To my parents and brother, who have always been supportive of my scientific pursuits
ACKNOWLEDGMENTS

I would like to thank Gordon for his continued support, pushing me to constantly develop as a scientist. Mike and Michele have been especially kind to me over the years, always available to help me pursue evolutionary biology as well as build skills to be an effective educator for material I am passionate about. The Braun and Kimball labs have had a tremendous influence on me and the development of my research. Brad and Matias have been great committee members, helping me develop research avenues in plant genomics and think about compelling questions to have a long-term career as an evolutionary biologist. A special thanks to Charlie for his many insightful discussions in population genetics and playing and never hesitating to help me out. I am grateful to my collaborators. The research in Chapter 3 would not have been possible without Cecile Ané. Mike Barker has made contributions to the research in Chapter 4 and we have continued working together. I am also thankful to my friend and collaborator Li at Arizona. Jing, Lorena, Natya, and Jay have all been great friends during our time together in the Burleigh lab. I sincerely appreciate all of the great faculty, staff, and friends at UF that have made working on my Ph.D. a fun and special experience.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>4</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>8</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>9</td>
</tr>
<tr>
<td>LIST OF OBJECTS</td>
<td>11</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>12</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>13</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td><strong>1 PLANT GENE FAMILY EVOLUTION</strong></td>
<td>15</td>
</tr>
<tr>
<td><strong>2 THE RELATIONSHIP OF RECOMBINATION RATE, GENOME STRUCTURE, AND PATTERNS OF MOLECULAR EVOLUTION ACROSS ANGIOSPERMS</strong></td>
<td>18</td>
</tr>
<tr>
<td>Introduction</td>
<td>18</td>
</tr>
<tr>
<td>Methods</td>
<td>22</td>
</tr>
<tr>
<td>- Recombination Rate Estimates and Genome Architecture</td>
<td>22</td>
</tr>
<tr>
<td>- Sequence Data and Genome Content</td>
<td>25</td>
</tr>
<tr>
<td>- Species Tree for Comparative Analyses</td>
<td>26</td>
</tr>
<tr>
<td>- Phylogenetic Structure</td>
<td>27</td>
</tr>
<tr>
<td>- Phylogenetic Independent Contrasts</td>
<td>28</td>
</tr>
<tr>
<td>- Selection Analyses</td>
<td>28</td>
</tr>
<tr>
<td>Results</td>
<td>29</td>
</tr>
<tr>
<td>- Phylogenetic Structure of Traits</td>
<td>29</td>
</tr>
<tr>
<td>- Recombination Rate and Genome Architecture</td>
<td>30</td>
</tr>
<tr>
<td>- Addressing Uncertainty in Genomic Architecture</td>
<td>32</td>
</tr>
<tr>
<td>- Recombination Rate and Selection</td>
<td>33</td>
</tr>
<tr>
<td>- Composition Biases across Genes Associated with Recombination Rate</td>
<td>34</td>
</tr>
<tr>
<td>Discussion</td>
<td>35</td>
</tr>
<tr>
<td>Conclusions</td>
<td>41</td>
</tr>
<tr>
<td><strong>3 EVALUATING AND CHARACTERIZING ANCIENT WHOLE GENOME DUPLICATIONS IN PLANTS WITH GENE COUNT DATA</strong></td>
<td>48</td>
</tr>
<tr>
<td>Introduction</td>
<td>48</td>
</tr>
<tr>
<td>Results</td>
<td>50</td>
</tr>
<tr>
<td>- Evidence for Ancient WGDs in Plants</td>
<td>50</td>
</tr>
<tr>
<td>- Challenges in Testing WGD Hypotheses</td>
<td>52</td>
</tr>
<tr>
<td>- Evaluating Model Performance with Simulations</td>
<td>53</td>
</tr>
</tbody>
</table>
4 ASSESSING THE PERFORMANCE OF Ks PLOTS FOR DETECTING ANCIENT WHOLE GENOME DUPLICATIONS .................................................. 85

Introduction ........................................................................... 85
Results .................................................................................... 89
Detecting WGDs Using Node-Averaged Ks ...................... 89
Detecting WGDs Using Pairwise Ks ..................................... 92
Analyzing “Syntenic” Data Instead of All Paralogs ............ 93
Performance of Mixture Models on Empirical Transcriptomic Data .................................................. 94
Discussion ............................................................................ 95
Methods .............................................................................. 103
Simulating Gene Family Evolution .................................... 104
Estimating Synonymous Substitution Rates ..................... 106
Discriminate Analyses and Fitting Mixture Models .......... 107
Comparisons with Empirical Data ...................................... 108

5 CONSERVED RATES OF MOLECULAR EVOLUTION AMONG DUPLICATE GENE COPIES AFTER PALEOPOLYPLOIDY AND A COMPLEMENTARY HYPOTHESIS TO GENE DOSAGE BALANCE .................................................. 119

Introduction ........................................................................... 119
Results .................................................................................... 123
Rates of Molecular Evolution Following Gene Duplications in Grasses ...... 123
Rates of Molecular Evolution among Gene Duplication Mechanisms in Legumes .................................................. 125
Biased Retention of Genes Following WGD ............................ 127
Discussion ............................................................................ 129
Methods .............................................................................. 135
Data, Alignments, and Gene Trees ...................................... 135
Finding Syntenic Genes within and between Genomes ........ 136
Hypothesis Testing with Rates of Molecular Evolution ........ 137
Enrichment of Gene Ontological Categories ...................... 139
Associating Synteny across Species with Rates of Molecular Evolution ...... 139
Evaluating Time-Dependence of Asymmetry after Gene Duplication .......... 140
6 FUTURE DIRECTIONS IN PLANT GENOME EVOLUTION ........................................ 149
LIST OF REFERENCES .................................................................................................. 151
BIOGRAPHICAL SKETCH .............................................................................................. 174
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>All trait data used in the study are displayed.</td>
<td>46</td>
</tr>
<tr>
<td>4-1</td>
<td>Detection and age of WGD peaks for distributed duplication and loss rates with node $K_s$.</td>
<td>117</td>
</tr>
<tr>
<td>5-1</td>
<td>Probabilities for counts of symmetric and asymmetric WGD and non-WGD gene pairs from a two-sided Fisher exact test.</td>
<td>146</td>
</tr>
<tr>
<td>5-2</td>
<td>Counts of symmetric and asymmetric gene pairs on syntenic blocks maintained across species and syntenic blocks not maintained across species.</td>
<td>147</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>1-1</td>
<td>Selective sweeps and background selection are both results of selection acting on linked variation.</td>
<td>17</td>
</tr>
<tr>
<td>2-1</td>
<td>Ultrametric species tree of taxa used for comparative analyses.</td>
<td>42</td>
</tr>
<tr>
<td>2-2</td>
<td>Phylogenetic independent contrasts analyses.</td>
<td>43</td>
</tr>
<tr>
<td>2-3</td>
<td>Distributions of correlation coefficients from COEVOl.</td>
<td>44</td>
</tr>
<tr>
<td>2-4</td>
<td>Composition biases among orthologous groups.</td>
<td>45</td>
</tr>
<tr>
<td>3-1</td>
<td>Ultrametric land plant phylogeny with putative ancient WGD or WGT events plotted as circles on branches.</td>
<td>79</td>
</tr>
<tr>
<td>3-2</td>
<td>Distributions of gene family sizes.</td>
<td>80</td>
</tr>
<tr>
<td>3-3</td>
<td>Testing multiple WGD hypotheses on a single branch using results for <em>Musa acumininata</em> α, β, γ for the land plant data set.</td>
<td>81</td>
</tr>
<tr>
<td>3-4</td>
<td>Gene count data was simulated on 4-taxon trees to test WGD hypotheses and estimate power.</td>
<td>82</td>
</tr>
<tr>
<td>3-5</td>
<td>Testing effects of background rates on WGD detection.</td>
<td>83</td>
</tr>
<tr>
<td>4-1</td>
<td>Distributions of the optimal numbers of components across WGDs of different ages for both known and estimated values of ( Ks ).</td>
<td>109</td>
</tr>
<tr>
<td>4-2</td>
<td>Distributions of the mean ( Ks ) of gene pairs comprising the components closest to the true age of the WGD at ( Ks = 0.5 ) and ( Ks = 1.0 ), when the number of components for each mixture model is optimized by ML.</td>
<td>110</td>
</tr>
<tr>
<td>4-3</td>
<td>Distributions of ( Ks ) for node-averaged and pairwise estimation.</td>
<td>111</td>
</tr>
<tr>
<td>4-4</td>
<td>Distributions of the optimal numbers of components for normal mixture models fit to syntenic node-averaged ( Ks ) data.</td>
<td>112</td>
</tr>
<tr>
<td>4-5</td>
<td>Distributions of the means of components closest to the true age of the WGD at ( Ks = 0.5, 1.0, 3.0, ) and ( 5.0 ), when the number of components for a normal mixture model is optimized by ML.</td>
<td>113</td>
</tr>
<tr>
<td>4-6</td>
<td>Distributions of ( Ks ) for node-averaged a pairwise estimation when a WGD has an age of ( Ks = 0.5, 1.0, 3.0, ) and ( 5.0 ) with among gene family variation in ( \lambda ) and ( \mu ).</td>
<td>114</td>
</tr>
</tbody>
</table>
Ks plots for five previously published transcriptome sequences or EST datasets...

Stylized workflow for simulating gene trees with a known WGD under a known retention rate..

Distributions of $dN/dS$ among speciation, and duplication branches across gene trees in grasses..

Each duplication node in the 2-rate, 3-rate, and 4-rate gene trees were tested for different $dN/dS$ following duplication in grasses..

Distributions of $dN/dS$ among speciation, and duplication branches across gene trees in legumes..

Each duplication node in the 2-rate, 3-rate, and 4-rate gene trees were tested for different $dN/dS$ following duplication in legumes..

Distributions of synteny between species and gene duplication mechanisms within species.
# LIST OF OBJECTS

<table>
<thead>
<tr>
<th>Object</th>
<th>Supplementary material for Chapter 2</th>
<th>page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Supplementary material for Chapter 2</td>
<td>47</td>
</tr>
<tr>
<td>3-1</td>
<td>Supplementary material for Chapter 3</td>
<td>84</td>
</tr>
<tr>
<td>4-1</td>
<td>Supplementary material for Chapter 4</td>
<td>118</td>
</tr>
<tr>
<td>5-1</td>
<td>Supplementary material for Chapter 5</td>
<td>148</td>
</tr>
</tbody>
</table>
### LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>cM</td>
<td>Centimorgan</td>
</tr>
<tr>
<td>$dN/dS$</td>
<td>Nonsynonymous to synonymous substitution rate ratio</td>
</tr>
<tr>
<td>HRI</td>
<td>Hill-Robertson interference</td>
</tr>
<tr>
<td>$K_s$</td>
<td>Synonymous substitution rate</td>
</tr>
<tr>
<td>LTR</td>
<td>Long terminal repeat</td>
</tr>
<tr>
<td>Mb</td>
<td>Megabase</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum likelihood estimate</td>
</tr>
<tr>
<td>MYA</td>
<td>Millions of years ago</td>
</tr>
<tr>
<td>$N_e$</td>
<td>Effective population size</td>
</tr>
<tr>
<td>pg</td>
<td>Picogram</td>
</tr>
<tr>
<td>WGD</td>
<td>Whole genome duplication</td>
</tr>
</tbody>
</table>
Plant genomes are distinguished among other eukaryotes for their high variability in size, genome architecture, and gene families. Understanding the mechanisms driving this variation is fundamental to understanding the broader context of plant evolution and phylogeny as well as the molecular genetic basis for trait evolution. One confounding factor for understanding plant evolution though is the prevalence of whole genome duplications throughout the plant tree of life. Whole genome duplications have certainly contributed to gene family evolution, but their role in driving plant diversity as well as their contributions to gene family evolution in comparison to other gene duplication mechanisms, and how various population-level processes drive the fixation and evolution of duplicate genes is still a subject of debate. The contributions of different gene duplication mechanisms and processes, such as recombination, have been difficult to reconcile, as most inferences to date have relied on observations made within individual genomes. Here, I utilize comparative phylogenetic methods and model-based inference to develop formal probabilistic hypothesis tests, which utilize genomic data across species. I argue that whole genome duplication alone does not drive novelty in
gene family evolution, and that recombination is important, having contributed to patterns of gene retention that are spatially constrained rather than functionally biased.
Understanding genome complexity across eukaryotes is a fundamental goal in molecular evolution. Despite the many aspects genome evolution driving variation at the macroevolutionary scale, the area perhaps of most interest to evolutionary biologists is gene family evolution. Gene families have evolved over time through complex patterns of gene duplication and loss. Duplicate gene copies have long been hypothesized to be a driver of genetic novelty, as redundant genetic material provides opportunity for new genetic pathways and traits to evolve (Ohno 1970). The accumulation of whole genome sequences across many organisms has revealed that the spatial distributions, rates of molecular evolution, and ultimate fates of duplicate gene copies are non-random. Additionally, there several different mechanisms of gene duplication, which may have profound impact on the retention and evolutionary fates of duplicate genes. Plants are ideal for testing how different gene duplication mechanisms and other genomic features affect gene family evolution, due to their tolerance of genome rearrangements and chromosomal mutations. Plants have typically exhibited larger gene family sizes and larger variation in gene duplication and loss rates compared to other well-studies eukaryotic groups such as mammals and birds. The large numbers of duplicate genes in plants may be due, at least in part, to their prevalence of whole genome duplications (WGDs). The importance of WGDs in plant evolution has been a popular topic in light of comparative genomic data, and here we aim to provide a rigorous treatment of both the detection of WGDs and their biological importance. We provide a perspective on how neutral population genetic processes have played an important role in the biased
retention of genes following WGD opposed to adaptation, which has been a common feature of recent plant WGD literature.

By population genetic processes, we mean the population genetic parameters that ultimately affect the fixation of new mutations entering a population. These are relevant to the fixation and fates of duplicate genes, as we can consider gene duplication events as de novo mutations, similar to the mutation of a single nucleotide base pair. Population genetic processes are effective population size ($N_e$), recombination, selection, and mutation. The least well-characterized affect here is recombination. One way that we can relate recombination to the other processes here though is through the concept of Hill-Robertson interference (HRI; Felsenstein 1974). HRI generally refers to the effects of linkage on genetic variation, which includes selective sweeps (Hill and Robertson 1969) and background selection (Charlesworth et al. 1993). Both cases imply that lower recombination rates lead to lower $N_e$ and reduced efficacy of selection on individual sites (Fig. 1-1). The effects of recombination have often been overlooked in comparative plant genomics, due to its difficulty to measure and high variability due to dependencies on genome architecture (Gaut et al. 2007). However, recombination may play an important role in both the birth and retention of duplicate genes (Tiley and Burleigh 2015), especially those produced by WGDs. I first make a case for recombination and $N_e$ playing a role in gene gain and loss, then characterize methods for detecting WGDs, and finally use a phylogenetic framework to draw rigorous statistical inferences on the differences in gene duplication mechanisms within a multi-species framework.
Figure 1-1. Selective sweeps and background selection are both results of selection acting on linked variation. Selective sweeps quickly drive beneficial mutations and any linked variation to fixation in a population. Background selection purges variation linked to deleterious mutations, including beneficial variation from populations. Both cases reduce diversity as a result of high linkage disequilibrium.
CHAPTER 2
THE RELATIONSHIP OF RECOMBINATION RATE, GENOME STRUCTURE, AND PATTERNS OF MOLECULAR EVOLUTION ACROSS ANGIOSPERMS

Introduction

Meiotic recombination has been a topic of interest in evolutionary biology since Fisher first addressed the effects of linkage on substitutions in a population (Fisher 1930), yet the macroevolutionary consequences of recombination on plant genomes are still poorly understood. Comparative studies of the effects of recombination rate on genome architecture and sequence evolution across distantly related species require both whole genome sequences and data-intensive estimates of recombination rates (McVean et al. 2002; Stumpf and McVean 2003; Wang and Rannala 2008). Recent genome sequencing and genetic mapping efforts, which provide physical measurements of genome size and map length, make studies of global recombination rate possible in plants. In this study, we take advantage of these new data to explore the relationship between recombination rate, genome structure, and patterns of molecular evolution throughout angiosperms in order to better characterize the broad macroevolutionary patterns of recombination rate variation and its possible consequences for genome evolution.

