana*, and *C. rhizopendulum* homolog pairs (Fig 4.3). An analysis of 49 single copy genes in a supermatrix of 56,372 base pairs was used to estimate relative substitution rates at silent sites on branches leading to *H. venusta*, *A. tequilana*, and *C. rhizopendulum*. K_s values for branches leading to *Agave*, *Hosta*, and *Chlorophytum* from their most recent common ancestor were 0.174, 0.159, and 0.410, respectively. In order to compare the K_s frequency plots, K_s values for *Agave* and *Chlorophytum* were normalized by multiplying raw values by relative rate ratios of 0.915 and 0.389, respectively. Correction factors of 0.955 and 0.578 were applied to K_s values of putatively orthologous *Agave*/*Hosta* and *Chlorophytum*/*Hosta* gene pairs, respectively, to account for variation in synonymous substitution rates on lineages leading to each of these species.

After applying rate corrections, maximum K_s values for paralog pairs were 77.49, 136.41, and 52.04 for *H. venusta*, *A. tequilana*, and *C. rhizopendulum*, respectively, but to avoid effects of saturation, paralog pairs with K_s values over 2.0 were not included in frequency plots (Fig 4.3). K_s frequency plots include 437 gene duplicates for *H. venusta*, 2374 for *A. tequilana*, and 1704 for *C. rhizopendulum*. For comparison, uncorrected K_s frequency plots can be found in

Supplementary Figure 1. Due to its placement as the sister to the rest of the ABK clade (Fig 4.1), putative ortholog sets including *Hosta* are most informative for understanding the timing of gene duplications relative to the origin of the ABK clade. After correction, K_s values for cross-species homolog pairs ranged from 0.0 to 113.54 for *Hosta*/*Agave* and 0.0 to 46.48 for *Hosta/Chlorophytum* gene pairs. K_s plots for putative ortholog pairs with K_s values less than 2.0 included 1656 *Hosta-Agave* ortholog pairs and 3639 *Hosta-Chlorophytum* ortholog pairs.

Mixture model analyses reveal distinct components in the K_s frequency plots that we interpret as background single-gene duplications or polyploidy associated duplication events. The Bayesian information criterion (BIC) was used to choose the optimal number of normal distributions that fit data for each K_s plot based on the EMMIX output. The component with the smallest K_s values was interpreted as the background duplication. K_s frequency plots estimated from *Hosta* and *Agave* paralog pairs show a concentration of paralog pairs with modal K_s value around 0.2 (green lines in Fig 4.3A and 4.3B, respectively). A slightly larger mode of 0.25 is seen in *Chlorophytum* K_s plot (green line in Fig 4.3C). These peaks in the K_s distributions for all three species are suggestive of whole genome duplication events (WGD). The larger K_s value for the hypothesized duplication events in *C. rhizopendulum* may indicate an older but independent WGD on the branch leading to *C. rhizopendulum* (H3) or possibly a WGD on the lineage leading to the last common ancestor of *Chlorophytum* and members of the ABK clade (H1). Gene tree analyses were used to evaluate each of these alternative hypotheses (see below).

K_s frequency plots include at least one component representing ongoing single-gene duplication events (Fig 4.3A, 4.3B and 4.3C; denoted by blue lines). *Agave* and *Chlorophytum* K_s plots also show a population of putative paralog pairs with a normal distribution centered around $K_s \sim 0.025$ (Fig 4.3B and 4.3C; denoted by red lines). These components may represent

pairs of alleles or sequencing errors that resulted in assembly of distinct unigenes with high sequence identity. Identification of a separate low-K_s component in *Chlorophytum* K_s distribution may also be an artifact of rate correction. The relative-rate correction increased the number of pairs found in lowest K_s values (0.0-0.025) and may have resulted in the identification of a distinct low-K_s component in the mixture model analysis (compare Fig 4.3 and Fig.4.4). EMMIX identified a fourth component in the *Agave* K_s distribution centered on $K_s \sim 0.5$ (Fig 4.3B; denoted by the purple line). This group of paralog pairs may represent of a second, older WGD event that was not detected in the *Hosta* K_s distribution. Alternatively, background gene duplications may have been split into multiple components.

The K_s frequency plot for *Hosta-Agave* and *Hosta-Chlorophytum* homolog pairs show peaks of 0.1 and 0.25, respectively (Fig 4.3D). Using normalized K_s values as a proxy for time, these values suggest that the polyploid event inferred from the *Hosta* and *Agave* paralog pairs ($K_s \sim 0.20$) occurred before divergence of lineages leading to *Hosta* and *Agave* and after divergence of the ABK clade and *Chlorophytum* ($K_s \sim 0.25$; Fig 4.3D). It is important to note, however, that if this was an allopolyploid event, the estimated relative age of the paralog plots of the ABK clade would reflect timing of divergence of the parental species involved in the event rather than the actual WGD event (Doyle and Egan, 2010). This would not, however, affect our interpretation that the polyploid event occurred just before divergence of *Hosta* and *Agave*.

The peak in the *Chlorophytum* paralog K_s distribution described above, with a mode of 0.25 (Fig 4.3C), largely overlaps the putative *Hosta-Chlorophytum* ortholog peak (Fig 4.3D; $K_s \sim 0.25$), suggesting that the divergence of *Chlorophytum* and the ABK clade lineages occurred just before (H1) or after (H3) the WGD inferred from the *Chlorophytum* K_s plot. If H1 is correct, the paleopolyploid event in a common ancestor of the ABK clade and *Chlorophytum* may be masked

by the later event (H2) inferred from the *Hosta* and *Agave* K_s plots. We analyzed nuclear gene trees to more rigorously characterize the timing of gene and genome duplications relative to speciation events in Agavoideae.

Phylogenetic analysis of gene families—To further elucidate the timing of polyploidy in the species sequenced here, we conducted phylogenetic analyses of gene families that included the duplicated genes identified in the K_s analyses described above. A total of 12,724 putative gene families were circumscribed through OrthoMCL clustering of the transcripts assembled for the six species included in this study. We focused on OrthoMCL clusters containing *H. venusta*, *A. tequilana*, or *C. rhizopendulum* paralog pairs with K_s falling under the peaks interpreted as representing paleopolyploid events. These included 288, 1047 and 789 paralog pairs for *H. venusta*, *A. tequilana* and *C. rhizopendulum*, respectively. Genes were placed in separate OrthoMCL clusters for 81 of the 2124 paralog pairs considered. In these cases, the two OrthoMCL clusters were combined before conducting phylogenetic analyses. After combining these clusters, gene sets that did not contain at least three species were removed from further consideration. The 2124 paralog pairs identified in *H. venusta*, *A. tequilana* and *C. rhizopendulum* were distributed among 555 OrthoMCL gene sets that were aligned for gene tree estimation. All 555 alignments and ML gene trees have been deposited in the DRYAD database (http://dx.doi.org/10.5061/dryad.7pg045t2).

Relationships of genes within RAxML gene trees were analyzed within the context of species relationships (Fig 4.2) to assess the placement of paleopolyploid events inferred from the ABK clade and *Chlorophytum* K_s frequency plots. The species tree estimated from a supermatrix analysis of 49 putatively single-copy nuclear genes (Fig 4.2) is fully consistent with relationships

inferred from analyses of plastid genes (Steele et al., submitted) and is supported with bootstrap percentages of 100% at all nodes.

We queried gene tree topologies to determine whether gene trees supported duplication events on the branch leading to the last common ancestors of the ABK clade and *Chlorophytum* (H1), the branch leading to the ABK clade after divergence from *Chlorophytum* (H2) or the branch leading to *Chlorophytum* (H3) (Fig 4.2). These hypotheses are not mutually exclusive because the WGDs inferred through inspection of the K_s frequency plots could represent multiple events with homeologous gene pairs that have overlapping K_s distributions. Since our primary interest was in testing whether the origin of the *Yucca-Agave* bimodal karyotype coincided with a polyploid event, characterization of possible genome duplication events in our outgroup lineages or predating divergence of Agavoideae (including *Agave*, *Hosta* and *Chlorophytum*) and the lineage leading to *Asparagus* (the closest outgroup used here) were outside the scope of this study. Trees were disregarded when one or more outgroup transcripts were nested within the Agavoideae paralog clade while other outgroup transcripts rooted the clade including Agavoideae paralog pairs. These trees could reflect artifacts due to poor alignment or sparse gene sampling or duplication events predating divergence of the outgroup and ingroup (Agavoideae) taxa.

Of 555 gene trees that passed our filtering steps, 183 were informative for testing our hypotheses concerning phylogenetic placement of WGD events (i.e. > 50% bootstrap support duplication events concordant with H1, H2 or H2). Of these, 102 trees contained paralog pairs from either *Hosta* or *Agave* that were used to evaluate support for H1 or H2. There were 81 trees that contained paralog pairs from *Chlorophytum*, and these were used to evaluate support for H1 or H3. Bootstrap percentages (BP) supporting clades defined by the most recent common

ancestor of duplicate genes included in focal paralog pairs were used to evaluate the degree of support for one of the three hypotheses being considered. Trees were classified as providing 50 - 80 BP support or greater than 80 BP support for a given hypothesis.

Gene family trees that contained either *Hosta* or *Agave* paralog pairs showed clear evidence for an Agavaceae-specific ancestral WGD event. Of 102 informative trees, 54 exhibited greater than 80 BP support for H2 and another 30 showed at least 50 BP for gene duplication after divergence of *Chlorophytum* and ABK clade genes (Fig 4A). There were 18 trees that suggested gene duplication before the divergence of *Chlorophytum* and the ABK clade (consistent with H1), 13 of these with greater than 80 BP support.

Chlorophytum paralog pairs found in 81 gene trees showed clear support for duplication events after divergence of the ABK clade and *Chlorophytum*-lineage (H3; Fig 4.5B). There were 52 trees that supported H3 with at least 80 BP and another 19 with at least 50 BP. Gene duplications on the lineage leading to *Chlorophytum* and the ABK clade (H1) were seen in 10 trees, 8 with at least 80 BP. In sum, we interpret these results as favoring two separate WGD events in the lineages leading to the ABK clade and *Chlorophytum*. Whereas a bimodal karyotype is associated with the WGD on the lineage leading to the ABK clade, bimodal karyotypes have not been reported in *Chlorophytum* or related taxa within Anthericaceae (Chase et al., 1996), Behniaceae or Herreriaceae sensu APG II (2003).

Discussion

Polyploidy is a ubiquitous and recurring phenomenon in angiosperms and a recent study by Jiao et al. (2011) demonstrated that all flowering plants share at least two ancestral whole

genome duplications events. Detection of ancient polyploid events can be difficult as characteristics of recent polyploids, such the doubled of chromosome number or genome size, are typically lost over time (Devos et al., 2002; Leitch and Bennet, 2004), often rapidly (e.g. Song et al., 1995). K_s analyses have been used to characterize ancient polyploidization in a number of taxa (Blanc and Wolfe, 2004; Schlueter et al., 2004; Cui et al., 2006; Barker et al., 2008, 2009; Shi et al., 2010; Jiao et al., 2011). Peaks observed in K_s plots have been interpreted in the context of species relationships in an effort to identify shared and independent polyploid events and their evolutionary implications. However, interpretation of cross-species comparisons of K_s plots are complicated by the fact that substitution rates may vary among lineages. Previous work has shown that with sufficient taxon sampling, analyses of gene trees can resolve the relative timing of hypothesized paleopolyploidy in the face of variable substitution rates (Pfeil et al., 2005; Cui et al., 2006; Barker et al., 2008, 2009, Jiao et al., 2011). In this study, K_s analyses of *Agave*, *Hosta*, and *Chlorophytum* transcriptomes revealed evidence of paleopolyploidy, but timing of one or more WGD events was unclear based on K_s plots alone. We resolved this uncertainty through phylogenetic analysis of gene families including paralog pairs hypothesized to represent WGD events. These analyses confirmed the existence of two paleopolyploidizations within Agavoideae, one on the branch leading to the ABK and another within the *Chlorophytum* lineage. As has been seen in previous studies (Pfeil et al., 2005; Cui et al., 2006; Barker et al., 2008, 2009), variation in substitution rates among lineages led to ambiguity in interpretation of cross-species comparisons of analysis of uncorrected K_s plots (Fig. 4.4). A relative rate correction was applied to K_s values in an effort to resolve this ambiguity, but analysis of gene tree topologies provided the clearest evidence for timing of gene duplications relative to speciation events.

Phylogenetic analysis of gene families constructed from the six study species supported a WGD event having occurred on the same branch as the origin of the *Yucca-Agave* bimodal karyotype, which was first described over a century ago (reviewed in Whitaker, 1934; Sato, 1935; Granick, 1944) (H2; Fig 4.3A). In addition, gene trees constructed with *Chlorophytum* paralog pairs showed support for a second *Chlorophytum*-specific WGD event (H3, Fig 4.3B). Gene duplications were also evident on the branch leading to the last common ancestor of *Agave*, *Hosta*, and *Chlorophytum*, but these were much less common than duplications on branches leading either to the ABK clade or *Chlorophytum* (Fig 4.5).

Determining relationships of taxa, including genera, within Agavoideae has been a long-standing problem in plant systematics. The use of molecular markers in both the nuclear and plastid genomes have yielded strong support for monophyly of ABK clade but not for relationships between genera within the group (Eguiarte et al., 1994, 2000; Bogler and Simpson, 1995, 1996; Pires et al., 2004; Bogler et al., 2006). Phylogenetic analysis of whole plastid genome alignments has resulted in a well resolved tree with high support (McKain et al., in prep.), but short internodes separating basal nodes in the ABK clade suggest that the group diversified rapidly after its origin. The process of diploidization, including gene loss (i.e. fractionation (Freeling, 2009)), following polyploidization can spur reproductive isolation and speciation (Werth and Windham, 1991; Lynch and Force, 2000; Taylor et al., 2001; Scannell et al. 2006). This process has been hypothesized as a driver of angiosperm diversification (Soltis et al. 2009). Ecological factors, including range expansion, colonization of arid habitats, and plant-pollinator interactions, are thought to have contributed to diversification of the ABK clade (Good-Avila et al., 2006), but divergence in genome structure following polyploidization may have also played role in the earliest speciation events in this group.

In this context, it is noteworthy that K_s values for hypothesized *Agave* and *Hosta* homeologs (~0.2; Figs. 4.3A and 4.3B) are significantly greater than those for putative orthologs (~0.1; Fig 4.3D), which may be due to a gap between polyploidization and the inferred radiation in the early history of the ABK clade. Alternatively, if the radiation was spurred by an allopolyploid event, large differences between ABK clade paralog and ortholog K_s values (Fig 4.3) may be due to divergence between parental genomes before hybridization.

Distinguishing between autopolyploidy and allopolyploidy, especially in ancient polyploid events, can be quite difficult and at times impossible (Doyle and Egan, 2010). Here, we have demonstrated that a paleopolyploid event occurred on the lineage leading to the ABK clade after the divergence of the *Chlorophytum* clade, consistent with the hypothesis that the *Yucca-Agave* bimodal karyotype originated with an allopolyploid event. If the last common ancestor of the ABK clade was in fact an allopolyploid, however, the progenitor species seem to be extinct and it is not possible to definitively distinguish between autopolyploidy and allopolyploidy through gene tree analyses (Doyle and Egan, 2010). Future work will test whether homeologous gene pairs are consistently segregating on small and large chromosomes as would be expected if the last common ancestor of the ABK clade was an allopolyploid hybrid of now extinct parental species with small and large chromosomes.

The origin of chromosome bimodality in Agavoideae has been under investigation since the *Yucca-Agave* karyotype was first described, though the character was long considered diagnostic for this group (e.g. Whitaker, 1934). The 5L + 25S karyotype has long been viewed as diagnostic for evolutionary affiliations with *Yucca* and *Agave* (e.g. *Hosta*; Whitaker, 1934). Further, when combined with embryological and other morphological characters, the cytogenetic analyses led Cave (1948) to posit that *Hesperocallis* with 4 large, 2 medium large, and 18 small

chromosomes is allied with *Hosta*, *Yucca* and *Agave*, a hypothesis that would gain molecular support 60 years later (Pires et al., 2004). The objective of this study was to test whether a paleopolyploid event was associated with the origin of the 5L + 25S karyotype. The results are consistent with the hypothesis that the last common ancestor of the ABK clade was an allopolyploid. Whereas the bimodal karyotype is suggestive of allopolyploidy, the possibility of chromosomal fusion and fission cannot be discounted.