Recombination affects both genome architecture and evolutionary rates. Lynch (2006) showed that generation scaled global recombination rate (centimorgans/basepairs/generation) decreases as species genome size increases in unicellular eukaryotes, invertebrates, vertebrates, and land plants. Similarly, in plants, Cavalier-Smith (1985) proposed that the recombination rate is higher in smaller angiosperm genomes than in larger genomes. Rees and Durrant (1986) corroborated this result in a study of the genera Lathyrus, Lolium, and Petunia and by Narayan and
McIntyre (1989) in *Lathyrus*. Both of these studies estimated nuclear genome size in picograms per haploid genome (C-values) and recombination rates based on observable chiasma from pachytene chromosomes. Typically, one observable chiasma is expected per chromosome arm for segregation to proceed normally. However, the number of crossovers per chromosome arm is variable (Martini et al. 2006), and Ross-Ibarra (2007) demonstrated a positive correlation between genome size and the number of chiasmata per chromosome arm across 279 angiosperm species from 22 families.

One potential mechanism for a negative association between global recombination rate and genome size is that recombination either deletes LTRs by chance or it facilitates selection against transposable element insertions (Langley et al. 1998). Much of the genome size variation in flowering plants can be attributed to changes in repetitive element content, and specifically long terminal repeat (LTR) retrotransposons (Bennetzen 2002; Bennetzen et al. 2005). The loss of LTR retrotransposon content can occur through unequal homologous recombination (Devos et al. 2002). Thus, lineages with higher recombination rates are expected to have lower LTR retrotransposon content, and hence smaller genomes (International *Brachypodium* Initiative 2010), as well as higher gene densities. It is not clear if recombination preferentially removes specific families of LTR retrotransposons; since LTR retrotransposons are removed by unequal crossing over due to high sequence identity, we might expect all LTR retrotransposon families to be removed equally. Additionally, all LTR retrotransposon families appear to have similar life histories in rice (Baucom et al. 2009), suggesting LTR retrotransposons vary only in abundance. Regions of the genome with little or no recombination (i.e., mainly heterochromatin during crossing
over) have longer transposable elements and lower gene density when compared to regions with frequent recombination (Gaut et al. 2007), and recombination rate and gene density are positively correlated in the genomes of maize, rice, wheat, and Arabidopsis thaliana (Anderson et al. 2006; International Rice Genome Sequencing Project 2005; Dvorak et al. 2004; Wright et al. 2003).

Recombination rate also has been linked to the GC content and codon usage bias of genes due to GC biased gene conversion (Marais 2003). Although DNA mismatch repair during crossover resolution can be GC biased (Lesecque et al. 2013), the strength of selection for a site and the effects of linkage alone can alter local codon usage landscapes (Loewe and Charlesworth 2007). GC biased gene conversion drives a positive relationship between local recombination rate and codon bias within the Caenorhabditis elegans and Drosophila melanogaster genomes (Marais et al. 2001). There is also a positive correlation between GC content and local recombination rate across mammals (Lartillot 2013) and within humans (Fullerton et al. 2001; Meunier and Duret 2004), which may indicate the strength of GC biased gene conversion. However, the relationship between recombination and compositional biases in angiosperms is unclear. Local recombination rate is weakly negatively correlated with GC content in Medicago truncatula (Paape et al. 2012), but not within self-fertilizing populations of Arabidopsis thaliana, likely due to reduced heterozygosity (Marais et al. 2004). Correlations between recombination rate and GC content appear to be a feature of exclusively outcrossing species (Pessia et al. 2012). Despite the lack of an obvious relationship between recombination rate and GC content across most plant species (Marais et al. 2004), there is evidence that GC biased gene conversion is occurring in
some lineages. For instance, individual gene families in grasses show evidence of nucleotide composition biases and gene conversion (Zhang et al. 2001).

Within populations, recombination can create favorable combinations of alleles that may have a selective advantage in future generations, while linkage between sites may reduce the efficacy of selection (Muller 1964; Crow and Kimura 1965), a phenomenon known as Hill-Robertson effects (Felsenstein 1974). Hill-Robertson effects include hitchhiking (Hill and Robertson 1966), fixation of sites linked to a beneficial mutation, and background selection (Charlesworth et al. 1993) or loss of variation linked to a deleterious mutation, which lead to reduced effective population size for a genomic region with a low recombination rate. Recombination rate is negatively associated with the ratio of nonsynonymous to synonymous substitution rates ($dN/dS$) within genomes and positively correlated with $dS$ in model organisms such as *Drosophila melanogaster* (Campos et al. 2014). However, evidence of Hill-Robertson effects is typically weaker in plant genomes (Tenaillon et al. 2002; Wright et al. 2006), even when considering variation in life history traits (Baudry et al. 2001). Moreover, it is not clear if the effects of recombination rates are pervasive over long evolutionary time periods, since recombination landscapes can vary over time (Baudet et al. 2010) and across populations (Colomé-Tatché et al. 2012; Bauer et al. 2013), and if recombination rate is associated with $dN/dS$ between species. For example, there was no association between recombination rate and rates of molecular evolution in comparisons between *Arabidopsis thaliana* and *A. lyrata* (Yang and Gaut 2011).

Both tandemly duplicated genes (Zhang and Gaut 2003; Akhunov et al. 2003; Rizzon et al. 2006) and dispersed duplicates (Rizzon et al. 2006) are more prevalent in
regions of the genome with high recombination. The long-term survival of duplicate genes may be enhanced by purifying selection, which is more effective in regions of high recombination (Hill and Robertson 1966). The probability of subfunctionalization or neofunctionalization of a duplicate gene increases with recombination rate (Lynch and Force 2000; Xue et al. 2010), and once the new gene copy has reached fixation, the probability of the duplicate gene’s survival also increases with recombination rate (Lynch et al. 2001). Given these expectations and observations of more duplicate genes in regions of high recombination, we hypothesize that species with higher global recombination rates may have more duplicate genes, resulting in larger gene families.

Here we make a first attempt to characterize the potential macroevolutionary role of recombination rate in shaping plant genomes. We examine correlates of global recombination rate across thirty phylogenetically diverse angiosperm species, with respect to genome architecture, compositional biases, and \(dN/dS\) in 3748 single-copy nuclear genes.

**Methods**

**Recombination Rate Estimates and Genome Architecture**

We assembled data for thirty angiosperm species with sequenced genomes and linkage maps from the primary literature (Fig. 2-1; Object 2-1, Table S1). Only genetic maps where the numbers of linkage groups correspond to the haploid chromosome number were used to estimate global recombination rate, and we used multiple maps for each species and calculated recombination rate from average map lengths (Object 2-1, Table S2). We corrected map lengths for each species for marker density using method 4 of Chakravarti et al. (1991), as implemented by Hall and Willis (2005) and Dumont and Payseur (2008). Global recombination rate was measured by taking the
corrected map length divided by the genome size in megabases ($cM/Mb$), where genome size is the total mapped and unmapped scaffold assembly size. Genome sizes were obtained from primary literature and early release statistics available on Phytozome (Table 2-1; citations are provided in Object 2-1, Table S1).

Our analyses assume that recombination rates estimated from genetic maps covary with rates of unequal crossing over. The rates of allelic homologous recombination appears to be a reasonable indicator of the rates of non-allelic crossing over in *Saccharomyces cerevisiae* (Jinks-Robertson and Petes 1985; Kupiec and Petes 1988; Lichten et al. 1987), but this remains to be broadly shown in plants. We might not expect allelic crossing over to always be a reasonable predictor of non-allelic crossing over though, since non-allelic crossing over is dependent on genome spatial complexity (Goldman and Lichten 1996).

Estimates of genome size in megabases may contain error due to the genome assembly. Therefore, we also calculated global recombination rates using C-values (cM/pg) as the estimates of genome size. C-values were taken from the Kew C-Value Database (http://data.kew.org/cvalues/; last accessed 22 August 2014; Object 2-1, Table S3). Some species used in this study can have different ploidy levels, resulting in multiple, distinct C-values (Object 2-1, Table S3). To test the effects of the different ploidy levels on the correlations between recombination rate and genome size, we generated 100 datasets by randomly selecting a single C-value for each species and performed phylogenetically corrected correlations for both cM/pg and pg and cM/pg and Mb. We performed a meta-analysis of the correlation coefficients with fixed effects using the R package metacor (Laliberté 2011).
In addition to genome size, we also looked at the relationship between recombination rate and genome compactness, defined as the genome size over the haploid chromosome number. Recombination rates should be higher on shorter chromosome arms, since at least one crossover per chromosome arm is expected (Martini et al. 2006). Chromosome size has been used as an indirect measure of recombination rate variation within a genome (Pessia et al. 2012), so we tested if the overall genome compactness (genome size / haploid chromosome number) was correlated with global recombination rate. We also tested if haploid chromosome number was correlated with global recombination rate.

Global recombination rates estimated using the map length over the total genome size might not be directly comparable between species because recombination generally occurs in euchromatic regions of chromosomes during meiosis (e.g., Thuriaux 1977). For example, 97%, 98%, and 95% of the genetic maps correspond to euchromatin in *Sorghum bicolor*, *Oryza sativa*, and *Zea mays* respectively (Paterson et al. 2009). The amount of the genome that is euchromatic during crossing-over can vary greatly between species. Thus, we also estimated a corrected recombination rate based on the euchromatic proportion of the genome for the 19 species (Table 2-1) in which fluorescence *in situ* hybridization or other analyses of pachytene chromosomes were performed to differentiate the chromosomal characteristics during meiosis. We used relative percentages of heterochromatin from the literature and subtracted that from the genome assembly size for each species (citations for differential chromatin studies are in Object 2-1, Table S1). We assumed 95% of the genetic map lies in the euchromatic portion of the genome during crossing over. Thus, the euchromatin corrected
recombination rate is equal to the total scaffold size minus the estimated percentage of heterochromatin in megabases over 95% of the marker-density corrected map length.

For the 29 species with published genomes and available transposable element data, we obtained the proportion of the genome consisting of all LTR retrotransposons (Table 2-1). This was used to calculate genome size without LTR retrotransposons by subtracting the percent content of LTR retrotransposons from the total genome size. Estimating genome size without LTR retrotransposon content was done to address if an association between recombination rate and genome size can be explained by LTR retrotransposon content alone. Detailed transposable element classification was available for 20 species, which allowed us to investigate if relationships between recombination rate and LTR retrotransposon content could be explained by the proportion of copia or gypsy superfamilies. The copia and gypsy superfamilies were selected because they are generally the most abundant LTR retrotransposon classes and constitute most of the variation in LTR retrotransposons in plants.

Finally, gene density was obtained from the literature or early release statistics by dividing the number of predicted genes by the genome size. These data were available for all 30 species used in the study (Table 2-1).

**Sequence Data and Genome Content**

Gene families for the 30 angiosperms with recombination rate data were downloaded from Phytozome v9.1 (www.phytozome.net, Last accessed 29 September 2013). Gene sequences were clustered into families based on reciprocal BLASTP distances, with full details described in Goodstein et al. (2012); clusters are provided by JGI through Phytozome using the BioMart tool. We translated the nucleotide sequences from Phytozome into amino acids and then aligned the amino acid sequences with
MUSCLE 3.8.31 (Edgar 2004). We obtained in-frame nucleotide alignments by mapping the codons to the aligned amino acid sequences using in-house Perl scripts. Perl scripts were also used to calculate GC content at 3\textsuperscript{rd} position 4-fold degenerate sites (3GC\textsuperscript{S}) and codon bias, measured as effective number of codons (ENC; Wright 1990), for each sequence for each gene family. We used ENC to measure codon bias because it is not biased by functional constraints of amino acid composition or gene length (Wright 2004; Comeron and Aguadé 1998). We were interested in 3GC\textsuperscript{S} because we wanted to test if GC biased gene conversion is detectable throughout the genome in plants. If GC biased gene conversion is generally occurring, then we would expect a stronger bias in 3GC\textsuperscript{S} for genomes with higher recombination rates. Additionally, we calculated the average gene family size from the number of genes in each gene family to test if recombination facilitates gene duplication or the preservation of duplicate genes. Only gene families that spanned the root of the tree in Figure 2-1 were used; this included 11250 of the 12748 Phytozome gene families.

**Species Tree for Comparative Analyses**

For the phylogenetically informed analyses, we used a species tree (Fig. 2-1) with a topology that corresponds to our current understanding of angiosperm phylogeny between species (www.phytozome.net; e.g., Soltis et al. 2011). While accounting for phylogenetic uncertainty is important in many studies, the relationships of the 30 taxa used here are mostly well established, and it is computationally prohibitive to repeat some analyses in this study using a distribution of trees. Full chloroplast genomes were not available for all species. Therefore, molecular branch lengths were estimated from an alignment of *matK* sequences (aligned length of 2036 bp) downloaded from Genbank (http://www.ncbi.nlm.nih.gov) using the GTR model implemented in HYPHY
2.1.2.28 (Kosaakovsky Pond et al. 2005). \textit{MatK} is noted for providing reasonable topology and branch length estimates across angiosperms (Hilu et al. 2003). We transformed the branch lengths to make them ultrametric using penalized likelihood in \textit{r8s} (Sanderson et al. 2003). \textit{Amborella trichopoda} was used as the outgroup, and the age of the most recent common ancestor of angiosperms was fixed to 150 million years ago (mya). Minimum age constraints were placed on Poaceae (65 mya; Kellogg 2001; Magallón and Castillo 2009), Fabidae (94 mya; Magallón and Castillo 2009; Davis et al. 2005), and Malvidae (94 mya; Magallón and Castillo 2009). Maximum age constraints were also placed at the most recent common ancestor of core Eudicots (124 mya; Magallón and Castillo 2009; Soltis and Soltis 2004; Anderson et al. 2005), and Eudicots and Monocots (130 mya; Magallón and Castillo 2009). The best smoothing parameter for the penalized likelihood analysis, 3200, was determined by cross validation. For the phylogenetic independent contrast analyses, to make comparisons consistent with the assumption that a contrast’s mean is independent of its standard deviation (Garland et al. 1992), contrasts were analyzed using the PDAP package in \textit{MESQUITE} (Midford et al. 2005; Maddison and Maddison 2006) and a base-10 logarithmic transformation was performed on the ultrametric branch lengths.

**Phylogenetic Structure**

We calculated Blomberg’s $K$ (Blomberg and Garland 2002; Garland et al. 2003) to test for a phylogenetic signal for recombination rate, genome size, LTR retrotransposon proportions, gene density, gene family size, global composition biases, genome compactness, and haploid chromosome number under a Brownian motion model of evolution (Object 2-1, Table S4). Blomberg’s $K$ is the ratio of the mean squared error (MSE) of trait values on the tips of the phylogeny and the MSE expected
under Brownian motion. A value of $K$ greater than one implies phylogenetic clustering of traits, and values less than one are consistent with a lack of structure or overdispersion. The significance of a calculated $K$ value was determined by 999 random permutations of the tips on the phylogeny performed using the R package picante (Kembel et al. 2010).

**Phylogenetic Independent Contrasts**

To account for the covariance of character states due to the shared phylogenetic history of different species, we performed phylogenetic independent contrasts (PICs; Felsenstein 1985) to examine the correlation between recombination rate and various characteristics of the genome. All PICs were obtained using the R package APE 3.0-11 (Paradis et al. 2004). We used Pearson’s correlation coefficient to test for associations of the phylogenetic independent contrast values of traits. All correlation analyses were performed in R (R Core Team 2013). All contrast analyses were performed for both raw recombination rate and euchromatin corrected recombination rate. We examined the relationship between global recombination rate and total genome size, genome size without LTR retrotransposon content, the relative abundance of LTR retrotransposons, gene density, and gene family size. We also calculated the power for all correlation analyses using the R package pwr (Champely 2012), which uses the calculations provided by Cohen (1988).

**Selection Analyses**

Single-copy gene families from Phytozome with sequences from more than ten species were evaluated to determine the relationship between recombination rate and $dN/dS$ with COEVOL (Lartillot and Poujol 2011), using partial correlations that control for covariation in $dS$. This was done to investigate differences in levels of selection in
genomes with variable rates of recombination. Since the method of Lartillot and Poujol (2011) relies on a given species tree, only single-copy gene families containing no obvious paralogs were analyzed. To maximize taxonomic sampling, only raw recombination rates were used. $dN$ and $dS$ were optimized along the species tree, and ultrametric branch lengths for the species tree were fixed for performing contrasts. This was done to help the convergence of chains and reduce computational complexity. Two chains were run up to 72 hours, each with geodesic averaging of traits, and convergence of chains was determined by an effective sample size (ESS) greater than 300 for all parameters, with the exception of the ancestral state at the root, where the sampling state is especially difficult over large evolutionary time periods. An ESS of 50 was used for the root ancestral state, which yields qualitatively similar runs (Lartillot and Poujol 2011). Parameter estimates for the largest alignments had converged by 72 hours, and parameter estimates that did not converge by this point likely indicated uncertainty in the data, possibly due to alignment, clustering, or annotation. Gene families with chains that did not converge were not used in the independent contrast analyses. For chains that did converge, 25% of the chain was discarded as burn-in, and partial correlation coefficients for recombination rate and $dN/dS$ were calculated for independent contrasts in COEVOL.

**Results**

**Phylogenetic Structure of Traits**

We first asked if there was a phylogenetic signal for recombination rate as well as several features of genome architecture. Global recombination rate, euchromatin corrected recombination rate, genome size, the proportion of the genome that consists of LTR retrotransposons, and average gene family size do not deviate
significantly from Brownian motion. Both global ENC and GC3S have significant phylogenetic structure, meaning that trait values are more similar amongst closely related species. Genome size without LTR retrotransposons, gene density, genome compactness, and haploid chromosome numbers are phylogenetically overdispersed, such that there is more variation than expected under Brownian motion (Object 2-1, Table S4).

**Recombination Rate and Genome Architecture**

In the next analyses, we evaluate if there is a correlation between global recombination rate and genome size, LTR retrotransposon content, and gene density. If recombination enables the elimination of LTR retrotransposons, we may expect negative correlations between recombination rate and genome size and LTR retrotransposon content and a positive correlation between recombination rate and gene density.

Phylogenetic independent contrast analyses show a strong negative correlation between the global recombination rate and genome size (Fig. 2-2a; \( r = -0.65, p < 0.001 \)). However, this strong negative correlation breaks down after removing the LTR retrotransposon content (Fig. 2-2a; \( r = 0.15, p = 0.460 \)). Plant genome size is strongly, positively correlated with total LTR retrotransposon content (\( r = 0.72, p < 0.001 \)), and global recombination rate is negatively associated with LTR retrotransposon content (Fig. 2-2b; \( r = -0.56, p = 0.002 \)). The correlation between recombination rate and LTR retrotransposon content is not biased towards either the *gypsy* (\( r = -0.31, p = 0.200 \)) or *copia* (\( r = -0.34, p = 0.160 \)) superfamilies. Recombination rate also is positively correlated with gene density (Fig. 2-2c; \( r = 0.57, p = 0.001 \)). There is no detectable correlation
between the global recombination rate and the average gene family size (Fig. 2-2d; \( r = 0.16, p = 0.410 \)).

The global recombination rate and euchromatin corrected recombination rate are correlated (\( r = 0.69, p = 0.001 \)), but euchromatin corrected recombination rate is not significantly correlated with most of the genomic traits including genome size (Fig. 2-2a; \( r = -0.28, p = 0.265 \)), genome size without LTR retrotransposon content (Fig. 2-2a; \( r = -0.31, p = 0.212 \)), LTR retrotransposon content (\( r = -0.11, p = 0.673 \)) including both the gypsy (\( r = -0.08, p = 0.789 \)) and copia (\( r = -0.23, p = 0.435 \)) contributions, and gene density (Fig. 2-2c; \( r = 0.34, p = 0.173 \)). However, there is a strong positive correlation between euchromatin corrected recombination rate and average gene family size (Fig. 2-2d; \( r = 0.82, p < 0.001 \)). Global recombination rate is not correlated with genome compactness or haploid chromosome number (Object 2-2, Fig. S1). However, euchromatin corrected recombination rate is negatively correlated with genome compactness (Object 2-2; \( r = -0.63, p = 0.027 \)).

The lack of significant correlations using euchromatin corrected recombination rate, but not global recombination rate, possibly due to the smaller sample size. For example, we have power of 0.98 to refute the null hypothesis for raw recombination rate and genome size with a sample of 30, but only power of 0.32 to refute the null hypothesis for euchromatin corrected recombination rate and genome size with a sample of 19. To have power of 0.8 for correlations of PICs and sample size of 19, the correlation coefficient would need to be 0.62. Some of the differences may be biological as well, considering the correlation coefficient between global recombination rate and genome size is -0.57 (\( p = 0.013 \)) for the same sample of 19 taxa.
Addressing Uncertainty in Genomic Architecture

Next, we addressed some possible sources of uncertainty and error in the correlation analyses. The heterogeneous sources of data may produce uncertainty or error in our results. For example, the difficulty of assembling repetitive DNA in a genome sequence can lead to underestimates of genome size (Shearer et al. 2014) and inaccurate estimates of transposable element content. Although this may introduce error into the estimates of genome size or transposable element content, we do not think it is biasing the analyses. Correlations between global recombination rates estimated with C-values, which are not affected by the ability to assemble repetitive DNA, and genome size in Mb yield similar results (Object 2-1, Tables S5 and S6).

An additional concern when investigating plant genome size is the history of polyploidy. Not all taxa examined in this study are diploid, and they have different histories of lineage-specific polyploidy events. This concern was addressed by Ross-Ibarra (2007), who demonstrated that including or excluding polyploids from analyses had little to no effect on the relationship between recombination rate and genome size. Our results also suggest that polyploidy does not influence the general relationship between recombination rate and genome size. Analysis of C-values for non-heterochromatin recombination rates reveals consistent negative correlations between recombination rate and genome size, no matter which ploidy level was selected. For the 100 permutations of genome size (pg) for species with multiple ploidy levels, the metacorrelation between recombination rate (cM/pg) and genome size in pg has mean $r = -0.70$ with a 95% confidence interval (-0.683, -0.721) and $p < 0.001$. The metacorrelation of recombination rate (cM/pg) and genome size in Mb has mean $r = -0.35$ with a 95% confidence interval (-0.31, -0.38) and $p < 0.001$. This suggests that
genome size is negatively correlated with recombination rate, and this result is robust to both different ploidy levels between and within species as well as error in genome size estimates from assemblies.

**Recombination Rate and Selection**

We were interested if purifying selection is associated with increasing recombination rate more frequently than positive selection, which is observed within populations of model organisms such as *Drosophila melanogaster* (Campos et al. 2014). Low $dN/dS$ values can be interpreted as evidence for purifying selection. Therefore, if recombination rate is also associated with purifying selection at a macroevolutionary scale, we should find more evidence for negative correlations between recombination rate and $dN/dS$ than positive correlations across genes.

Recombination rate is correlated with $dN/dS$ for 6.9% of 3748 genes using nominal two-tailed posterior probability cutoffs of 0.025 and 0.975. Since we are using posterior probabilities, it is inappropriate to correct for the family-wise error rate or the false discovery rate by treating them as p-values. Therefore, we correct our nominal posterior probability cutoffs to achieve desired 5% significant results. Of the significant results, 16% of these are positive correlations (posterior probability > 0.983), while 84% are negative correlations (posterior probability < 0.017) (Figure 3). This indicates that most genes experience more effective purifying selection as global recombination rate increases, while few genes experience higher $dN/dS$, which could be due to relaxed selective pressures or to increased efficacy of positive selection with increasing recombination rate (Betancourt and Presgraves 2002). Hidden paralogy is always a concern when investigating plant nuclear genes, but this likely would only make convergence more difficult, generating uncertainty and lowering ESS at duplication
nodes. Overall, increasing recombination rate is associated with stronger purifying selection for both the significant pool and non-significant pool of corrected tests ($X^2 = 262.44, p < 0.001$). These results imply that the rate of recombination, a population genetic process, can influence the patterns of molecular evolution across species.

We also found evidence of correlations between recombination rate and the $dN/dS$ of several genes linked to recombination. Ring finger domains play a critical role as ubiquitin ligases (Ruffner et al. 2001), and it is thought that ring finger domain containing proteins assist with initiating double stranded breaks (Wright et al. 2005). Ring finger domain containing proteins may take part in the meiotic recombination mechanism in plants, since they are associated with early protein-protein interactions for crossover formation in model systems like yeast and Caenorhabditis elegans (Perry et al. 2005; Lynn et al. 2007; Jantsch et al. 2004). Sequence variation in the ring finger domain containing protein RNF212 also is correlated with recombination rate in humans (Kong et al. 2008). Three ring fingers are significantly positively correlated with recombination rate in this study (Fig. 2-3), but there is no overall enrichment of zinc fingers in the COEVOL results (Fisher exact test, $p = 0.515$; Object 2-1, Table S7).

Likewise, the well characterized Mre11, which participates in heteroduplex resolution and possibly telomere maintenance, and the nonhomologous end joining protein Ku70 both experience stronger purifying selection (i.e., lower $dN/dS$) as recombination increases (Fig. 2-3), but meiotic recombination proteins are not enriched in the COEVOL results either (Fisher exact test, $p = 0.071$; Object 2-1, Table S8).

**Composition Biases across Genes Associated with Recombination Rate**

Finally, we tested for associations between recombination rate and nucleotide compositional biases, in an attempt to understand the degree that GC biased gene
conversion might influence covariation between recombination rate and $dN/dS$.
Recombination rate was not correlated with the genome average ENC or 3GC$^S$ (Object 2-2, Fig. S2). However, recombination rate could still affect sequence variation in genes in which $dN/dS$ is correlated with global recombination rate. We used the alignments analyzed with COEVOL to examine the distribution of ENC and 3GC$^S$ for genes in which the $dN/dS$ is not associated with recombination rate and the pool of genes in which $dN/dS$ is associated with recombination rate (Object 2-1, Tables S9-S12). For simplicity, we binned distributions into recombination rate quartiles (Fig. 2-4). We do not observe a pattern, suggesting that genes in which $dN/dS$ is correlated with global recombination rate have a stronger bias in codon usage or 3GC$^S$ than genes in which $dN/dS$ is not correlated with global recombination rate. Detectable composition biases in our data appear to be lineage specific, such as strong biases in both ENC and 3GC$^S$ in grasses (Fig. 2-4).

**Discussion**

Global recombination rate appears to be evolutionarily labile across angiosperms (Fig. 2-1; Object 2-1, Table S4), and even relatively closely related congeners can have very different recombination rates (e.g., *Solanum lycopersicum* and *S. tuberosum*). Jaramillo-Correa et al. (2010) found evidence of phylogenetic structure in recombination rate estimates (cM/Mb) across 81 seed plant species. However, Jaramillo-Correa et al. (2010) included conifers, which have low recombination rates and likely contributed heavily to the observed phylogenetic structure. Still, this study indicates that global recombination rate is strongly associated with the evolution of genome structure and patterns of molecular evolution in angiosperms (Fig. 2-2 and Fig. 2-3).
We find a strong negative correlation between global recombination rate and genome size as well as the total proportion of LTR retrotransposons. This result is consistent with many previous studies and expectations (Fig. 2-2; Lynch 2006; Devos et al. 2002; Shirasu et al. 2000; Vitte et al. 2003). In contrast, Ross-Ibarra (2007) observed that recombination rate increases with genome size in angiosperms, when measuring global recombination rate as the number of chiasmata per chromosome arm. It is unclear what mechanism would produce a positive correlation between recombination rate and genome size, but still, this does not necessarily contradict our results, as the recombination rate estimates are not directly comparable. While there may be some error in our estimates of recombination rate, linkage map lengths were averaged across multiple populations for each species, and the correlations were similar whether we used the scaffold assembly size or C-values to represent genome size (Object 2-1, Tables S5 and S6). Furthermore, any error in recombination rate estimates should not strengthen the correlation with genome size. The negative correlation between recombination rate and genome size also makes sense mechanistically. LTR retrotransposons, which play a large role in expanding the genome size of plants, may generally have deleterious effects in gene rich euchromatin, and meiotic recombination can facilitate removal of LTR retrotransposons by unequal or intra-strand crossing over. The relationship between recombination rate and LTR retrotransposon content may entirely explain the relationship between recombination rate and genome size, as recombination rate is not negatively correlated with genome size after removing LTR retrotransposon content (Fig. 2a). However, many correlates of global recombination
rate are also likely correlates of genome size, and therefore, our analyses do not prove a causal relationship between global recombination rates and genome size.

We found no evidence linking recombination rate and $3GC^S$ or ENC globally (Object 2-2, Figure S2) or across angiosperm genes where recombination rate is correlated with $dN/dS$ (Fig. 2-4). However, grasses have strong composition biases compared to the non-monocots represented in this study (Fig. 2-4). GC biased gene conversion has been observed in grasses (Zhang et al. 2001; Haudry et al. 2008; Muyle et al. 2011), which may indicate that some effects of recombination are specific to certain clades but not pervasive across all angiosperms. For example, GC biased gene conversion appears to be reduced in self-fertilizing species (Marais et al. 2004; Glémin et al. 2006; Escobar et al. 2010) Thus, there may be a combination of biological factors necessary for GC biased gene conversion to occur.

Recombination is generally restricted to gene rich regions of the genome (Thuriaux 1977), and therefore, it is sensible to correct for the nonrandom distribution of crossovers when calculating recombination rates (e.g., Ross-Ibarra 2007). There was a positive correlation between global and euchromatin corrected recombination rates, and correlation coefficients estimated using the euchromatin corrected recombination rate were generally weaker, but consistent with those estimated using the global recombination rate. One exception is that euchromatin corrected recombination rate was negatively correlated with genome compactness (Object 2-2, Fig. S1), which may better characterize the euchromatic portion of the genome.