The utility of next generation sequencing to gain insight into the genomes and evolutionary history of nonmodel species is obvious. The next-generation sequence data presented here allowed us to assess the plausibility of a long-standing hypothesis that relates the origin of chromosome bimodality to polyploidy. This work will aid in the understanding of the evolution of Agavoideae while providing an improved framework for future phylogenetic, ecological and crop improvement studies. There is also great potential for investigating bimodal karyotypes across Asparagales and their implications for understanding causes and consequences of polyploidy.

Acknowledgements

The authors thank E. Wafula (Penn State) for help with the transcriptome assembly. We are thankful for the comments of three anonymous reviewers that greatly improved this manuscript. We are also grateful to the National Science Foundation for support this research through an Assembling the Tree of Life (DEB 0829868) grant to J.L-M., C.W.D. and J.C.P., and a Doctoral Dissertation Improvement Grant to M.R.M. and J.L-M. The quality of the manuscript was greatly improved due to the suggestions of the editors of this special issue and anonymous reviewers.

References

AGP 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.

AKEMINE, T. 1935. Chromosome studies on *Hosta* I. The chromosome numbers in various species of *Hosta*. *Journal of the Faculty of Science, Hokkaido Imperial University. Ser. 5, Botany* 5(1): 25–32.

BALDWIN, J.T. AND B.M. SPEESE. 1951. Cytogeography of *Chlorophytum* in Liberia. *American Journal of Botany* 38: 153-156.

BARKER, M.S., K.M. DIUGOSCH, L. DINH, R.S. CHALLA, N.C. KANE, M.G. KING AND L.H. RIESEBERG. 2010. EvoPipes.net: Bioinformatic tools for ecological and evolutionary genomics. *Evolutionary Bioinformatics Online* 6: 143-149.

BARKER, M.S., N.C. KANE, M. MATVIENKO, A. KOZIK, R.W. MICHELMORE, S.J. KNAPP, AND L.H. RIESEBERG. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution* 25: 2445 -2455.

BARKER, M.S., H. VOGEL, AND M.E. SCHRANZ. 2009. Paleopolyploidy in the Brassicales: analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. *Genome Biology and Evolution* 1: 391-399.

BENNETT, S.T., A.Y. KENTON, AND M.D. BENNETT. 1992. Genomic in situ hybridization reveals the allopolyploid nature of *Milium montianum* (Gramineae). *Chromosoma* 101(7): 420-424.

BJORÅ, C.S., A. HEMP, G. HOELL, AND I. NORDAL. 2008. A taxonomic and ecological analysis of two forest *Chlorophytum* taxa (Anthericaceae) on Mount Kilimanjaro, Tanzania. *Plant Systematics and Evolution* 274: 243-253.

BLANC, G., AND K.H. WOLFE. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant Cell* 16: 1667-1678.

BOGLER, D.J., J.C. PIRES, AND J. FRANCISCO-ORTEGA. 2006. Phylogeny of Agavaceae based on *ndhF*, *rbcL*, and ITS sequences: implications of molecular data for classification. *Aliso* 22: 313–328.

BOGLER, D.J., AND B.B. SIMPSON. 1995. A chloroplast DNA study of the Agavaceae. *Systematic Botany* 20: 191–205.

BOGLER, D.J., AND B.B. SIMPSON. 1996. Phylogeny of Agavaceae based on ITS rDNA sequence variation. *American Journal of Botany* 83: 1225–1235.

BRANDHAM, P.E. 1983. Evolution in a stable chromosome system. In *Kew Chromosome Conference II*: 251-260.

BRANDHAM, P.E. 1971. The chromosomes of the Liliaceae: II: Polyploidy and karyotype variation in the Aloïneae. *Kew Bulletin* 25: 381-399.

BRANDHAM, P.E., AND M.J. DOHERTY. 1998. Genome size variation in the Aloaceae, an angiosperm family displaying karyotypic orthoselection. *Annals of Botany* 82: 67-73.

CAVE, M.S. 1948. Sporogenesis and embryo sac development of *Hesperocallis* and *Leucocrinum* in relation to their systematic position. *American Journal of Botany* 35: 343–349.

CHASE, M.W., M.F. FAY, D.S. DEVEY, O. MAURIN, N. RØNSTED, T.J. DAVIES, Y. PILLON, ET AL. 2006. Multigene analyses of monocot relationships: A summary. *Aliso* 22: 63–75.

CHASE, M.W., J.L. REVEAL, AND M.F. FAY. 2009. A subfamilial classification for the expanded asparagalean families Amaryllidaceae, Asparagaceae and Xanthorrhoeaceae. *Botanical Journal of the Linnean Society* 161: 132-136.

CHASE, M.W., P.J. RUDALL AND J.G. CONRAN. 1996. New circumscriptions and a new family of Asparagoid lilies: genera formerly included in Anthericaceae. *Kew Bulletin* 51: 667-680.

CHEN, Z.J., AND Z. NI. 2006. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *BioEssays* 28: 240-252.

COMAI, L., A. MADLUNG, C. JOSEFSSON, AND A. TYAGI. 2003. Do the different parental "heteromes" cause genomic shock in newly formed allopolyploids? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 358: 1149-1155.

CRONUIST, A. 1981. An integrated system of classification of flowering plants. Columbia University Press, New York, New York.

CROSA, O. 2004. Segunda especie y justificación del género *Zoellnerallium* (Alliaceae). *Darwiniana* 42: 165–168.

CUI, L., P.K. WALL, J.H. LEEBENS-MACK, B.G. LINDSAY, D.E. SOLTIS, J.J. DOYLES, P.S. SOLTIS, ET AL. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Research* 16: 738-749.

DAHLGREN, R.M.T., H.T. CLIFFORD, AND P.F. YEO. 1985. The families of the monocotyledons: structure, evolution, and taxonomy. Springer-Verlag, Berlin, Germany.

DEVOS, K.M., J.K.M. BROWN AND J.L. BENNETZEN. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Research* 12: 1075-1079.

DOYLE, J.J. AND A.N. EGAN. 2010. Dating the origins of polyploidy events. *New Phytologist* 186: 73-85.

DUARTE, J.M., P.K. WALL, P.P. EDGER, L.L. LANDHERR, H. MA, J.C. PIRES, J. LEEBENS-MACK, AND C.W. DEPAMPHILIS. 2010. Identification of shared single copy nuclear genes in *Arabidopsis, Populus, Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology* 10:61.

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

EGUIARTE, L.E., M.R. DUVALL, G.H. LEARN JR, AND M.T. CLEGG. 1994. The systematic status of the Agavaceae and Nolinaceae and related Asparagales in the monocotyledons: An analysis based on the *rbcL* gene sequence. *Boletín de la Sociedad Botánica de México* 54: 35–56.

EGUIARTE, L.E., V. SOUZA, AND A. SILVA MONTELLANO. 2000. Evolución de la familia Agavaceae: filogenia, biología reproductiva y genética de poblaciones. *Boletín de la Sociedad Botánica de México* 66: 131–151.

FERNÁNDEZ, A., AND J.R. DAVIÑA. 1991. Heterochromatin and genome size in *Fortunatia* and *Camassia* (Hyacinthaceae). *Kew Bulletin* 46: 307–316.

FISHBEIN, M., S.R. KEPHART, M. WILDER, K.M. HALPIN, AND S.L. DATWYLER. 2010. Phylogeny of *Camassia* (Agavaceae) inferred from plastid *rpl16* intron and *trnDtrnYtrnEtrnT* intergenic spacer DNA sequences: implications for species delimitation. *Systematic Botany* 35: 77–85.

FREELING, M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual Review of Plant Biology* 60*:* 433–453.

GITAÍ, J., R. HORRES, AND A.M. BENKO-ISEPPON. 2005. Chromosomal features and evolution of Bromeliaceae. *Plant Systematics and Evolution* 253: 65-80.