Interestingly, the euchromatin corrected recombination rate is strongly positively associated with the average number of genes in a gene family, while global
recombination rate is not. Cook’s distance applied to a linear regression model indicates that the relationship between euchromatic recombination rate and average gene family size is largely influenced by contrasts between *Solanum tuberosum* and *S. lycopersicum*, *Manihot esculenta* and *Populus trichocarpa*, *Zea mays* and *Sorghum bicolor*, and the ancestral state of Asterids (represented only by *S. tuberosum* and *S. lycopersicum* in the euchromatin corrected data) and Rosids. We applied Kendall’s tau to the data because a nonparametric test should be less sensitive to possible outliers, and a significant positive association remains ($\tau = 0.4118, p = 0.017$). Whole genome duplication alone also cannot explain these contrasts. For example, *S. bicolor* has a larger average gene family size than *Z. mays*, despite *Z. mays* having undergone a lineage specific whole genome duplication since its divergence from *S. bicolor* (Gaut and Doebley 1997). Additionally, *S. tuberosum* and *S. lycopersicum* share a whole genome triplication (Tomato Genome Consortium 2012), yet the genome of *S. lycopersicum* is composed of much more heterochromatin and has a larger average gene family size. The positive correlation between the euchromatin corrected recombination rate and average gene family size also persists even if the largest gene families are removed (Object 2-1, Table S13). Thus, this association is not due to massive expansion of a few families or clustering errors.

The positive association between euchromatin corrected recombination rate and gene family size may be due to a link between recombination and tandem duplication rate, since more duplicate genes are located near sites of recombination (Zhang and Gaut 2003; Akhunov et al. 2003; Rizzon et al. 2006). Conversely, gene loss rates also may be lower near regions of high recombination due to the presence of strong purifying
selection. Recombination rate can affect the time to fixation and the efficacy of selection for duplicate genes (Lynch and Force 2000; Innan and Kondrashov 2010). Specifically, the probability of subfunctionalization or neofunctionalization of a newly duplicated gene and the preservation of that gene is maximized under free recombination (Lynch and Force 2000; Xue et al 2010; Lynch et al. 2001). Lower levels of recombination will ultimately reduce the probability of preservation of any given duplicate, assuming the duplication event itself is not selected for due to additive dosage effects (Lynch et al. 2001). Variation in angiosperm gene content is often discussed in the context of whole genome duplications (Jiao et al. 2011; Jiao et al. 2012; Amborella Genome Project 2013), but our results suggest that recombination, independent of whole genome duplications, may be critical for creating and maintaining gene copy variation.

Recombination is linked to efficacy of purifying selection in populations (Campos et al 2014) and also within genomes. Generally, genes in regions of the genome with high recombination rates should have lower $dN/dS$ than genes in regions of low recombination. Campos et al. (2014) found that regions with crossovers have seven times the synonymous nucleotide diversity of regions without crossovers in a population of *Drosophila melanogaster*, which corresponds with findings in populations of *Arabidopsis lyrata* (Kawabe et al. 2008), and patterns of SNP variation in humans (Lercher and Hurst 2002). Our analyses suggest that the increased efficacy of selection due to recombination is also observable on a macroevolutionary scale across angiosperms. Our analyses linking lower $dN/dS$ with increased recombination rates in many genes support the hypothesis that purifying selection acts more effectively in species with higher global recombination rates (Fig. 2-3). Although we find a small
proportion of genes where $dN/dS$ increases with global recombination rate, it is not certain that this is the result of hitchhiking. However, patterns of $dN/dS$ variation across genes imply a role for background selection in plant genome evolution. While several genes involved with meiotic recombination were analyzed in this study, these follow the broader pattern of correlations between recombination rate and genes. These results suggest that while a small proportion of recombination associated genes are correlated with recombination rate, there is not likely any selective pressures acting on these genes as a group to modify recombination rate.

While our results suggest a role for recombination in shaping macroevolutionary patterns of genome architecture and molecular evolution in plants, well-known covariates of recombination rate, genome size, and substitution rates, such as effective population size (Hazzouri et al. 2013), could strongly affect the results. Obtaining estimates of effective population size can be challenging (Lanfear et al. 2014), and we could not incorporate effective population size into this study due to the limited availability of these estimates. Regardless, our results suggest that Hill-Robertson effects may have macroevolutionary consequences on both the interspecific rates of molecular evolution and the average size of gene families among species. Other gene-specific factors, such as gene function, tissue specificity, expression level, and architectural features of the genes, may further elucidate the possible relationship between interspecific recombination rate evolution and patterns of variation in $dN/dS$, as they have for intraspecific studies (e.g., Larracuente et al. 2008; Weber and Hurst 2009; Slotte et al. 2011), since these also covary with rates of molecular evolution and recombination rate (Pál et al. 2001; Webster and Hurst 2012). Associations between
recombination rate and selection are typically weaker in plants than other eukaryotes (Slotte 2014; Cutter and Payseur 2013), so codon models that allow for among site rate heterogeneity in $dN/dS$ on branches might also help further reveal the relationship between local recombination rate and $dN/dS$ across species (Gossman et al. 2014).

**Conclusions**

Although genomic data has enabled many insights into plant evolution, the role of population level evolutionary processes on macroevolutionary patterns is still largely unknown. Understanding the impacts of recombination rate variation, in addition to effective population size, selection, and mutation, is necessary for elucidating genome evolution. The results presented in this study are largely consistent with previous intraspecific studies (Gaut et al. 2007; Anderson et al. 2006; International Rice Genome Sequencing Project 2005; Dvorak et al. 2004; Paape et al. 2012; Marais et al. 2004). While it is unclear if our results reflect the role of recombination in genome evolution or a covariate of recombination rate, taken together with previous research, they suggest recombination rate affects genome architecture and the distribution of $dN/dS$ across angiosperm species. The effects include removal of LTR retrotransposons and influencing gene duplication and loss. Recombination rate variation may not only explain the rate at which tandem duplicates arise, but also the preservation of duplicate genes through increased efficacy of purifying selection.
Figure 2-1. Ultrametric species tree of taxa used for comparative analyses. Divergence times were estimated based on molecular branch lengths and fossil calibrations using r8s. Asterisks denote nodes with fossil calibrations. The distribution of recombination rate (cM/Mb) is given along the tips.
Figure 2-2. Phylogenetic independent contrast analyses. a) Phylogenetic independent contrasts for both global recombination rate and euchromatin corrected recombination rate with genome size and genome size without LTR retrotransposons. Size estimates were log2 transformed for normality to satisfy assumptions of phylogenetic independent contrasts and parametric correlations. b) Global recombination rate plotted against total LTR retrotransposon content as well as the proportions of *copia* and *gypsy* superfamilies. c) Gene density is based on gene number and genome size estimate from literature. d) Plots of recombination rate vs average gene family size based on both the 11250 families that span the root as well as smallest 75% of gene families. The best-fit linear regression is given for each contrast plot.
Figure 2-3. Distributions of correlation coefficients from COEVOL. Only results from chains that converged are displayed. A negative correlation coefficient indicates a relationship between global recombination rate and purifying selection for a gene while a positive correlation indicates a relationship between recombination rate and relaxed selection for a gene. The total distribution is displayed in grey, while the significant distribution for posterior probabilities of 0.025 and 0.975 are shown in red, and corrected posterior probabilities for 5% significant results are in blue.
Figure 2-4. Composition biases among orthologous groups. a) Distributions of ENC binned by species with the bottom, second, third, and top quartiles of global recombination rate, pooled from not significant and significant genes from the 3748 alignments used for COEVOL. b) The total distributions of ENC for the 3748 COEVOL genes for species from the best represented plant families in this study. c) Distributions of 3GC\textsuperscript{S} ordered by recombination rate. d) Complementary distribution of 3GC\textsuperscript{S} shows similar patterns of codon bias and GC bias at 3\textsuperscript{rd} position 4-fold degenerate sites in Poaceae vs other plant families best represented in this study.
Table 2-1. All trait data used in the study are displayed. When trait data was unavailable for certain species NA is used.

<table>
<thead>
<tr>
<th>Species</th>
<th>cM/ Mb</th>
<th>Euchromatin cM/Mb</th>
<th>Genome size (MB)</th>
<th>Proportion of LTRs</th>
<th>Proportion Copia</th>
<th>Proportion Gypsy</th>
<th>Gene Density</th>
<th>Average Gene Family Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis lyrata</td>
<td>2.26</td>
<td>2.87</td>
<td>207.00</td>
<td>0.200</td>
<td>NA</td>
<td>NA</td>
<td>157.83</td>
<td>0.768</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>4.17</td>
<td>4.27</td>
<td>125.00</td>
<td>0.040</td>
<td>NA</td>
<td>NA</td>
<td>248.91</td>
<td>1.209</td>
</tr>
<tr>
<td>Brachypodium distachyon</td>
<td>5.52</td>
<td>NA</td>
<td>272.00</td>
<td>0.214</td>
<td>0.049</td>
<td>0.161</td>
<td>93.87</td>
<td>0.952</td>
</tr>
<tr>
<td>Brassica rapa</td>
<td>4.83</td>
<td>5.10</td>
<td>283.80</td>
<td>0.271</td>
<td>NA</td>
<td>NA</td>
<td>144.99</td>
<td>0.79</td>
</tr>
<tr>
<td>Capsella rubella</td>
<td>4.15</td>
<td>NA</td>
<td>134.80</td>
<td>0.094</td>
<td>NA</td>
<td>NA</td>
<td>196.74</td>
<td>0.751</td>
</tr>
<tr>
<td>Carica papaya</td>
<td>2.82</td>
<td>3.22</td>
<td>372.00</td>
<td>0.333</td>
<td>0.055</td>
<td>0.278</td>
<td>66.52</td>
<td>0.712</td>
</tr>
<tr>
<td>Citrus clementina</td>
<td>2.77</td>
<td>3.31</td>
<td>301.40</td>
<td>0.199</td>
<td>0.079</td>
<td>0.120</td>
<td>81.40</td>
<td>0.787</td>
</tr>
<tr>
<td>Citrus sinensis</td>
<td>2.19</td>
<td>2.62</td>
<td>320.50</td>
<td>0.153</td>
<td>0.071</td>
<td>0.082</td>
<td>91.87</td>
<td>0.784</td>
</tr>
<tr>
<td>Cucumis sativus</td>
<td>3.15</td>
<td>3.41</td>
<td>243.50</td>
<td>0.104</td>
<td>0.054</td>
<td>0.038</td>
<td>109.58</td>
<td>0.749</td>
</tr>
<tr>
<td>Eucalyptus grandis</td>
<td>2.09</td>
<td>NA</td>
<td>691.00</td>
<td>0.160</td>
<td>NA</td>
<td>NA</td>
<td>52.64</td>
<td>0.79</td>
</tr>
<tr>
<td>Fragaria vesca</td>
<td>1.86</td>
<td>NA</td>
<td>240.00</td>
<td>0.110</td>
<td>NA</td>
<td>NA</td>
<td>104.38</td>
<td>1.924</td>
</tr>
<tr>
<td>Glycine max</td>
<td>2.40</td>
<td>3.38</td>
<td>1060.00</td>
<td>0.430</td>
<td>0.130</td>
<td>0.300</td>
<td>43.80</td>
<td>0.867</td>
</tr>
<tr>
<td>Gossypium raimondii</td>
<td>1.96</td>
<td>3.97</td>
<td>761.40</td>
<td>0.449</td>
<td>0.338</td>
<td>0.111</td>
<td>53.82</td>
<td>0.821</td>
</tr>
<tr>
<td>Linum usitatissimum</td>
<td>3.42</td>
<td>NA</td>
<td>318.30</td>
<td>0.184</td>
<td>NA</td>
<td>NA</td>
<td>136.30</td>
<td>0.823</td>
</tr>
<tr>
<td>Malus domestica</td>
<td>1.52</td>
<td>NA</td>
<td>742.30</td>
<td>0.307</td>
<td>0.055</td>
<td>0.252</td>
<td>77.31</td>
<td>0.845</td>
</tr>
<tr>
<td>Manihot esculenta</td>
<td>1.91</td>
<td>3.03</td>
<td>760.00</td>
<td>0.111</td>
<td>NA</td>
<td>NA</td>
<td>40.35</td>
<td>0.811</td>
</tr>
<tr>
<td>Medicago truncatula</td>
<td>3.00</td>
<td>3.35</td>
<td>257.60</td>
<td>0.242</td>
<td>0.410</td>
<td>0.057</td>
<td>171.33</td>
<td>0.708</td>
</tr>
<tr>
<td>Mimulus guttatus</td>
<td>4.96</td>
<td>NA</td>
<td>321.70</td>
<td>0.200</td>
<td>0.100</td>
<td>0.100</td>
<td>87.47</td>
<td>0.79</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>4.73</td>
<td>5.41</td>
<td>372.00</td>
<td>0.235</td>
<td>0.025</td>
<td>0.120</td>
<td>114.66</td>
<td>1.084</td>
</tr>
<tr>
<td>Panicum virgatum</td>
<td>1.11</td>
<td>NA</td>
<td>1358.00</td>
<td>0.111</td>
<td>NA</td>
<td>NA</td>
<td>72.17</td>
<td>0.805</td>
</tr>
<tr>
<td>Phaseolus vulgaris</td>
<td>2.67</td>
<td>2.78</td>
<td>521.10</td>
<td>0.367</td>
<td>0.094</td>
<td>0.251</td>
<td>52.19</td>
<td>0.794</td>
</tr>
<tr>
<td>Populus trichocarpa</td>
<td>4.96</td>
<td>6.73</td>
<td>485.00</td>
<td>0.065</td>
<td>0.016</td>
<td>0.049</td>
<td>94.13</td>
<td>1.431</td>
</tr>
<tr>
<td>Prunus persica</td>
<td>2.45</td>
<td>2.54</td>
<td>227.30</td>
<td>0.186</td>
<td>0.086</td>
<td>0.100</td>
<td>122.53</td>
<td>0.813</td>
</tr>
<tr>
<td>Setaria italica</td>
<td>3.28</td>
<td>NA</td>
<td>405.70</td>
<td>0.250</td>
<td>NA</td>
<td>NA</td>
<td>87.43</td>
<td>0.736</td>
</tr>
<tr>
<td>Solanum lycopersicum</td>
<td>1.81</td>
<td>7.50</td>
<td>760.00</td>
<td>0.618</td>
<td>0.063</td>
<td>0.197</td>
<td>45.69</td>
<td>2.156</td>
</tr>
<tr>
<td>Solanum tuberosum</td>
<td>1.15</td>
<td>1.63</td>
<td>727.00</td>
<td>0.522</td>
<td>0.038</td>
<td>0.152</td>
<td>48.15</td>
<td>1.232</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>2.12</td>
<td>4.02</td>
<td>738.50</td>
<td>0.544</td>
<td>0.052</td>
<td>0.190</td>
<td>37.43</td>
<td>1.464</td>
</tr>
<tr>
<td>Theobroma cacao</td>
<td>2.69</td>
<td>NA</td>
<td>326.90</td>
<td>0.160</td>
<td>0.070</td>
<td>0.090</td>
<td>88.09</td>
<td>1.537</td>
</tr>
<tr>
<td>Vitis vinifera</td>
<td>2.99</td>
<td>NA</td>
<td>487.10</td>
<td>0.145</td>
<td>0.084</td>
<td>0.032</td>
<td>68.80</td>
<td>0.76</td>
</tr>
<tr>
<td>Zea mays</td>
<td>0.72</td>
<td>1.24</td>
<td>2066.40</td>
<td>0.751</td>
<td>0.218</td>
<td>0.377</td>
<td>15.75</td>
<td>0.738</td>
</tr>
</tbody>
</table>
Object 2-1. Supplementary material for Chapter 2
CHAPTER 3
EVALUATING AND CHARACTERIZING ANCIENT WHOLE GENOME DUPLICATIONS
IN PLANTS WITH GENE COUNT DATA

Introduction

Whole genome duplications (WGDs) play an important role in shaping the genomes of plants (e.g., Adams and Wendel 2005). The availability of large-scale genomic data and fully sequenced genomes has revealed much evidence for ancient WGDs, or paleopolyploidy (e.g., Vision et al. 2000; Cui et al. 2006; Jallion et al. 2007; Jiao et al. 2011; Vanneste et al. 2014). In fact, recent evidence suggests at least two WGDs preceded the diversification of angiosperms (Jiao et al. 2011; Amborella Genome Project, 2013; Li et al. 2015). WGDs in plants often are followed by rapid fractionation, in which many homeologous gene copies are lost (Otto 2007; Mandáková et al. 2010; Schnable et al. 2012), and this diminishes evidence of WGDs over time. Consequently, although WGDs appear to be pervasive throughout the evolutionary history of plants, it can be extremely difficult to detect, let alone characterize, these ancient events (e.g., Burleigh 2012).

Perhaps the most direct evidence for ancient WGDs is the presence of large syntenic regions within a genome (e.g., Kellis et al. 2004; Jaillon et al. 2004). However, few studies have performed statistically rigorous tests of WGD hypotheses based on syntenic data, and many measures of synteny are presented without estimates of uncertainty. Although recent models of gene family evolution use syntenic data and account for many complexities of WGD, these are used for probabilistic orthology prediction and not explicitly testing WGDs (Conant and Wolfe 2008; Conant 2014). Additionally, reoccurring WGDs throughout plant evolution can make the interpretation
of syntenic data difficult, especially without well-assembled genomes that can be used to detect synteny between species.

In the absence of well-annotated genomes assembled to the chromosome level, ancient WGDs often are inferred from the distribution of the rate of synonymous substitution per synonymous site \( (dS) \) among duplicate genes in a genome (e.g., Lynch and Conery 2000; Raes et al. 2003). It is generally assumed that gene duplication and loss follows a steady-state birth-death process, with constant rates of duplication (birth) or loss (death) per gene family (Lynch and Conery 2003). WGDs violate this assumption, and consequently, WGDs produce peaks in cumulative distributions of pairwise \( dS \) between paralogs within a genome (e.g., Gu et al. 2002; McLysaght et al. 2002; Jaillon et al. 2004; Vandepoele et al., 2004; Maere et al. 2005). In the case of an ancient WGD, the steady-state birth-death process alone cannot describe the distribution of \( dS \) between paralogs. However, the signal of a WGD in the distribution of \( dS \) may degrade through time, and it can be extremely difficult to identify ancient WGDs based on \( dS \) distributions (e.g., Blanc and Wolfe 2004a; Paterson et al. 2004).

Furthermore, multiple substitutions at the same site can create peaks in \( dS \) plots that do not correspond to WGDs (Vanneste et al. 2013).

Recently, Rabier et al. (2014) described a new approach to identify WGDs (or WGTs – whole genome triplications) on a phylogeny based on gene count data, the number of gene copies in various gene families across a group of taxa. Hahn et al. (2005) originally developed a maximum likelihood approach to estimate gene birth (i.e., duplication) and death (i.e., loss) rates on a phylogeny with gene count data. Rabier et al. (2014) extended this approach to estimate background rates of gene duplication (\( \lambda \))
and loss ($\mu$) throughout the tree and account for the probability of a WGD when reconstructing ancestral gene copy numbers along the nodes of a phylogeny. When the user defines WGDs (or WGTs) within the tree, the model also assumes a proportion of the extra duplicated genes are lost immediately following the WGD event. The fraction of extra genes that survive from a WGD is the retention rate ($q$), and the model estimates independent gene retention rates for each WGD in the tree. This approach is appealing because the user can explicitly test for WGDs along specific branches of a phylogeny with a likelihood ratio test by comparing models with and without WGDs. The user also can estimate the timing of the WGD along a branch; the likelihood is maximized when the WGD is placed in its optimal position along an edge in the species phylogeny.

In this study, we evaluate the evidence for numerous ancient WGDs across land plants and estimate gene retention rates following WGDs using gene copy number data from fully sequenced genomes. We explore the effects of estimates of gene duplication rates ($\lambda$) and gene loss rates ($\mu$) on estimates of WGD retention rates ($q$) and use simulations and empirical tests to assess our ability to detect WGD events across sequenced plant genomes.

Results

Evidence for Ancient WGDs in Plants

Using the full land plant data, we detected nine WGDs from of the 19 ancient WGD hypotheses tested (Fig. 3-1). We detected WGD and WGT events with high $q$ values relative to the estimated $\lambda$ and $\mu$ consistently across all of our analyses (Fig. 3-1; Object 3-1, Tables S1-S4). For example, we detected WGDs that occurred in the *Physcomitrella patens* lineage ($q = 0.098$) and the *Aquilegia coerulae* lineage ($q =$
0.073), both of which are also supported by dS distributions (Rensing et al 2007; Cui et al 2006). We also detected evidence for a WGT that is shared by Solanum lycopersicum and S. tuberosum (q = 0.102), a WGD preceding Populus truchocarpa (q = 0.597), a WGD shared by Papilionoids (q = 0.163), and a WGD that was specific to Glycine max (q = 0.763), all of which have some evidence based on synteny (The Tomato Genome Consortium 2012; Tuskan et al 2006; Schmutz et al 2012). We detected the oldest of three proposed WGDs that occurred in the Musa acuminate lineage (q = 0.258) as well as the older of two proposed events that preceded the diversification of Poaceae (q = 0.062). We also detected the older WGD in the Arabidopsis thaliana lineage, Arabidopsis β WGD (q = 0.063), using the full land plant data set, but we detected both Arabidopsis α (q = 0.023) and β (q = 0.088) using Theobroma cacao and the Brassicales only (i.e., the Brassicales data set; Fig. 3-1).

The estimates of the proportion of genes retained from WGDs varied greatly among putative ancient WGD events (Fig. 3-1). These retention rates were robust to both the choice of prior and the specific gene set being used (Object 3-1, Tables S1-S4). The prior in these analyses is the probability distribution of the number of gene copies at the root of the species tree, which accounts for uncertainty in ancestral gene family sizes. A geometric mean of 1.5 is a reasonable prior choice based on the distribution of average gene family sizes (Fig. 3-2a). Prior choice had little effect on estimates of λ and μ in the land plant analyses, but it affected our ability to detect some WGD or WGT events in data sets that span the root of Amborella trichopoda and the eudicots (the eudicot data set) as well as Theobroma cacao and the Brassicales (the Brassicales data set; Fig. 3-2b; Object 3-1, Tables S2 and S4). Specifically, we did not
detect the WGT preceding *Solanum* when we assumed the number of genes at the root of *Amborella trichopoda* and the eudicots was geometrically distributed with a mean of one (i.e., only one gene copy at the root of the tree for each gene family; Object 3-1, Table S2). An inappropriately high prior for the analyses of *T. cacao* and Brassicales led to a failure to detect both the *Arabidopsis* α and β WGDs, but we detected both events when we assume the number of genes at the root is geometrically distributed with a mean of 1.5 (Object 3-1, Table S4).

**Challenges in Testing WGD Hypotheses**

In several cases, our analyses did not detect multiple WGDs on the same branch on a phylogeny. Tests of the three consecutive WGDs on the branch leading to *Musa acuminata* (*M. acuminata* α, β, and γ) favored a scenario with only the oldest (γ) WGD. Placing a single WGD close to 96 million years ago (mya) on the *M. acuminata* branch resulted in a significant improvement in the likelihood score (Fig. 3-3). LRTs showed no statistical support for WGDs in *M. acuminata* at 65 mya (β) and 60 mya (α). When testing these WGDs with data that includes only the monocots and *Amborella trichopoda*, a single WGD on the branch of *M. acuminata* near 96 mya, with $q = 0.275$ was also preferred.

Testing the two WGD hypotheses before the diversification of Poaceae with the land plant data set also suggested only a single WGD near 96 mya (Poaceae σ; Fig. 3-1). However, we did not detect the Poaceae σ WGD with the monocot data set (Object 3-1, Table S3). The more recent (70 mya) Poaceae ρ WGD was not statistically significant and had estimated retention rates of 0.001 and 0 using both the land plant (Fig. 3-1; Object 3-1, Table S1) and monocot data sets (Object 3-1, Table S3) respectively.
We detected the *Arabidopsis* β WGD based on the hypothesis in Fig. 3-1, where we placed the *Arabidopsis* α and β WGDs are placed on separate branches. However, we did not detect the *Arabidopsis* β WGD when both *Arabidopsis* α and β WGDs were on the same branch (Object 3-1, Table S5). In addition, we did not detect evidence of the most ancient WGDs, including those preceding the diversification of angiosperms and seed plants (Fig. 3-1). We explored the timing of the angiosperm and seed plant WGDs by adjusting the age of the putative WGDs along 5 mya intervals on the branch prior to the divergence between *Amborella trichopoda* and other angiosperms, and in all cases we were still unable to detect the angiosperm and seed plant WGDs (Object 3-1, Table S6). We also did not find evidence of a WGT common to all eudicots in both the land plant and the eudicot data sets (Fig. 3-1; Object 3-1, Tables S1 and S2); however, we detected the eudicot WGT using 4-taxon trees (see Results “Evaluating model performance with simulations”). A WGD common to all monocots had a small, but insignificant $q = 0.007$ in the analysis of the land plant data set and $q = 0$ in the monocot data set (Fig. 3-1; Object 3-1, Tables S1 and S3).

**Evaluating Model Performance with Simulations**

We performed simulation experiments to test model performance in cases where we did not detect a well-documented ancient WGD and determine if the failure to detect a WGD or WGT was due to a lack of power. For the eudicot WGT, we tested three simulation scenarios: 1) a WGT on a terminal branch represented by *Vitis vinifera* (Fig. 3-4a), 2) a WGT on an internal branch, without any WGDs following the eudicot WGT (Fig. 3-4b), and 3) a WGT on an internal branch with a WGD on the tip leading to *Populus trichocarpa* and WGT on the tip leading to *Solanum lycopersicum* after the eudicot WGT (Fig. 3-4c). For testing the eudicot WGT on a single terminal branch (Fig.
we had power of 0.76 to detect the WGT when $q = 0.01$ with 5,000 gene families (Object 3-1, Table S7), and power of 0.89 with 10,000 gene families. Similarly, when the eudicot WGT is on an internal branch with no other WGDs or WGTs on the tips (Fig. 3-4b), power = 1 when $q = 0.10$ and there were at least 500 gene families, and power = 0.97 for $q = 0.01$ when there were 10,000 gene families (Object 3-1, Table S8). When the eudicot WGT was placed on an internal branch with an additional WGT and WGD occurring on terminal branches (Fig. 3-4c), we had power of 0.96 and 1 for 10,000 gene families when $q = 0.01$ and $q = 0.10$ respectively for all WGD and WGT events (Fig. 3-4c; Object 3-1, Table S9). However, when the eudicot WGT had $q = 0.01$ and the more recent events on terminal branches had $q = 0.10$, we have power of 0.76 to detect the eudicot WGT for 10000 gene families (Fig. 3-4c; Object 3-1, Table S9). In contrast to the analyses using the plant data and the full tree, we detected the eudicot WGT in the observed data using the three 4-taxon trees. The first case, the WGT occurs on the *V. vinifera* branch without any consecutive WGDs or WGTs (Fig. 3-4a; Object 3-1, Table S7). Secondly, the WGT occurs before the divergence of *V. vinifera* and *Theobroma cacao* and is not obstructed by any other WGT or WGD on the tree (Fig. 3-4b; Object 3-1, Table S8). In the third case, the eudicot WGT occurred before the divergence of *P. trichocarpa* and *S. lycopersicum*; afterward, there is also a WGD in the tip leading to *P. trichocarpa* and a WGT in the tip leading to *S. lycopersicum* (Fig. 3-4c; Object 3-1, Table S9). The eudicot WGT had an estimated $q$ of 0.070, 0.067 and 0.104 for these analyses respectively.

Simulation results suggest that detecting multiple WGDs on a single branch is difficult, if not impossible, when retention rates are low (Object 3-1, Table S10).
Specifically, we had no power to detect both the *Arabidopsis* α and β WGDs when \( q_\alpha = q_\beta = 0.01 \), even with 10,000 gene families (Fig. 3-4d). However, the power to detect both the *Arabidopsis* α and β WGDs goes to 1 when \( q_\alpha = q_\beta = 0.10 \) for 10,000 gene families. When we simulated gene count data with \( q_\alpha = 0.10 \) and \( q_\beta = 0.01 \), we had power of 0.02 – 0.62 and 0.04 – 0.19 to detect the *Arabidopsis* α and β WGDs respectively for 500 – 10,000 gene families. When testing the observed data on the same phylogeny used for simulating the *Arabidopsis* α and β WGDs on the same branch, we did not detect the *Arabidopsis* β WGD, but we detected the *Arabidopsis* α WGD with an estimated \( q = 0.092 \) (Object 3-1, Table S5 and S10). This is consistent with our simulation results, which suggests that with 10,000 gene families, we had power of 0.62 to detect the *Arabidopsis* α WGD when \( q_\alpha = 0.10 \) and only power of 0.19 to detect the *Arabidopsis* β WGD for \( q_\beta = 0.01 \) (Fig. 3-4d; Object 3-1, Table S10).

Testing the WGD hypotheses associated with monocots presents a complex scenario with nested WGD hypotheses (Fig. 3-4e). For simulated data with \( q = 0.01 \) for all WGD events, we had no power to detect both the Poaceae ρ and the Poaceae σ WGDs, power of 0.04 – 0.15 to detect the *Phoenix dactylifera* WGD, and power of 0.05 – 0.58 to detect the monocot WGD for 500 – 10,000 gene families (Object 3-1, Table S11). However, when \( q = 0.10 \) for all the monocot WGDs, we had power of 0.73, 0.75, 1, and 1 to detect the Poaceae ρ, Poaceae σ, *P. dactylifera*, and monocot WGDs respectively with 10,000 gene families (Object 3-1, Table S11). When testing the 10,795 monocot gene families on the 4-taxon tree used for simulation (Fig. 3-4e), we detected the monocot WGD with estimated \( q = 0.309 \) and the Poaceae ρ WGD with estimated \( q = 0.053 \); however, we did not detect the *Phoenix dactylifera* WGD or the Poaceae σ
WGD (Object 3-1, Table S11). Notably analyses of the larger phylogenetic hypotheses with land plant data (Fig. 3-1; Object 3-1, Table S1) and monocot data (Object 3-1, Table S3) did not detect the putative ancient WGD preceding the diversification of monocots, but analysis of the 4-taxon tree (Fig. 3-4e) detected the WGD (Object 3-1, Table S11).

We also simulated data to address the WGDs preceding the diversification of seed plants and angiosperms. Due to the available whole genome sampling, the angiosperm and seed plant WGDs were located on the same branch, and they were the oldest WGDs tested in this study (Fig. 3-4f). Simulations indicated that we had power of 0.09, 0.41, and 1 to detect the angiosperm WGD for 500, 1000, and 5,000 or more gene families when $q = 0.01$ (Object 3-1, Table S12). However, we had power of 1 to detect the angiosperm WGD for $q = 0.10$ when there were at least 500 gene families (Object 3-1, Table S12). In the same simulations, we had no power (i.e., power = 0) to detect the seed plant WGD regardless of $q$ or the number of gene families (Object 3-1, Table S12). We did not detect either the angiosperm or seed plant WGD when testing the observed land plant data on the 4-taxon tree used for simulations (Fig. 3-4f; Object 3-1, Table S12).

In the four simulation experiments that included included *Nelumbo nucifera* in the 4-taxon tree, we detected the WGD preceding *Nelumbo nucifera* with $q$ ranging from 0.171 to 0.250 (Fig. 4; Object 3-1, Table S7-S9 and S12).