GIUSEPPINA, B., C. BRULLO, S. PULVIRENTI, A. SCRUGLI, M.C. TERRASI, AND S. D'EMERICO. 2010. Advances in chromosomal studies in Neottieae (Orchidaceae): constitutive heterochromatin, chromosomal rearrangements and speciation. *Caryologia* 63(2): 184-191.

GIVNISH, T.J., J.C. PIRES, S.W. GRAHAM, M.A. MCPHERSON, L.M. PRINCE, T.B. PATTERSON, H.S. RAI, ET AL. 2006. Phylogenetic relationships of monocots based on the highly informative plastid gene *ndhF*: Evidence for widespread concerted convergence. *Aliso* 22: 28–51.

GOLDBLATT, P., AND M. TAKEI. 1997. Chromosome cytology of Iridaceae-patterns of variation, determination of ancestral base numbers, and modes of karyotype change. *Annals of the Missouri Botanical Garden* 84(2): 285-304.

GOLDMAN, N., AND Z. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* 11: 725 -736.

GONÇALVES, L. DE A., W.R. CLARINDO, C.R. DE CARVALHO, AND W.C. OTONI. 2007. Cytogenetics and flow cytometry-based DNA quantification in *Herreria salsaparilha* Martius (Herreriaceae): A medicinal species. *Cytologia* 72: 295-302.

GOOD-AVILA, S.V., V. SOUZA, B.S. GAUT, AND L.E. EGUIARTE. 2006. Timing and rate of speciation in *Agave* (Agavaceae). *Proceedings of the National Academy of Sciences* 103: 9124 -9129.

GOULD, F.W. 1942. A systematic treatment of the genus *Camassia* Lindl. *American Midland Naturalist* 28: 712-742.

GRANICK, E.B. 1944. A karyosystematic study of the genus *Agave*. *American Journal of Botany* 31(5): 283–298.

HALPIN, K.M. 2011. A chloroplast phylogeny of Agavaceae subfamily Chlorogaloideae with a focus on species relationships in *Hastingsia*. M.S. dissertation, Oklahoma State University, Stillwater, Oklahoma, USA.

ISELI, C., C.V. JONGENEEL, AND P. BUCHER. 1999. ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*: 138–148.

JIAO, Y., N.J. WICKETT, S. AYYAMPALAYAM, A.S. CHANDERBALI, L. LANDHERR, P.E. RALPH, L.P. TOMSHO, ET AL. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97-100.

JONES, K., AND J.B. SMITH. 1967. The chromosomes of the Liliaceae: I: The karyotypes of twenty-five tropical species. *Kew Bulletin* 21(1): 31-38.

KIM, J.-H., D.-K. KIM, F. FOREST, M.F. FAY, AND M.W. CHASE. 2010. Molecular phylogenetics of Ruscaceae sensu lato and related families (Asparagales) based on plastid and nuclear DNA sequences. *Annals of Botany* 106: p.775 -790.

LEITCH, I.J. AND M.D. BENNET. 2004. Genome downsizing in polyploid plants. *Biological Journal of the Linnean Society* 82:651-663.

LI, L., C.J. STOECKERT AND D.S. ROOS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research* 13: 2178-2189.

LOTTAZ, C., C. ISELI, C.V. JONGENEEL, AND P. BUCHER. 2003. Modeling sequencing errors by combining Hidden Markov models. *Bioinformatics* 19: 103-112.

LYNCH, M. AND J. S. CONERY. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151–1155.

LYNCH, M. AND A. FORCE. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154: 459-473.

MARTÍNEZ, A. 1985. The chromosomes of Orchids VIII. Spiranthinae and Cranichidinae. *Kew Bulletin* 40(1): 139-147.

MCKELVEY, S.D. AND K. SAX. 1933. Taxonomic and cytological relationships of *Yucca* and *Agave*. *Journal of the Arnold Arboretum* 14: 76–81.

MCLACHLAN, G.J., D. PEEL, K.E. BASFORD, AND P. ADAMS. 1999. The EMMIX software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software* 4: 1–14.

PAGE, R.D.. 2002. Visualizing phylogenetic trees using TreeView. *Current Protocols in Bioinformatics*.

PALOMINO, G. AND V. ROMO. 1988. Karyotypic studies in two Mexican species of *Echeandia* Ort. (Liliaceae). *The Southwestern Naturalist* 33: 382-384.

PFEIL, B.E., J.A. SCHLUETER, R.C. SHOEMAKER, AND J.J. DOYLE. 2005. Placing paleopolyploidy in relation to taxon divergence: A phylogenetic analysis in legumes using 39 gene families. *Systematic Biology* 54: 441 -454.

PIRES, J.C., I.J. MAUREIRA, T.J. GIVNISH, K.J. SYTSMA, O. SEBERG, G. PETERSEN, J.I. DAVIS, ET AL. 2006. Phylogeny, genome size, and chromosome evolution of Asparagales. *Aliso* 22: 287–304.

PIRES, J.C., I.J. MAUREIRA, J.P. REBMAN, G.A. SALAZAR, L.I. CABRERA, M.F. FAY AND M.W. CHASE. 2004. Molecular data confirm the phylogenetic placement of the enigmatic *Hesperocallis* (Hesperocallidaceae) with Agave. *Madroño* 51: 307–311

PIRES, J.C., J. ZHAO, M.E. SCHRANZ, E.J. LEON, P.A. QUIJADA, L.N. LUKENS AND T.C. OSBORN. 2004. Flowering time divergence and genomic rearrangements in resynthesized *Brassica* polyploids (Brassicaceae). *Biological Journal of the Linnean Society* 82: 675–688.

PLANT LIST. 2010. The plant list, version 1 [online]. Available at: http://www.theplantlist.org/ [Accessed September 8, 2012].

ROBERT, M.L., K.Y. LIM, L. HANSON, F. SANCHEX-TEYER, M.D. BENNETT, A.R. LEITCH AND I.J. LEITCH. 2008. Wild and agronomically important Agave species (Asparagaceae) show proportional increases in chromosome number, genome size, and genetic markers with increasing ploidy. *Botanical Journal of the Linnean Society* 158: 215–222.

RUDALL, P.J., E.M. ENGLEMAN, L. HANSON, AND M.W. CHASE. 1998. Embryology, cytology and systematics of *Hemiphylacus*, *Asparagus* and *Anemarrhena* (Asparagales). *Plant Systematics and Evolution* 211: 181-199.

SATO, D. 1935. Analysis of the karyotypes in *Yucca*, *Agave* and the related genera with special reference to the phylogenetic significance. *The Japanese Journal of Genetics* 11: 272–278.

SCANNELL, D.R., K.P. BYRNE, J.L. GORDON, S. WONG AND K.H. WOLFE. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440: 341-345.

SCHLUETER, J.A., P. DIXON, C. GRANGER, D. GRANT, L. CLARK, J.J. DOYLE, R.C. SHOEMAKER. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* 47: 868-876.

SCHUBERT, I. AND M. LYSAK. 2011. Interpretation of karyotype evolution should consider structural constraints. *Genetics* 27: 207-216.

SEN, S. 1975. Cytotaxonomy of Liliales. *Feddes Repertorium* 86: 255–305.

SHERMAN, H., AND R. BECKLING. 1991. The generic distinctness of *Schoenolirion* and *Hastingsia*. *Madroño* 38: 130-138.

SHI, T., H. HUANG, AND M.S. BARKER. 2010. Ancient genome duplications during the evolution of kiwifruit (*Actinidia*) and related Ericales. *Annals of Botany* 106: 497 -504.

SIMPSON, J., A. MARTÍNEZ HERNÁNDEZ, M. JAZMÍN ABRAHAM JUÁREZ, S. DELGADO SANDOVAL, A. SÁNSCHEZ VILLARREAL, AND C. CORTÉS ROMERO. 2011. Genomic resources and transcriptome mining in *Agave tequilana*. *Global Change Biology Bioenergy* 3: 25-36.

SMITH,C.I., O. PELLMYR, D.M. ALTHOFF, M. BALCÁZAR-LARA, J. LEEBENS-MACK, AND K.A. SEGRAVES. 2008. Pattern and timing of diversification in *Yucca* (Agavaceae): specialized pollination does not escalate rates of diversification. *Proceedings of the Royal Society B: Biological Sciences* 275: 249-258.

SOLTIS, D.E., V.A. ALBER, J. LEEBENS-MACK, C.D. BELL, A.H. PATERSON, C. ZHENG, D. SANKOFF, ET AL. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* 96: 336-348.

SONG, K., P. LU, K. TANG, AND T.C. OSBORN. 1995. Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. *Proceedings of the National Academy of Sciences* 92: 7719-7723.

STAMATAKIS, A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22: 2688-2690.

STEDJE, B. 1989. Chromosome evolution within the *Ornithogalum tenuifolium* complex (Hyacinthaceae), with special emphasis on the evolution of bimodal karyotypes. *Plant Systematics and Evolution* 166: 79-89.

STEDJE, B. 1996. Karyotypes of some species of Hyacinthaceae from Ethiopia and Kenya. *Nordic Journal of Botany* 16: 121–126.

SUYAMA, M., D. TORRENTS, AND P. BORK. 2006. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* 34: W609-W612.

TALAVERA, S., P. GARCÍA-MURILLO, AND J. HERRERA. 1993. Chromosome numbers and a new model for karyotype evolution in *Ruppia* L. (Ruppiaceae). *Aquatic Botany* 45: 1-13.

TAMURA, M.N. 1995. A karyological review of the orders Asparagales and Liliales (Monocotyledonae). *Feddes Repertorium* 106: 83-111.

TAYLOR, W.R. 1925. Cytological studies on *Gasteria*. II. A comparison of the chromosomes of *Gasteria*, *Aloë*, and *Haworthia*. *American Journal of Botany* 12(4): 219-223.

TAYLOR, J.S., Y. VAN DE PEER AND A. MEYER. 2001. Genome duplication, divergent resolution and speciation. *Trends in Genetics* 17: 299-301.

TREUTLEIN, J., G.F. SMITH, B.E. VAN WYK, AND M. WINK. 2003. Phylogenetic relationships in Asphodelaceae (subfamily Alooideae) inferred from chloroplast DNA sequences (*rbcL*, *matK*) and from genomic fingerprinting (ISSR). *Taxon* 52: 193–207.

TUSKAN, G.A., S. DIFAZIO, S. JANSSON, J. BOHLMANN, I. GRIGORIEV, U. HELLSTEN, N. PUTANM, ET AL. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313: 1596-1604.

VAN D PEER, Y. 2011. A mystery unveiled. *Genome Biology* 12: 113.

VOSA, C.G. 1983. Chromosome evolution in *Ornithogalum*. In *Kew Chromosome Conference II*: 370.

VOSA, C.G. 2005. On chromosome uniformity, bimodality and evolution in the tribe Aloineae (Asphodelaceae). *Caryologia* 58: 83–85.

WALL, P.K., J. LEEBENS-MACK, K.F. MÜELLER, D. FIELD, N.S. ALTMAN AND C.W. DEPAMPHILIS. 2008. PlantTribes: A gene and gene family resource for comparative genomics in plants. *Nucleic Acids Research* 36: D970-D976.

WANG, Z., M. GERSTEIN, AND M. SNYDER. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10: p.57-63.

WENDEL, J.F. 2000. Genome evolution in polyploids. *Plant Molecular Biology* 42: 225–249.

WERTH, C.R. AND M.D. WINDHAM. 1991. A model for divergent, allopatric speciation of polyploid Pteridophytes resulting from silencing of duplicate-gene expression. *American Naturalist* 137: 515-526.

WHITAKER, T.W. 1934. Chromosome constitution in certain monocotyledons. *Journal of the Arnold Arboretum* 15: 135–153.

WURDACK, K.J.J. AND L.J. DORR. 2009. The South American genera of Hemerocallidaceae (*Eccremis* and *Pasithea*): Two introductions to the New World. *Taxon* 58: 1122-1134.

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

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

YANG, Z. AND R. NIELSEN. 1998. Models of amino acid substitution and appliations to mitochondrial protein evolution. *Molecular Biology and Evolution* 15: 1600-1611.

YANG, Z., AND R. NIELSEN. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* 17: 32-43.

ZONNEVELD, B.J.M., AND F. IREN. 2001. Genome size and pollen viability as taxonomic criteria: Application to the genus *Hosta*. *Plant Biology* 3: 176-185.

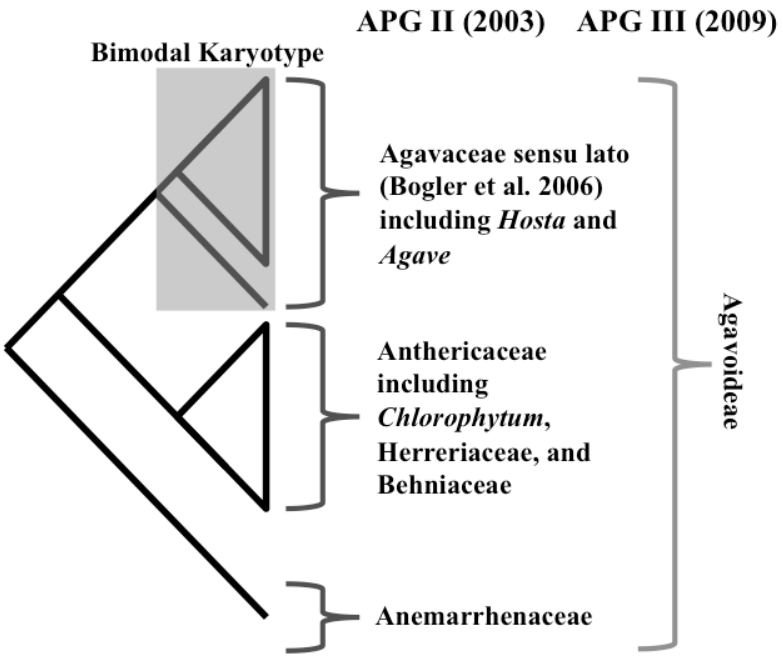

Figure 4.1. The currently accepted phylogeny of Agavoideae (Asparagaceae) with both APG II and APG III nomenclature.

The grey box indicates the former Agavaceae (senus Bogler et al. 2006), which we define as the Agavoideae bimodal karyotype clade. The sister clade to the ABK clade includes the former Anthericaceae, Herreriaceae and Behniaceae

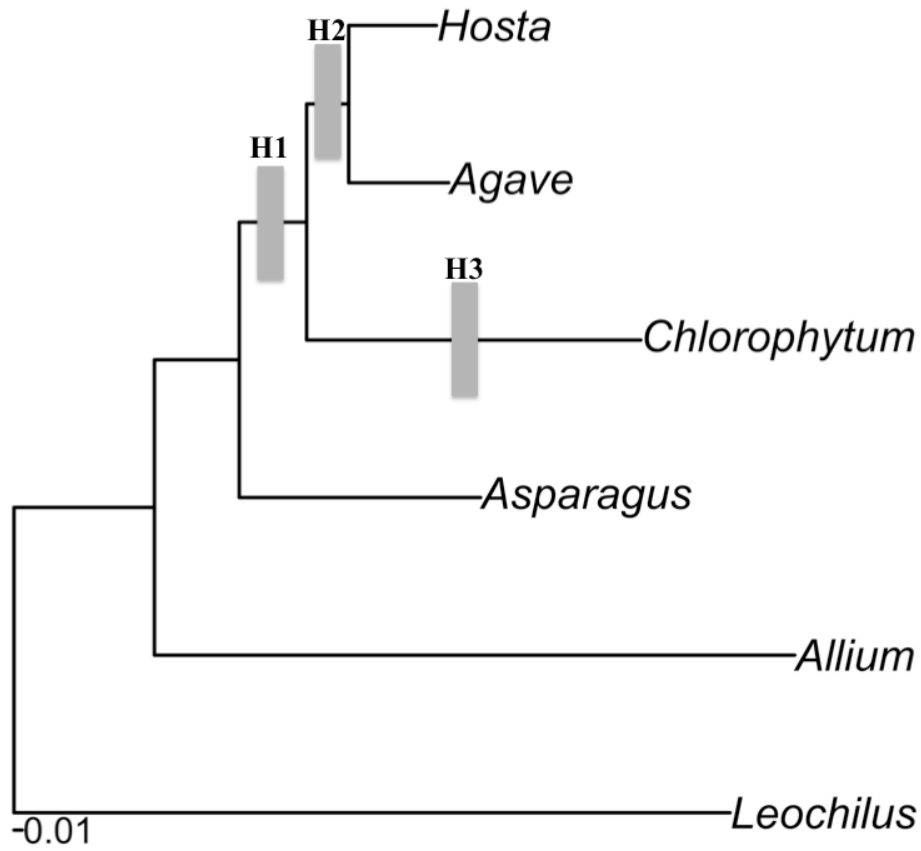

Figure 4.2. The maximum likelihood species tree derived from a supermatrix analysis of putatively single-copy genes

The maximum likelihood species tree derived from a supermatrix analysis of putatively single-copy genes (Duarte et al. 2010) extracted from transcriptome data. All nodes have 100% bootstrap values and branch lengths are represented by K_s values. The hypothesized timing of genome duplication with the Agavoideae are marked as H1, H2 and H3 (see Methods).