**Heterogeneity in Gene Duplication and Loss across Lineages and Gene Families**

The *Arabidopsis* α WGD was not significant in analyses using the land plant or eudicot data sets; however, it was detected when testing WGDs with the 4-taxon Brassicales data set (Fig. 3-1). Using the 4-taxon tree requires a different gene count
data set to span the root; this includes 12967 gene families, 6843 of which are shared with the land plant data set. Depending on the root of the tree and the conditioning of the data, different gene families can be considered in the analysis (Fig. 3-1). Apart from simulations, we tested if failure to detect the Arabidopsis α WGD when using the land plant data set was due to 1) lower power, because there were fewer genes in the land plant data set than in the Brassicales data set, 2) a lack of signal in older gene families as opposed to de novo gene families in the Brassicales data set or angiosperm data, or 3) poor estimates of the single gene duplication and loss rates (λ and μ) in Brassicales when using λ and μ estimates from across the entire land plant tree.

We randomly resampled without replacement 7567 gene families from the Brassicales data set (i.e., the number of gene families in the land plant data set) 500 times to test if the land plant data lacked power (see Methods). When we estimated λ and μ for the 500 randomly resampled Brassicales data sets (Object 3-1, Table S13), the Arabidopsis α WGD was always significant (Fig. 3-5a). Therefore, a lack of power does not explain why we fail to detect the Arabidopsis α WGD with the land plant data set. Additionally, the Arabidopsis α was detected when using only the 6843 gene families that span the root of both the land plant tree and the 4-taxon tree. Thus, Brassicales de novo genes alone cannot explain the significant Arabidopsis α LRT in the 4-taxon tree.

Since the Arabidopsis α WGD was detected on the 4-taxon tree using the 6843 gene families in common between the land plant data and Brassicales data, the estimates of λ and μ from the land plant tree may not be appropriate for optimizing the retention rate of the Arabidopsis α WGD. Specifically, the global gene duplication rate
was lower and the global gene loss rate was higher in the Brassicales data set ($\lambda = 0.0013$ and $\mu = 0.0021$) than the global estimates from the full land plant tree ($\lambda = 0.0016$ and $\mu = 0.0019$). Calculating the likelihoods based on the retention rates from the 4-taxon Brassicales tree, but using estimates of $\lambda$ and $\mu$ based on the land plant tree (See Methods for details) led to a non-significant LRT for the analysis using the 6843 gene families that span both land plants as well as *Theobroma cacao* and Brassicales in addition to nine out of the 500 resampled data sets (Fig. 3-5b). Therefore, inappropriate estimates of $\lambda$ and $\mu$ and a stronger signal for WGD from more recent gene families contributed to the failure to detect the *Arabidopsis* $\alpha$ WGD in the land plant data set. Estimates for $q$ from the *Arabidopsis* $\alpha$ WGD on the 4-taxon tree using land plant data, angiosperm data, and Brassicales-only data were 0.023, 0.027, and 0.043 respectively (Fig. 3-5a).

The *Arabidopsis* $\beta$ WGD was significant in all analyses of the resampled data sets as well as analyses using gene families shared by land plants and Brassicales as well as angiosperms and Brassicales, but not Brassicales alone (Fig. 3-5c). When optimizing $\lambda$ and $\mu$ with the Brassicales specific gene families, the *Arabidopsis* $\beta$ WGD had an estimated $q$ of 0. The *Arabidopsis* $\beta$ WGD was significant when fixing $\lambda$ and $\mu$ to rates obtained from the land plant analysis for all data sets (the 500 resampled Brassicales data sets, the land plant data set, the angiosperm data set, and the Brassicales-only data set) using the previous estimates of $q$ (Fig. 3-5d). Therefore, evidence for the *Arabidopsis* $\beta$ WGD was stronger in older gene families that were better characterized by $\lambda$ and $\mu$ optimized from the global land plant data than more recent gene families that had a higher estimated $\lambda$. Retention rates for the *Arabidopsis* $\beta$
Discussion

The growth of genomic data in plants has prompted much interest in identifying evidence of ancient WGDs and associating ancient WGD with diversification and trait evolution (e.g., Vanneste et al. 2014; Cannon et al. 2015). Our analyses suggest that in many cases we can evaluate ancient WGD hypotheses and assess the impact of ancient WGDs on gene content using only gene count data. The gene count model of Rabier et al. (2014) provides an effective and statistically rigorous test for ancient WGDs in plants, while also providing estimates of gene retention rates following individual WGD events. This model does not require sequences to be assembled at the chromosome level, or dating based on $dS$, but only annotation of genome sequence data and circumscribed gene families.

The WGD retention rates ($q$) following WGDs vary tremendously across the tree (Fig. 3-1). While a few WGDs, including those associated with *Populus trichocarpa* and *Glycine max* appear to retain a majority of the duplicated genes, only a small percentage of duplicated genes survive most ancient WGDs (Fig. 3-1). Some of the variation in retention rates may be due to the type of WGD event, and consequently, whether fractionation is biased or unbiased (Freeling 2009). During fractionation, there may be biases regarding the types of genes that are retained following WGD (Thomas et al. 2006; Li et al. 2016) as well as which genome copy’s genes are retained (Schnable et al. 2011). Unbiased fractionation suggests that gene expression and deletion affects both parental genomes equally. Recent data supports unbiased fractionation in *Glycine max*, *Populus trichocarpa*, and *Musa acuminata*, possibly
indicating that these WGDs represent autopolyploidy rather than allopolyploidy (Garsmeur et al. 2014). Interestingly, in our analyses, these lineages have WGDs with the highest $q$ estimates (Fig. 3-1).

Our analyses also suggest several reasons to be cautious about identifying ancient WGDs from gene count data. First, the model of Rabier et al. (2014) assumes that the global gene duplication and loss rates ($\lambda$ and $\mu$) are constant throughout the tree. When testing WGDs on a tree spanning a large phylogenetic distance, $\lambda$ and $\mu$ will be averaged across lineages, and this may lead to inappropriate estimates of $\lambda$ and $\mu$. For example, our analyses suggest that $\lambda$ and $\mu$ are different in the four-taxon tree (Brassicales and *Theobroma cacao*) than in other parts of the land plant tree, and this affects tests of the *Arabidopsis* $\alpha$ WGD (Fig. 3-1). Although most WGDs we tested, such as *Arabidopsis* $\beta$, appear to be relatively robust to poor estimates of these parameters, tests of WGDs on more limited phylogenies, or alternatively, tests allowing different $\lambda$ and $\mu$ values in different parts of the tree may improve the accuracy of the analyses. However, enabling local $\lambda$ and $\mu$ values on branches throughout the tree also may diminish the power to identify WGDs. For example, it may be difficult to distinguish a WGD on a branch from an increased $\lambda$ on the same branch. The gene count model also implicitly assumes that all gene families have the same $\lambda$ and $\mu$ and that all genes have equal probability of being retained following WGDs. There is strong evidence that gene retention after fractionation often is not random (Blanc and Wolfe 2004b; Seoighe and Gehring 2004; Maere et al. 2005; Rizzon et al. 2006; Makino and McLysaght 2012; Li et al. 2016) or that fractionation is not necessarily complete before a following speciation event (Schranz et al 2012; Conant 2014).
In our study, the retention rates of the *Arabidopsis* α and β WGDs for gene families that only span the root of Brassicales and *Theobroma cacao* were higher for the *Arabidopsis* α WGD and lower for the *Arabidopsis* β WGD than the retention rates for older gene families that span the root of the land plant tree. These inconsistent retention rate estimates suggest the possibility that a higher proportion of gene families that span the root of land plants were retained by the *Arabidopsis* β WGD as compared to the more recent *Arabidopsis* α WGD. This may explain why we detected this event with the land plant data set but not the 4-taxon analysis where both *Arabidopsis* α and β WGDs are on the same branch. Another explanation is that older WGDs and WGTs, such as the eudicot WGT, had not completed the fractionation process prior to the *Arabidopsis* β WGD.

A critical remaining question is whether our failure to detect several putative ancient WGDs in any analyses (Fig. 3-1) is due to the model or lack of power for the gene count method, or whether it suggests that the putative ancient WGDs did not happen. In nine out of the 19 putative ancient WGDs or WGTs tested in this study, including the WGT prior to the diversification of eudicots, we did not detect a WGD or WGT in analyses using the whole land plant tree (Fig. 3-1). However, we detected the eudicot WGT, the *Arabidopsis* α WGD, and the *Nelumbo* WGD using 4-taxon trees. There are multiple lines of evidence for the eudicot WGT, including syntenic comparisons between *Vitis vinifera* and other angiosperms (Jaillon et al. 2007; Argout et al. 2011) and comparisons between *V. vinifera* and *Amborella trichopoda* (*Amborella Genome Project 2013*). For example, a 1:3 ratio of syntenic blocks between *A. trichopoda* and *V. vinifera* would suggest a WGT occurred after the divergence of *A.*
trichopoda and other angiosperms, but before the divergence of eudicots (Amborella Genome Project 2013). Our ability to detect the eudicot WGT in analyses using the whole tree may be obscured by the presence of more recent WGDs and heterogeneity in retention rates across WGDs and WGTs when analyzing large phylogenies. Therefore, there may be benefits to refining hypothesis tests with smaller phylogenies, since we do detect the eudicot WGT in all three 4-taxon tree hypotheses (Fig. 3-4). Our simulation experiments also suggest it is possible that we did not detect the eudicot WGT in the full land plant (Fig.3-1; Object 3-1, Table S1) or eudicot (Object 3-1, Table S2) analyses because the eudicot WGT has a low retention rate in comparison to more recent WGD and WGT events (Fig. 3-4c; Object 3-1, Table S9). We might expect older WGDs to have more time for gene loss; therefore, it may be generally more difficult to detect WGDs and WGTs with increasing age. In any case, our results indicate that the retention rate associated with the eudicot WGT is low, which is consistent with previous analyses of gene family evolution (Amborella Genome Project 2013), and consequently, it is difficult to detect, especially with the complexities of testing nested WGDs and WGTs.

Simulations indicate that the gene count model lacks power to detect multiple WGDs on a single branch, especially when more recent WGDs have a higher retention rate compared to older WGDs (Object 3-1, Tables S5 and S10-S12). Therefore, some of our results should be interpreted with caution: an absence of evidence for a WGD is not evidence of the absence for a WGD. We resolved the difficulty of identifying both the Arabidopsis α and β WGDs by placing the Arabidopsis α WGD on a separate, neighboring branch. Although recent evidence suggests that this placement is not
correct (Dassanyake et al. 2011; Vanneste et al. 2014), this enables us to detect both the Arabidopsis α and β WGDs. Ideally it would be possible to include lineages that break up successive WGD events onto separate branches when testing WGD hypotheses.

Evidence for some putative ancient WGDs that we did not detect is largely circumstantial. For example, the WGD on the terminal branch leading to Phoenix dactylifera was proposed based on the distribution of dS between paralogs and some syntenic data for the largest scaffolds of the P. dactylifera genome sequence (Al-Mssallem et al. 2013). Our simulations suggest we have should have detected this WGD with 10,000 gene families if q = 0.1, but we only have power of 0.15 to detect the P. dactylifera WGD if q = 0.01. Thus, our results suggest that either the retention rate for the P. dactylifera WGD is extremely low, or that there is no ancient WGD in the lineage leading to P. dactylifera.

The hypothesis that two WGDs occurred before the diversification of grasses (i.e., Poaceae ρ and σ) is based on 2416 genes on 9 syntenic blocks in the Oryza sativa genome and 831 genes on 8 older syntenic blocks, inferred by ratios of duplicate genes, respectively (Simillion et al. 2002; Tang et al. 2010). Although genes on the Poaceae ρ and σ correspond to different median dS values (Tang et al. 2010), our results imply that only one WGD occurred before the divergence of grasses. The different dS values could be due to paralogs in genomic regions where synteny has eroded over time having higher substitution rates than paralogs maintained on syntenic blocks that are detectable by comparison of O. sativa chromosomes, and not two separate WGDs. However, in the case of Poaceae ρ and σ, our simulations indicate we have at most
power or 0.73 and 0.75 to detect the two WGDs assuming \( q = 0.1 \), and if \( q = 0.01 \), we have no power to detect both WGDs. Thus, although evidence for two WGDs preceding the diversification of Poaceae is far from conclusive, we cannot rule out the possibility that they occurred but have extremely low retention rates.

Similarly, there is some syntenic evidence for a WGD prior to the diversification of monocots (e.g., Tang et al. 2010); however, complexities of testing the monocot WGD prevent us from differentiating model performance and biology. We did not detect the monocot WGD in the observed data on the full land plant data set (Fig. 3-1; Object 3-1, Table S3). However, simulations suggest we have excellent power to detect the monocot WGD on the 4-taxon tree when \( q = 0.1 \), and power of 0.58 for \( q = 0.01 \) with 10,000 gene families (Fig. 3-4e; Object 3-1, Table S11). We detected the monocot WGD on the 4-taxon tree with the observed data (Fig. 3-4e; Object 3-1, Table S11); however the estimated \( q \) of the monocot WGD was 0.309. Based on other analyses of ancient WGDs with plant genomes (e.g., Tang et al. 2010), a retention rate of 0.309 is higher than expected, and it may be an artifact from testing many nested WGD hypotheses (Fig. 3-5e). Extensive testing revealed this result is not due to optimization error. Instead it may be due to an excess of duplicate genes in *Sorghum bicolor* compared to the other taxa in the 4-taxon tree.

We also find no evidence for WGDs that predated the divergence of angiosperms or seed plants based on the gene content of current genomes. This is not surprising, as there is only weak syntenic support for the angiosperm WGD and none for the seed plant WGD (*Amborella Genome Project* 2013). Furthermore, the distribution of \( dS \) based on a few hundred paralogs may be unreliable at such an old age, especially
since substitutional saturation alone can lead to false positives for WGDs (Vanneste et al. 2013). Our simulations indicate that even with a $q = 0.01$, we should have sufficient power to detect the angiosperm WGD with 5,000 – 10,000 gene families on the 4-taxon tree (Fig. 3-4f; Object 3-1, Table S12). In contrast, we have no power to detect the seed plant WGD, even if $q$ is as high as 0.1 (Fig. 3-4f; Object 3-1, Table S12). Thus, although again we cannot rule out an ancient WGD preceding the angiosperms if $q < 0.01$, our analyses provide some reasons to question the angiosperm WGD. Due to lack of power, we cannot evaluate the putative seed plant WGD based on gene count data, suggesting it may be difficult to find conclusive evidence for such extremely ancient WGDs.

Additional experiments with the Arabidopsis α and β WGDs indicate that gene families have different probabilities of being retained following WGD (Freeling 2009, Li et al. 2016). Although we did not consider differences in gene functions or gene dosage balance (Edger and Pires 2009), we used different data filtering strategies and gene family age to show differential background duplication and loss rates ($\lambda$ and $\mu$) as well as probabilities of retention following WGD across gene families. While we primarily use the gene count model as a means of detecting WGDs, it is possible to use a model testing approach for more targeted hypothesis testing of gene family evolution following WGDs, where gene function or ontological categories are known.