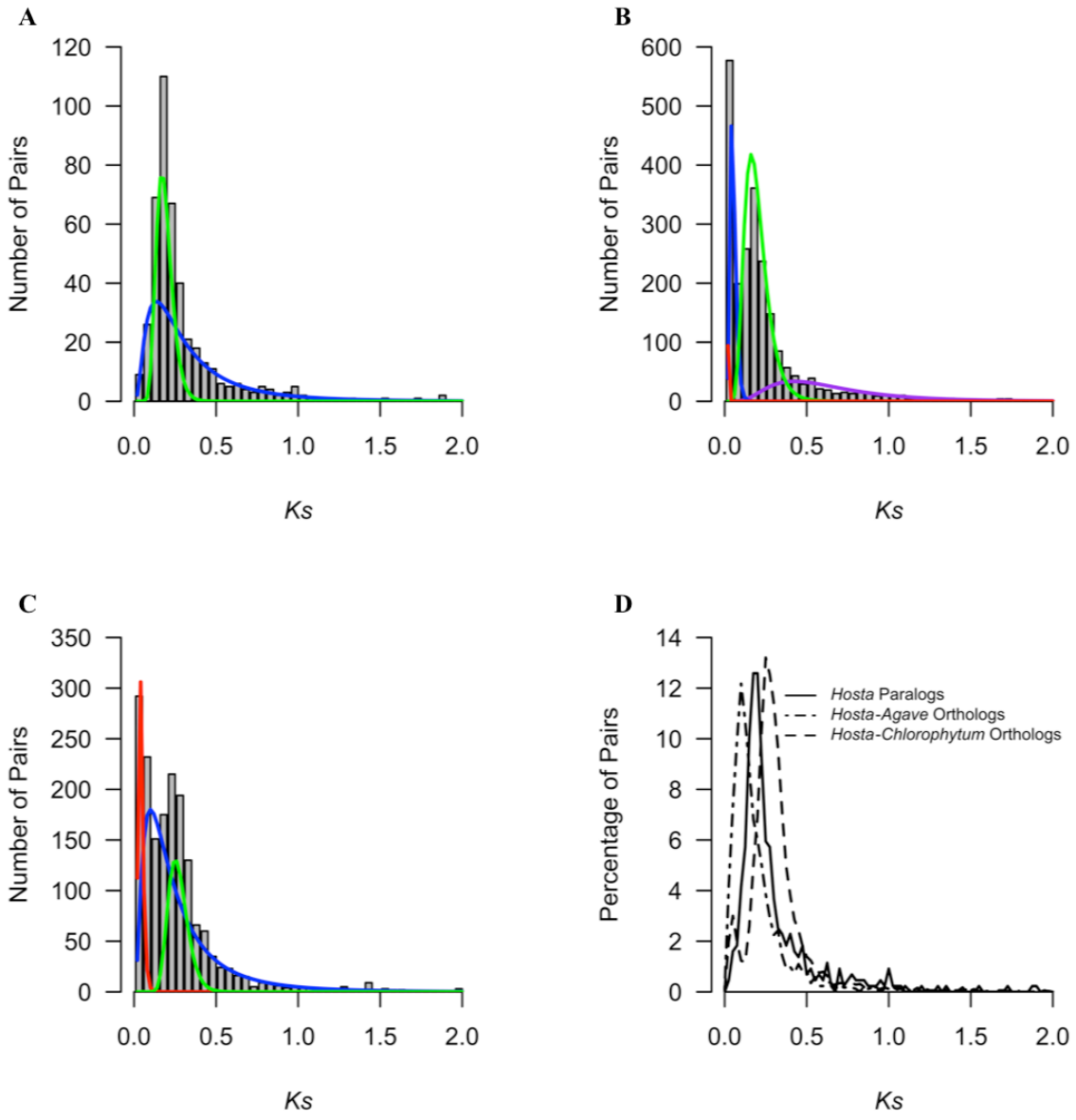

Figure 4.3. K_s frequency plots (corrected for rate variation)

Normalized K_s frequency plots (corrected for rate variation; see Results) for paralogous and orthologous duplicate pairs from *Hosta venusta*, *Agave tequilana*, and *Chlorophytum rhizopendulum* derived from transcriptome data. K_s distribution components estimating using EMMIX (see Methods) are superimposed on histograms for each paralog-pair K_s plot (A-C). These components are hypothesized to represent background gene duplications (blue), gene

duplications associated with polyploidy events (gene), allele or sequencing errors that resulted in assembly of distinct unigenes with high sequence identity (red), and older gene duplications in the *A. tequilana* genome (purple; see Results). (A) *H. venusta* and (B) *A. tequliana* paralog plots include secondary peaks (green lines) with modal K_s value indicative of a whole genome duplication event at $K_s = 0.2$. The *C. rhizopendulum* paralog plot (C) shows a secondary peak with a mode at $K_s = 0.25$. (D) The K_s distribution for *H. venusta* paralogs (solid line) exhibits a mode between modal K_s values for putative *Hosta-Agave* ($K_s = 0.1$) and *Hosta-Chlorophytum* ($K_s = 0.25$) orthologs.