Our results suggest the need to critically evaluate evidence for some ancient land plant WGD hypotheses, as well as the role these ancient WGDs have played in shaping plant genomes. Multiple lines of evidence, including this study, indicate that many duplicate gene copies in some genomes such as Populus trichocarpa and Glycine max
are products of WGDs (Tuskan et al. 2006; Schmutz et al. 2010). However, our extremely low retention rate estimates, which are consistent with relatively stable gene numbers across land plants, indicate that the contributions of many ancient WGDs in terms of gene content often appear to be relatively minor. In several cases, we cannot distinguish the absence of a putative ancient WGD from its presence with an extremely low gene retention rate. While this model makes some simplifying assumptions, such as global background λ and μ, it is one of the few statistically rigorous approaches that has been used to evaluate evidence for ancient plant WGDs, and it can be used with taxa that are not completely mapped or sequenced. This approach also could be extended to taxa with only transcriptome data if the number of missing gene copies can be estimated (see Rabier et al. 2014). Characterizing ancient WGDs may be inherently difficult as information is lost with age, but our results suggest that gene copy data can provide new insights into plant ancient WGDs.

**Material and Methods**

**Gene Family Data**

We obtained gene counts from 30,023 orthogroup clusters circumscribed by the *Amborella* Genome Project (2013) using OrthoMCL (Li et al., 2003). We first filtered the gene count data to remove any families that did not span the root of the land plant phylogeny to eliminate any gene families that arose *de novo* within the land plants (Fig. 3-1). This *land plant data set* includes gene families that have at least one copy in *Physcomitrella patens* and at least one copy in another taxon. Failing to condition gene count data with at least one copy in each clade spanning the root of the phylogeny can lead to biased estimates of λ and μ (Rabier et al. 2014). We also created gene count data sets that were filtered to include at least one copy in *Amborella trichopoda* and at
least one copy in a monocot lineage (the *monocot data set*), at least one copy in *Amborella trichopoda* and another in a eudicot lineage (the *eudicot data set*), and one in *Theobroma cacao* and another in a Brassicales taxon (the *Brassicales data set*; Fig. 3-1). Likelihood calculations based on probabilities of gene count data can be extremely memory intensive when there are no limits to ancestral gene family sizes. Therefore, we removed gene families with ≥100 copies in any taxon from all gene count data sets, and we set the ancestral gene family size to a maximum of 100. The filtering process resulted in data from 7564, 10795, 11249, and 12957 gene families in the land plant, monocot, eudicot, and Brassicales gene count data sets respectively.

**Phylogenetic Tree**

We used a species tree with a topology that corresponds to our current understanding of land plant relationships (Fig. 3-1; e.g., Soltis et al. 2011; Ruhfel et al. 2014). The relationships of the taxa used here are generally well supported, although there is some disagreement about the position of *Populus* (see Sun et al. 2015). We obtained estimates of the ultrametric branch lengths for most branches in the angiosperms based on the dating analysis of Bell et al. (2010), who used a BEAST analysis (Drummond and Rambaut 2007) based on 36 minimum fossil age constraints, which were treated as exponential distributions. Divergence times for *Physcomitrella* and the other land plants, *Selaginella* and the angiosperms, and the two *Solanum* species were obtained from TimeTree (http://www.timetree.org; Hedges et al. 2006). The species tree with branch lengths is available as supplemental data. We used ultrametric branch lengths in million years, since we assume gene duplication and loss is a function of calendar time as opposed to evolutionary time measured, for instance, as a number of substitutions per sites from selected genes.
Timing of Hypothesized WGD Events

We identified hypotheses for ancient WGD events along the land plant phylogeny from primary literature. Support for these hypotheses comes from various lines of evidence, including synteny, $dS$ plots, and mapping gene duplication events onto a phylogeny.

There is much evidence in support of multiple WGDs in the evolutionary history of Arabidopsis thaliana (e.g., Arabidopsis Genome Initiative 2000; Vision et al. 2000; Blanc et al. 2000; Raes et al. 2003). The most recent WGD in the A. thaliana lineage (Arabidopsis α) is thought to be shared with most members of the Brassicaceae (The Brassica rapa Genome Sequencing Project Consortium 2011); however, based on the Carica papaya genome, it is not shared across the Brassicales (Ming et al. 2008). An older WGD event in the A. thaliana lineage (Arabidopsis β) also is not shared by C. papaya (Ming et al. 2008; Argout et al. 2011), but is shared by all sequenced Brassicaceae (The Brassica rapa Genome Sequencing Project Consortium 2011; Dassanayake et al. 2011). Bowers et al. (2003) estimated Arabidopsis α at 14.5-20.4 mya based on the shared syntenic regions with other partially sequenced angiosperm genomes. More recently, analyses of syntenic regions shared between Arabidopsis thaliana and Thellungiella parvula revealed that the Arabidopsis α WGD occurred prior to the divergence of Arabidopsis and Thellungiella (Dassanayake et al. 2011), and Vanneste et al. (2014) dated the Arabidopsis α WGD at ~48 mya. We originally tested the Arabidopsis α and β on the branch prior to the divergence of A. thaliana and T. parvula, with Arabidopsis α placed at 48 mya and Arabidopsis β placed at 82.999999 mya. However, we could not detect Arabidopsis β under any conditions when Arabidopsis α and β were on the same branch (Object 3-1, Table S5 and S10).
Therefore, we placed *Arabidopsis* α on the tip leading to *A. thaliana* at 15.5 mya (Fig. 3-1). The timing of *Arabidopsis* β is uncertain, so we placed it at 57 mya, the mid-point of the branch leading to Brassicaceae (Fig. 3-1). Placing the *Arabidopsis* α and β WGDs on two separate branches allowed us to then detect both WGDs on the 4-taxon tree of *Theobroma cacao* and Brassicales.

*Populus trichocarpa* has over 45,000 genes, approximately 8,000 of which are thought to be retained from a recent WGD (Tuskan et al. 2006). Comparisons with *Salix* ESTs indicate the WGD occurred before the divergence of *Populus* and *Salix*, but this also suggests *P. trichocarpa* has a slow rate of synonymous substitution when compared to *Arabidopsis* (Tuskan et al. 2006). Estimating the age of the *P. trichocarpa* WGD based on the distribution of $dS$ between *P. trichocarpa* paralogs with the *A. thaliana* mutation rate calibration (Lynch and Conery 2000) places the WGD event at 8-13 mya, after the divergence of *Populus* and *Salix* (Tuskan et al. 2006). Therefore, we placed the putative *P. trichocarpa* WGD at 49.5 mya, the midpoint of the *P. trichocarpa* branch on the species tree.

WGDs are prevalent across the Fabaceae (Cannon et al. 2015), including a lineage specific WGD in *Glycine max* (Schmutz et al. 2010) and another WGD at the root of Papilionoideae (Schmutz et al. 2010; Young et al. 2011). Comparison of gene pairs on syntenic segments in *Glycine max* and other genomes indicated that these events occurred around 13 mya and 59 mya respectively (Schmutz et al. 2010). We tested the *G. max* WGD at 13 mya and moved the Papilionoideae WGD back to 68 mya so it could be on the branch preceding the divergence of *G. max* and *Medicago truncatula* in our species tree.
Comparisons of syntenic gene pairs shared between *Solanum lycopersicum* and *Solanum tuberosum* indicate a WGT predating the divergence of these lineages (The Tomato Genome Consortium, 2012). The age distribution of *dS* between syntenic paralogs in *S. lycopersicum* dates this event anywhere from 52 mya to 90 mya. We placed the WGT at 71 mya, on the branch shared by the *S. lycopersicum* and *S. tuberosum* lineages.

Columbine tetraploidy was proposed based on the distribution of *dS* from 178 paralogous gene pairs (Cui et al. 2006). This was one of the first studies to identify a potential ancient WGD in the Ranunculales. It is difficult to estimate an age for this putative WGD event because there are no other species with genomic data along a lineage to provide some bounds for the event. Therefore, we placed the Ranunculales WGD at 64.5 mya, halfway along the tip leading to *Aquilegia coerulea* (the only Ranunculales taxon in the tree).

Both syntenic evidence and distributions of *dS* support a WGD specific to the *Nelumbo* lineage, that is not shared by *Vitis*, and is much more recent than the WGD common to all angiosperms (Ming et al. 2013). Dating of paralogous genes in *N. nucifera* suggests the hypothesized WGD occurred 54-76 mya. Therefore, we placed the *Nelumbo* WGD at 65 mya. The *Nelumbo* WGD was only tested in analyses where the number of genes at the root of the species tree for each gene family was distributed as a geometric distribution with a mean of 1.5.

Comparisons of paralogs in *Vitis vinifera* and their homologs in *Oryza sativa*, *Populus trichocarpa*, and *Arabidopsis thaliana* suggest a paleohexaploidization event preceding the divergence of *V. vinifera* and eurosids (Jallon et al. 2007). Observations
of ratios of syntenic blocks across plant genomes also suggest a WGT before the divergence of eudicots (Lyons et al. 2008; Argout et al. 2011; Amborella Genome Project 2013) in addition to evidence from gene trees (Jiao et al. 2012).

Evidence suggests two whole genome duplications (\(\rho\) and \(\sigma\)) may have preceded the diversification of grasses. The \(\rho\) WGD was estimated to occur around 70 mya, and the \(\sigma\) WGD was estimated to have occurred around 130 mya (Tang et al. 2010). Both of these WGDs are supported by syntenic data, and the age of each event was estimated using distributions of \(dS\). However, estimating the precise timing of older WGDs, such as Poaceae \(\sigma\), is difficult due to saturation of synonymous substitutions. Thus, we placed \(\sigma\) near 96 mya, close to the divergence of \textit{Musa acuminata} and Poaceae.

Comparisons of the \textit{Oryza sativa} and \textit{Vitis vinifera} genomes by Tang et al. (2010) revealed some shared synteny, suggesting an additional WGD may have occurred in the monocot lineage before Poaceae \(\sigma\). We tested this hypothesis using gene count probabilities by placing a WGD on the branch leading to the most recent common ancestor of monocots near 130 mya.

Three WGDs in the \textit{Musa acuminata} lineage after the split between Zingiberales and Poales also have been proposed (D'Hont et al. 2012). Syntenic blocks of paralogous genes within \textit{M. acuminata} suggest these three duplications are not shared with the other grass species in our tree, and the distribution of \(dS\) between paralogous gene pairs dates two duplications to \(\sim 65\) mya (the most recent is referred to as \textit{M. acuminata} \(\alpha\) and the next \textit{M. acuminata} \(\beta\)), near the Cretaceous-Tertiary boundary. The third duplication (\textit{M. acuminata} \(\gamma\)) is estimated to have occurred \(\sim 100\) mya based on the age of paralogs in 12 syntenic blocks (D'Hont et al. 2012). These syntenic blocks
are not homologous to the syntenic blocks in grasses that indicate the grass ρ and σ
duplications (Tang et al. 2010).

An analysis of the distributions of $dS$ of *Musa acuminata, Oryza sativa, Sorghum
bicolor, Brachypodium distachyon, Phoenix dactylifera* and *Arabidopsis thaliana*
proteomes based on predicted gene models suggests an additional WGD within the *P.
dactylifera* lineage that occurred after the divergence of Zingiberales and Arecales (Al-
Mssallem et al. 2013). In the *P. dactylifera* genome sequence, 4215 genes out of 41660
annotated gene models were paralogous and were arranged in 411 collinear blocks.
There was a bimodal distribution of $dS$, but it is uncertain if this is due to a *P. dactylifera*
specific WGD or an older shared WGD, such as a WGD predating monocots or
angiosperms (Al-Mssallem et al. 2013). Because of the uncertainty in the date of this
event, we tested the hypothesized WGD at 59 mya on the branch leading to *P.
dactylifera*.

Phylogenetic approaches dating of gene families inferred that WGDs took place
before the diversification of angiosperms and seed plants. Based on different
orthogroup filtering methods, the WGD preceding angiosperms was estimated to have
occurred at 192, 210, or 234 mya, while the WGD predating seed plants was estimated
to have occurred at 319, 321, or 347 mya (Jiao et al. 2011). Comparison of *Amborella
trichopoda* genome assembly to itself revealed 47 syntenic blocks, which contained 466
gene pairs (*Amborella Genome Project 2013*). Shared synteny with *Vitis vinifera*
suggests that the WGD predated the divergence of *Amborella* and other angiosperms,
but there was no syntenic evidence of the duplication predating seed plants (*Amborella
Genome Project 2013*). The angiosperm WGD was set at 200 mya, and the age of the
seed plant WGD was set at 350 mya (but see subsequent analyses in section “Exploring the Timing of Ancient WGD events” in which we tested different dates).

Evidence for a WGD in the *Physcomitrella patens* lineage is based on a distribution of $dS$ from paralogs in the *P. patens* genome (Rensing et al. 2007). The age of the *P. patens* WGD was estimated to be 45 mya, which is where we placed it in our tree.

**Whole Genome Duplication Retention and Loss Rates**

We used the R package WGDgc (Rabier et al. 2014), run with R version 3.0.2 (R Development Core Team, 2013), to test hypotheses of ancient WGDs across land plants and to estimate the rates of retaining duplicated genes following independent WGD events. The land plant, eudicot, monocot, and Brassicales data sets were run under a range of priors. Specifically, the number of genes at the root of the tree is assumed to be a geometrically distributed random variable, and the expected number of genes at the root is our prior parameter. Based on distributions of average gene family sizes (Fig 3-2a), gene duplication and loss rates (Fig 3-2b), and the likelihood scores given the data and models, the most appropriate prior was a geometric mean of 1.5. Having a prior geometric mean >1 allows for more than one gene copy from a gene family at the root of the tree, while not favoring too many copies at the root. When using a higher prior, gene birth rates go to 0 in the monocot and Brassicales data sets (Figure 3-2b). We found that estimated retention rates following WGDs are relatively robust to the choice of prior (Object 3-1, Tables S1-S4), which is consistent with simulations in Rabier et al. (2014).

The likelihood function is calculated as $L(D | \lambda, \mu, e, q)$, where $D$ is the set of gene family sizes $\{D_1, D_2, \ldots, D_n\}$, $\lambda$ and $\mu$ are the gene duplication and loss rates, $e$ is the set
of WGD and WGT events of known placement on the tree, and \( q \) is the set of retention rates at these WGD events. Gene families are assumed to evolve independently. Therefore, the likelihood \( \prod_{i=1}^{n} L(D_i | \lambda, \mu, e, q) \) is the product of likelihoods for individual gene families. The likelihood for a single gene family is \( \mathbb{P}\{D_i\} / \mathbb{P}\{\text{family } i \text{ is retained}\} \). \( \mathbb{P}\{D_i\} \) is calculated on a given species tree from the tips to the root using a postorder tree traversal similarly to Felsenstein’s pruning algorithm (Felsenstein 1981), except that all values for the number of genes must be integrated at each node in the tree. On edges that do not include a WGD (or WGT) event, this algorithm was described by Csűrös and Miklós (2009). WGD events break edges into segments that include the background duplication and loss of genes only, and segments that include WGD events only (Rabier et al 2014). On edge segments with a single WGD event, the recursive algorithm uses transition probabilities governed by the retention rate at the event, as described by Rabier et al (2014). \( \mathbb{P}\{\text{family } i \text{ is retained}\} \) is dependent on the filtering decision. For instance, if all gene families retained for analysis are those with at least one gene copy, then \( \mathbb{P}\{\text{family } i \text{ is retained}\} = \mathbb{P}\{D_i \neq (0, \ldots, 0)\} \), which is calculated recursively on the species tree from the tips to the root as before (Rabier et al 2014). Here, we filtered gene count data to retain families such that at least one gene copy is present in the subtrees left and right of the root of the species tree. This filtering is reflected in the calculation of \( \mathbb{P}\{\text{family } i \text{ is retained}\} \), again using a recursive algorithm. Thus, the likelihood function and MLEs of \( \lambda, \mu, \) and \( q \) depend only the timing of a WGD on a fixed edge \( e \), as well as filtering gene count data to span the root of the species tree.
After convergence of the likelihood scores for all runs, we performed a series of LRTs to determine the significance of individual putative WGD events within the land plant phylogeny. Models were nested by removing only one WGD or WGT at a time and comparing this to the model with all WGD or WGT events. This was done to avoid WGD events of large effect from influencing $\lambda$ and $\mu$, creating dependence on the order WGD events were tested. Probabilities of likelihood ratio test statistics based on mixture densities are given in Rabier et al. (2014). A nominal probability of a type I error of 0.001, such that a significant LRT statistic must be $> 9.55$, was applied to all tests. WGD retention rates were based on the final model, where all putative ancient WGDs are in the tree.