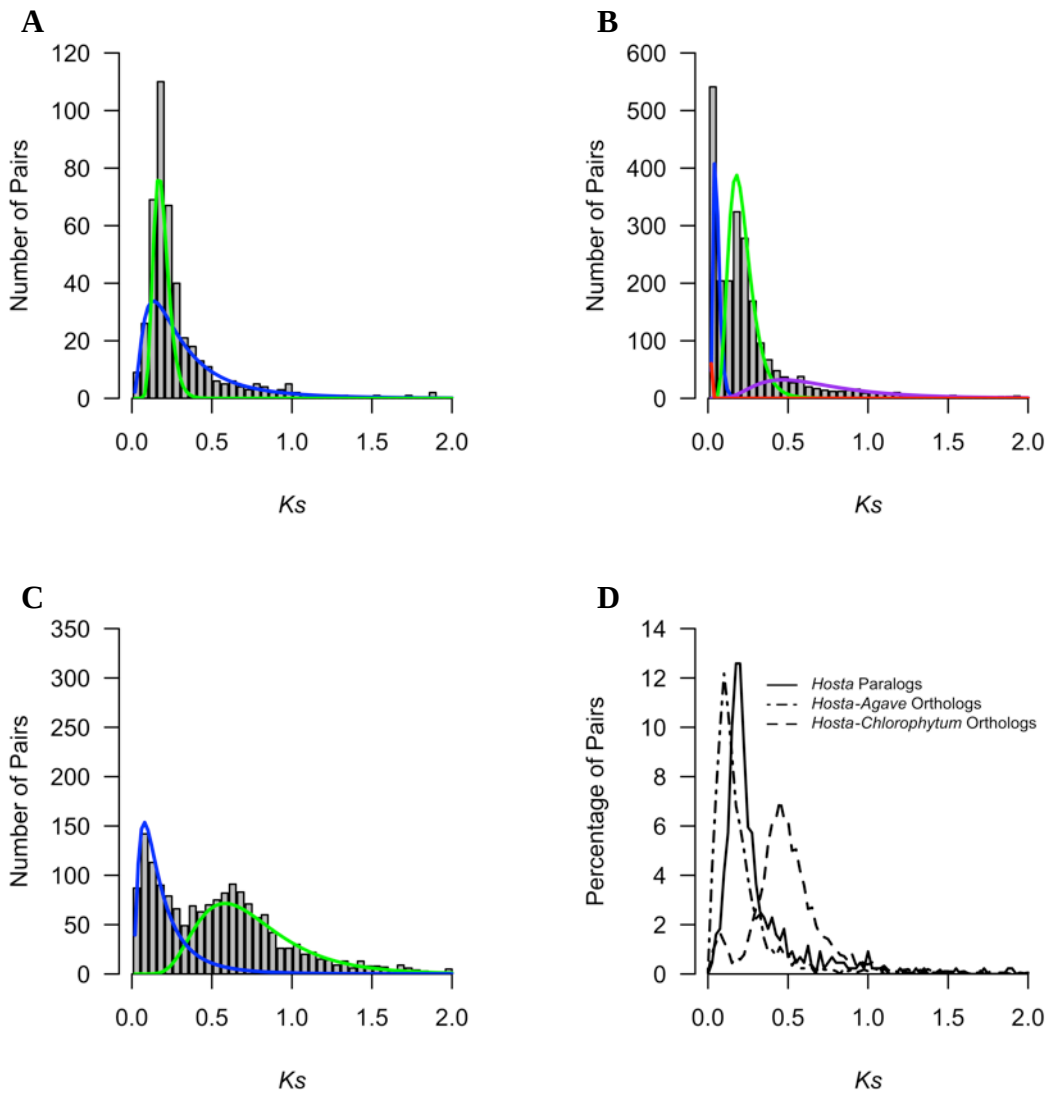

Figure 4.4. K_s frequency plots (uncorrected for rate variation; see Methods) for paralogous and orthologous duplicate pairs from *Hosta venusta*, *Agave tequilana*, and *Chlorophytum rhizopendulum* derived from transcriptome data. K_s value population distributions identified using the EMMIX mixture model analysis are shown for each paralog-pair K_s plot (A-C): blue lines depict background gene duplication, green lines depict duplications attributed to hypothesized whole genome duplications, red lines depict pairs of alleles or sequencing errors that resulted in assembly of distinct unigenes with high sequence identity, and purple lines depict either an older duplication event or a portion of the background gene duplications. *H. venusta*

(A) and *A. tequliana* (B) paralog plots include secondary peaks with modal K_s value indicative of a whole genome duplication event at K_s = 0.2 and 0.25, respectively. The *C. rhizopendulum* paralog plot (C) shows a secondary peak with a mode at K_s = 0.6. (D) *H. venusta* paralog K_s values plotted with *Hosta-Agave* and *Hosta-Chlorophyutm* best blast hit pair K_s values show that the K_s peak for *H. venusta* paralogs is lower than the *Hosta-Chlorophytum* ortholog K_s peak at 0.5 but higher than that of the *Hosta-Agave* ortholog K_s peak at 0.1.

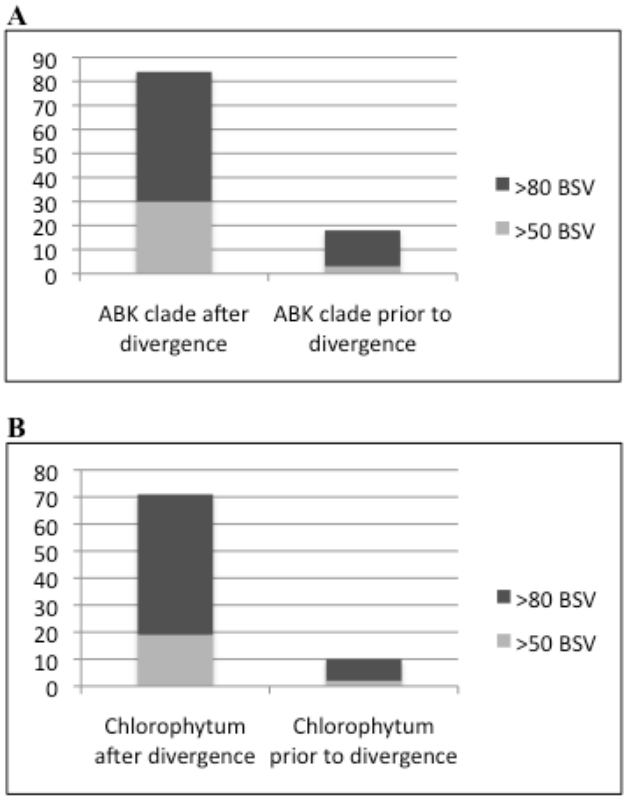

Figure 4.5. Total counts for duplication events in Agavoideae inferred from gene tree topologies

Total counts for duplication events in Agavoideae inferred from gene tree topologies suggest genome-wide duplications on the branches leading to the ABK clade (H2) and *Chlorophytum* (H3). Histograms are shown for duplications observed in gene trees including (A) *A. tequilana* and *H. venusta* paralog pairs, and (B) *C. rhizopendulum* paralog pairs.

Table 4.1. Contigs statistics for assemblies of six study species including the filtered counts.

Species	Contig total	Contigs >300bp	Contigs >500bp	Contigs used in gene family analyses
Agave tequilana	12972	12972	11087	9052
Chlorophytum rhizopendulum	58766	33369	19770	19879
Hosta venusta	57423	19054	3076	9810
Allium cepa	12990	8992	6683	8992
Asparagus officinalis	107254	62708	43093	31431
Leochilus labiatus	43860	18316	8073	10947

CHAPTER V

DISCUSSION AND CONCLUSIONS

Polyploidy has shaped the evolution of angiosperms providing novel genetic material for morphological and physiological innovations (Soltis et al., 2009; Fawcett and Van de Peer, 2010). Understanding exactly how polyploidy has influenced angiosperm evolution requires knowing when paleopolyploid events occurred on the angiosperm phylogeny. Though numerous studies have looked at paleopolyploid events in the history of angiosperms (Vision et al., 2000; Bowers et al., 2003; Blanc and Wolfe, 2004; Paterson et al., 2004; Barker et al., 2008, 2009; Shi et al., 2010; Tang et al., 2010) few have definitively identified the timing of these events (Jiao et al., 2011, 2012). Until recently, the main barriers in determining the timing of paleopolyploid events were taxon sampling and sequencing depth of taxa. With the advent of high-throughput sequencing, this is no longer an obstacle as transcriptomes, a favorite target of paleopolyploidy analyses, can be obtained relatively easily and cheaply. The work presented here utilizes a phylogenomic approach to determine the timing of paleopolyploid events. Two previously identified events in the history of Poales, *rho* and *sigma* (Tang et al., 2010), were placed on the Poales phylogeny, and a third event, *tau*, was identified to have occurred sometime within the monocot lineage. Two novel polyploid events were described within the Asparagaceae subfamily Agavoideae and one was placed in relation to the origin of the Agavoideae bimodal karyotype (ABK) clade. Additionally, a whole chloroplast genome phylogeny was generated for Agavoideae, resolving relationships so that paleopolyploid events could be placed in a phylogenetic context.

Two paleopolyploid events in the history of the grasses have been previously identified through synteny, K_s, and phylogenetic analyses (Bowers et al., 2003; Blanc and Wolfe, 2004; Paterson et al., 2004; Schlueter et al., 2004; Yu et al., 2005; Tang et al., 2010). Despite the extensive study of these events and their importance in the history of major crops, researchers have been encumbered by a paucity of sequence data from non-grass Poales members. Utilizing the sequenced transcriptomes and genomes of taxa across Poales and other monocots, the timing of the *rho* and *sigma* events were determined. The *rho* event was shown to have occurred prior to the diversification of the grass family Poaceae. This coincides with previous studies of MADS-box gene families that suggested a duplication event occurred prior to the diversification of Poaceae (Preston and Kellogg, 2006; Preston et al., 2009). The spikelet, a major innovation for the grass family, originated at about this time and may be the result of the *rho* event. With *rho* at the base of Poaceae, increased diversification rates, which are thought to be associated with polyploid events (Soltis et al., 2009), seem to follow the WGD radiation lag-time model (Schranz et al., 2012). The disparity in time between the increased diversification rate of the PACCMAD+BEP clade and the timing of *rho* could be explained by the delayed development of the PACCMAD+BEP clade's well-defined spikelet, which is not seen in the basal Poaceae subfamily Anomochlooideae (Sajo et al., 2008, 2012). It is possible that increased diversification rates often associated with polyploidy, at least in some cases, are the result of the development of a novel, advantageous features and not directly caused by the polyploidy event itself through mechanisms such as reciprocal gene loss (Lynch and Force, 2000; Scannell et al., 2006; Freeling, 2009). Now that *rho* has been placed on the Poales phylogeny, more study into the fate of duplicated genes and how they may have influenced the evolution of the grasses can

be done with the proper sampling of outgroups, such as Ecdeiocoleaceae, for comparative studies.

The *sigma* event, which was just recently identified (Tang et al., 2010), was placed prior to the diversification of Poales, but after the Poales lineage split from the commelinids. Comparative studies of pre- and post-*sigma* duplication taxa are fairly difficult at this position on the monocot phylogeny. The earliest branching lineages of Poales are separated from the closest extant clades in Zingiberales and Commelinales by at least 20 million years (Magallón and Castillo, 2009) leaving no close outgroup species to the *sigma* event. Nevertheless, comparative studies focusing on paralogs created at the *sigma* event could lead to potential genetic links to the morphological diversity found within Poales.

A surprising finding of the Poales duplication analysis was the strong evidence for a third event in the history of the grasses, *tau*. Based on this study, *tau* occurred sometime prior to the divergence of Asparagales and the commelinids. An older monocot event was speculated by Tang et al. (2010), though the evidence was weak, and a duplication event shared by Commelinales and Poales was found in the analysis of MADS-box genes, which could be this event (Litt and Irish, 2003). The identification and strong support found in this phylogenomic approach bodes well for its potential to identify other unknown paleopolyploid events. A stipulation of that potential, however, would be identification of paralog pairs. In this case, it was misidentification of paralogs as *sigma* that led to the identification of *tau*. Further refinement of methods to identify paralog pairs is needed to properly identify separate paleopolyploid events when they occur in succession as seen in the monocots.

In order to identify and phylogenetically place potential paleopolyploid events in Agavoideae, a resolved phylogeny is needed. Using whole chloroplast genomes, the

relationships within Agavoideae were resolved with high support. The phylogeny exhibited a number of short internodes towards the based of the ABK clade, indicating a rapid radiation. Analyses of transcriptome data show that there is a paleopolyploid event occurring after the divergence of the ABK clade from the *Behnia*/*Chlorophytum*/*Echeandia* clade but prior to the diversification of the ABK clade. Studies in Hawaiian silverswords have suggested that polyploidy can be linked to adaptive radiations (Barrier et al., 1999), which may be the case in Agavoideae. The "*Yucca-Agave*" bimodal karyotype (Whitaker, 1934; Sato, 1935; Watkins, 1936) originates sometime near the polyploid event, but analysis for a causative link was not conducted. The phylogeny suggests that there are two instances where the "*Yucca-Agave*" karyotype is lost, once in the *Hesperocallis*/*Chlorogalum*/*Camassia* clade and the other in the genus *Schoenolirion*. The role of the ABK clade paleopolyploid event in the evolution of agaves, yuccas, and their kin remains to be elucidated. Current studies are underway to examine the fate of paralogs linked to the ABK paleopolyploid event and their role in ABK evolution.

The data presented in this study demonstrate that a phylogenomic approach, where paralogs from a polyploid event are used to query gene family trees, can be used to place paleopolyploid events phylogenetically. This approach was used to place two important events in the history of the grasses, *rho* and *sigma*, furthering the understanding of how polyploidy has shaped the evolution of this economically important group. The identification of a paleopolyploid event in Agavoideae and of a third duplication event in the monocot lineage demonstrate the utility of this approach for discovery of previous unknown paleopolyploid events. With the timing of events in Poales and Agavoideae now known, further studies can begin to understand the consequences of these polyploid events and how they shaped the evolution of these lineages.

References

BARKER, M.S. ET AL. 2008. Multiple Paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution* 25: 2445–2455.

BARKER, M.S., H. VOGEL, AND M.E. SCHRANZ. 2009. Paleopolyploidy in the Brassicales: analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. *Genome Biology and Evolution* 1: 391–399.

BARRIER, M., B.G. BALDWIN, R.H. ROBICHAUX, AND M.D. PURUGGANAN. 1999. Interspecific hybrid ancestry of a plant adaptive radiation: allopolyploidy of the Hawaiian silversword alliance (Asteraceae) inferred from floral homeotic gene duplications. *Molecular Biology and Evolution* 16: 1105–1113.

BLANC, G., AND K.H. WOLFE. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant Cell* 16: 1667–1678.

BOWERS, J.E., B.A. CHAPMAN, J. RONG, AND A.H. PATERSON. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422: 433–438.

FAWCETT, J.A., AND Y. VAN DE PEER. 2010. Angiopserm polyploids and their road to evolutionary success. *Trends in Evolutionary Biology* 2: e3.

FREELING, M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual Review of Plant Biology* 60: 433–453.

JIAO, Y. ET AL. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biology* 13: R3.

JIAO, Y. ET AL. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.

LITT, A., AND V.F. IRISH. 2003. Duplication and diversification in the *APETALA1/FRUITFULL* floral homeotic gene lineage: implications for the evolution of floral development. *Genetics* 165: 821–833.

LYNCH, M., AND A. FORCE. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154: 459–473.

MAGALLÓN, S., AND A. CASTILLO. 2009. Angiosperm diversification through time. *American Journal of Botany* 96: 349–365.

PATERSON, A.H., J.E. BOWERS, AND B.A. CHAPMAN. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences* 101: 9903–9908.

PRESTON, J.C., A. CHRISTENSEN, S.T. MALCOMBER, AND E. A KELLOGG. 2009. MADS-box gene expression and implications for developmental origins of the grass spikelet. *American Journal of Botany* 96: 1419–1429.

PRESTON, J.C., AND E. A KELLOGG. 2006. Reconstructing the evolutionary history of paralogous *APETALA1/FRUITFULL*-like genes in grasses (Poaceae). *Genetics* 174: 421–437.

SAJO, M.G., H.M. LONGHI-WAGNER, AND J. RUDALL. 2008. Reproductive morphology of the early-divergent grass *Streptochaeta* and its bearing on the homologies of the grass spikelet. *Plant Systematics and Evolution* 275: 245–255.

SAJO, M.G., N. PABÓN-MORA, J. JARDIM, D.W. STEVENSON, AND J. RUDALL. 2012. Homologies of the flower and inflorescence in the early-divergent grass *Anomochloa* (Poaceae). *American Journal of Botany* 99: 614–628.

SATO, D. 1935. Analysis of the karyotypes in *Yucca, Agave* and the related genera with special reference to the phylogenetic significance. *The Japanese Journal of Genetics* 11: 272–278.

SCANNELL, D.R., K. BYRNE, J.L. GORDON, S. WONG, AND K.H. WOLFE. 2006. Multiple rounds of speciation associated wtih reciprocal gene loss in polyploid yeasts. *Nature* 440: 341–345.

SCHLUETER, J.A. ET AL. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* 47: 868–876.

SCHRANZ, M.E., S. MOHAMMADIN, AND EDGER. 2012. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Current Opinion in Plant Biology* 15: 147–153.

SHI, T., H. HUANG, AND M.S. BARKER. 2010. Ancient genome duplications during the evolution of kiwifruit (*Actinidia*) and related Ericales. *Annals of Botany* 106: 497–504.

SOLTIS, D.E. ET AL. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* 96: 336–348.

TANG, H., J.E. BOWERS, X. WANG, AND A.H. PATERSON. 2010. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proceedings of the National Academy of Sciences* 107: 472–477.

VISION, T.J., D.G. BROWN, AND S.D. TANKSLEY. 2000. The origins of genomic duplications in *Arabidopsis*. *Science* 290: 2114–2117.

WATKINS, G. 1936. Chromosome numbers and species characters in *Yucca*. *American Journal of Botany* 23: 328–333.

WHITAKER, T.W. 1934. Chromosome constitution in certain monocotyledons. *Journal of the Arnold Arboretum* 15: 135–143.

YU, J. ET AL. 2005. The genomes of *Oryza sativa*: a history of duplications. *PLoS Biology* 3: 1003–1006.