**Exploring the Timing of Ancient WGD Events**

While there is uncertainty in the precise timing of all ancient WGDs, we explored more thoroughly the timing of the putative WGDs preceding the diversifications of angiosperms and seed plants since they reside on a very long branch (Fig. 3-1). Specifically, we tested the angiosperm WGD and seed plant WGD on five million year intervals between 148-349 mya (angiosperm) and 350-420 mya (seed plant). All retention rates were optimized for each run, and starting gene duplication and loss rates were provided from the land plant analyses, 0.0016 and 0.0019 respectively, to speed computation.

**Testing Arabidopsis Alpha and Beta Duplication Events**

In our initial analysis of the land plant data set, we failed to detect the *Arabidopsis* $\alpha$ WGD, but we detected the *Arabidopsis* $\alpha$ WGD using the Brassicales data set. This could be due to: 1) power alone (because the Brassicales data set contains more gene families), 2) different rates of gene retention in Brassicales specific
gene families than in gene families that span the land plant root, or 3) differences in $\lambda$ and $\mu$ within the Brassicales than the rest of the land plant tree. To address power and retention rate differences among gene families, we randomly sampled without replacement 7567 (i.e., the number of gene families in the land plant data set) of the 12957 gene families in the Brassicales data set 500 times, and for each randomly sampled data set, we estimated $\lambda$, $\mu$, and each $q$ using a geometric mean of 1.5 as the prior distribution of the number of genes at the root. To observe if the Brassicales specific gene families were driving the result, the Brassicales model likelihood was optimized using only the 6843 gene families shared between the land plant data and the Brassicales data, as well as only 2170 Brassicales data set specific gene families (that is, families with no sequence outside of Brassicales and *Theobroma cacao* in our tree) and only the 10489 angiosperm specific gene families (that is, families with no sequences outside of angiosperms). To speed up optimization for these experiments and make them computationally tractable, we used the estimates of $\lambda$ and $\mu$ from the original Brassicales data set as starting values.

To address the effects of $\lambda$ and $\mu$ on the $q$ values of *Arabidopsis* $\alpha$ and $\beta$ and on the LRT statistics for these two WGD events, we calculated the likelihood score of each randomly resampled Brassicales data set in several ways. First, for each of the 500 resampled data sets, $\lambda$, $\mu$, and retention rates for both the *Arabidopsis* $\alpha$ and $\beta$ WGDs were unconstrained and optimized by maximum likelihood. Next, we calculated the likelihood score by fixing $q$ for both the *Arabidopsis* $\alpha$ and $\beta$ WGDs to their previously optimized values, and fixing $\lambda$ and $\mu$ to the estimates from the analysis of the land plant data set, 0.0016 and 0.0019 respectively. LRTs were calculated by removing either the
*Arabidopsis* α or the *Arabidopsis* β WGD event. If a WGD is not statistically significant using the land plant λ and μ, then it indicates that gene duplication and loss rates estimated from the land plant tree do not adequately describe gene duplication and loss rates in the Brassicales. Using the land plant estimates of λ and μ allowed us to observe the effects of enforcing inappropriate gene duplication and loss rates on our ability to detect WGDs.

**Testing Multiple Duplications on a Single Branch**

Testing the grass ρ and σ as well as the *Musa acuminata* α, β, and γ WGD events required further attention. For instance, since the estimation of q depends on the timing of the WGD event, a model without grass ρ is not equivalent to σ. Therefore, we had to explore and test all possible ways that WGD events can be removed from a branch. For testing *M. acuminata* α, β, and γ, all 9 nested hypotheses were explored (Fig. 3-3). For example, the *M. acuminata* branch with both the α and β WGDs is preferable to just α; however, having α, β, and γ improves the likelihood score further, which is equivalent to *M. acuminata* with γ alone.

**Simulation Experiments**

We extended the WGDgc R package (version 1.2) to simulate gene count data along a phylogeny. Simulations were used to estimate our power to detect a WGT or two WGDs on a single branch. Specifically, we ran experiments to test four sets of WGDs or WGTs: 1) The eudicot WGT, 2) *Arabidopsis* α and β, 3) The monocot WGD, *Phoenix dactylifera* WGD, Poaceae ρ WGD, and Poaceae σ WGD, and 4) Two WGDs predating the divergence of *Amborella trichopoda* and other angiosperms (i.e., the angiosperm and seed plant WGDs). All simulation experiments were performed with 4-taxon trees to make them computationally feasible. The 4-taxon trees used in the
simulation experiments are shown in Fig. 4, and newick and simmap formatted trees with branch lengths are provided in the supplementary material.

The eudicot WGT simulations were performed for the duplication and loss rates estimated from the eudicot data ($\lambda = 0.0022$, $\mu = 0.0018$). Likewise the *Arabidopsis* $\alpha$ and $\beta$ WGDs were simulated using the gene duplication and loss rates estimated from the Brassicales data ($\lambda = 0.00135$, $\mu = 0.00211$), monocot lineage WGDs were simulated under the monocot gene duplication and loss rates ($\lambda = 0.00249$, $\mu = 0.00288$), and the angiosperm and seed plant WGDs were only simulated using the land plant duplication and loss rates ($\lambda = 0.00162$, $\mu = 0.00196$). The eudicot WGT simulations tested $q$ of 0.01 or 0.10 such that either all WGDs/WGTs had equal retention rates or the eudicot WGT had $q = 0.01$ while more recent WGDs/WGTs had $q = 0.10$. We allowed unequal retention rates in the *Arabidopsis* $\alpha$ and $\beta$ simulations, such that $q_\alpha = q_\beta = 0.01$, $q_\alpha = q_\beta = 0.10$, or $q_\alpha = 0.10$ and $q_\beta = 0.01$. Because the monocot simulations were more complex than other simulation scenarios we only allowed all WGDs to have $q = 0.01$ or $q = 0.10$. The angiosperm and seed plant WGDs were simulated only with equal retention rates of $q_{\text{angiosperm}} = q_{\text{seed plant}} = 0.01$ or $q_{\text{angiosperm}} = q_{\text{seed plant}} = 0.10$.

For each experiment, we ran 100 simulation replicates with 500, 1000, 5000, and 10000 simulated gene families. All simulated data used a geometric distribution of genes at the root of each gene family, with a mean of 1.5. Data were conditioned to span the root of the species trees used to test the eudicot WGT, the *Arabidopsis* $\alpha$ and $\beta$ WGDs, and the angiosperm and seed plant WGDs respectively. The trees and R scripts used to run simulations are available as supplementary material.
Figure 3-1. Ultrametric land plant phylogeny with putative ancient WGD or WGT events plotted as circles on branches. Numbers displayed next to events are retention rates ($q$) optimized using a prior geometric mean of 1.5. Events associated with significant LRTs for analyses with land plant data are in blue and non-significant events are colored grey. Some events are detected in subsequent analyses but not the land plant analyses, while some events are not consistently detected across all analyses. An “*” next to a retention rate indicates a WGT rather than a WGD. The Venn diagram shows how many gene families were used for each data set and the overlap between data sets. The data sets depicted in the Venn diagram refer to the different data filtering strategies. A 4-taxon tree formed by *Theobroma cacao* and Brassicales is shown in the zoomed-in portion with the *Arabidopsis* α and β WGD events. 12957 gene families span the root of the 4-taxon tree (the Brassicales data), and optimizing model parameters causes a significant LRT for both *Arabidopsis* α and β as well, as higher retention rates.
Figure 3-2. Distributions of gene family sizes. a) The distribution of mean number of genes per gene family across the 22 taxa used in this study is shown using kernel densities for the land plant, eudicot, monocot, and Brassicales filtered data sets. The mean of each distribution is shown by a vertical dashed line. These are all close to 1.5, which provides some justification of a geometric mean of 1.5 as the best prior choice. b) The prior used for each analysis is plotted against the ratio of the estimated duplication and loss rates. A prior of one enforces a single gene at the root of the species tree for likelihood calculations, but this can lead to inflated duplication rates. Allowing a prior geometric mean of 1.5 allows for some uncertainty in the number of genes at the root, without driving duplication rates to 0 or inflating loss rates, which happens when using higher priors.
Figure 3-3. Testing multiple WGD hypotheses on a single branch using results for *Musa acuminata* α, β, γ for the land plant data set. Blue circles represent WGDs present in the model and grey circles indicate absence. Although *M. acuminata* α and β is preferable to a model with only *M. acuminata* β, the model with all three WGDs (top) is statistically equivalent to a model with only *M. acuminata* γ (bottom left). Retention rates for each WGD for each model are displayed below each event.
Figure 3-4. Gene count data was simulated on 4-taxon trees to test WGD hypotheses and estimate power. Events associated with significant LRTs in the observed data are in blue while non-significant events are colored grey. An "***" above a circle indicates a WGT rather than a WGD. Numbers above circles indicate the simulated retention rates and numbers below circles indicate power. Panels a, b, and c were designed to test the eudicot WGT. Panels d, e, f were designed to test 2 WGDs on a single branch. Panel d estimates power for testing the *Arabidopsis* α and β WGDs on the same branch. Panel e simulates a complex scenario of WGDs in the monocot lineage for the monocot WGD, *Phoenix dactylifera* WGD, Poaceae ρ WGD, and Poacea σ WGD. Panel f is used to assess our ability to detect the angiosperm and seed plant WGDs.
Figure 3-5. Testing effects of background rates on WGD detection. Circles represent data sets of 7564 genes randomly sampled without replacement from the Brassicales data set, while the square represents the land plant data set with 7564 gene families that also span the root of Brassicales and *Theobroma cacao*. The triangle represents gene families that span the root of angiosperms as well as Brassicales and *T. cacao*, and the diamond represents gene families that only span the root of Brassicales and *T. cacao*. The size of each point represents retention rate quartiles. A significant LRT statistic is colored blue and non-significant LRTs are grey. Panels a and c show the estimated λ, μ, and q for the *Arabidopsis* α and β WGDs respectively, as well as the result of the LRT statistic. In panels b and d, the likelihoods were calculated by fixing the retention rates to their values optimized in panels a and c and by further fixing λ and μ to 0.0016 and 0.0019 from the land plant data. Data points are plotted according to λ and μ estimated in panels a and b to separate the points and to better show effects of inappropriate parameter estimates on testing WGD hypotheses with gene count data.
Object 3-1. Supplementary material for Chapter 3
CHAPTER 4
ASSESSING THE PERFORMANCE OF Ks PLOTS FOR DETECTING ANCIENT WHOLE GENOME DUPLICATIONS

Introduction

Genomic data have revealed evidence for previously undetected ancient whole genome duplications (WGDs) in many eukaryotic lineages, including angiosperms (e.g. Schlueter et al. 2004; Cui et al. 2006; McKain et al. 2012; Vanneste 2014), gymnosperms (Li et al. 2015; Guan et al. 2016), ferns (Vanneste et al. 2015), mosses (Rensing et al. 2007; Szövényi et al. 2015; Johnson et al. 2016; Devos et al. 2016), teleost fishes (e.g. Taylor et al. 2003; Jaillon et al. 2004; Crête-Lafrenière et al. 2012), horseshoe crabs (Nossa et al. 2014; Kenny et al. 2016), and most recently spiders (Clark et al. 2015; Schwager et al. 2017). Ancient WGDs often are inferred from the distribution of synonymous substitution (Ks) rates among paralogous genes within a genome, which is visualized in a “Ks plot” (Cui et al. 2006; Barker et al. 2008; Vanneste et al. 2014). In the absence of WGDs or large episodic duplications, the synonymous substitutions between paralogs within a genome should follow an exponential distribution (Lynch and Conery 2003). WGDs should produce additional normally distributed peaks in the Ks plots (Blanc and Wolfe 2004; Schlueter et al. 2004). The age of the ancient WGDs can be estimated from the synonymous evolutionary distances at the peaks (Maere et al. 2005).

Ks plot analyses require only genomic or transcriptomic sequence data from a single taxon, and can be relatively quick and easy, especially with the aid of existing bioinformatic pipelines (e.g. Lyons et al. 2008; Barker et al. 2010). However, Ks plots can be difficult to interpret, and our ability to identify WGDs from Ks plots is unclear. For
example, substitutional saturation can produce peaks in the \( K_s \) plots that do not reflect WGDs (Vanneste et al. 2013), although the effects of saturation may be ameliorated by excluding high \( K_s \) estimates from the \( K_s \) plots (e.g., Lynch and Conery 2000; Cui et al. 2006; Barker et al. 2008; Tang et al. 2010; Vanneste et al. 2013; Vanneste et al. 2014). Analyses sometimes fail to detect peaks representing ancient WGDs from \( K_s \) plots as well (e.g. Johnson et al. 2016), but our ability to recover evidence of ancient WGDs from \( K_s \) plots may be due to methodological choices rather than biological variation in duplicate gene retention. Specifically, many studies use pairwise estimates of \( K_s \) (e.g. Schlueter et al. 2004; Ming et al. 2013; Nossa et al. 2014; Johnson et al. 2016). Pairwise \( K_s \) estimates are problematic as they are more susceptible to saturation than phylogenetic estimates (Yang 1994) and introduce many more data points into \( K_s \) plots than there are duplicated genes (Blanc and Wolfe 2004; Cui et al. 2006; Rensing et al. 2007; Barker et al. 2008). Thus, the average \( K_s \) estimates for nodes of hierarchical clusters computed from pairwise \( K_s \) estimates (Blanc and Wolfe 2004; Maere et al. 2005; Cui et al. 2006; Barker et al. 2008) or phylogenetic estimates (Rensing et al. 2007; Olsen et al. 2016) have been preferred and are expected to increase resolution of WGDs in \( K_s \) plots.

The age of a WGD may also impact the efficacy of \( K_s \) plot analyses. For example, WGDs with peaks at \( K_s = 0.5 \) can be difficult to differentiate in \( K_s \) plots when many genes are lost (Cui et al. 2006), because the small evolutionary distances between homeologs resulting from the WGD are masked by recently duplicated genes. A WGD also could be too old to be detected with a \( K_s \) plot if too many duplicate genes from the ancient WGD are lost (Paterson et al. 2004; Connant 2014). One approach to
highlight a weak WGD signal from Ks plots is to include only paralogs that likely emerged from a WGD. A WGD can produce large syntenic segments within a genome (e.g. Kellis et al. 2004; Tuskan et al. 2006), and examining Ks plots built from only syntenic paralogs can make the WGDs more apparent (e.g. Tang et al. 2008; Tang et al. 2010; Amborella genome project 2013; Myburg et al. 2014).

The null hypothesis for Ks plot analyses assumes exponentially distributed evolutionary distances among paralogs, implying that gene duplication and loss rates have remained constant over time (Cui et al. 2006; Soltis et al. 2011). Variation in gene duplication and loss rates through time, any episodic burst of gene duplication and not necessarily a WGD, could conceivably affect the interpretation of Ks plots. Large segmental duplications are often invoked as alternative explanations (e.g. Al-Mssallem et al. 2013), but also biased retention of other small-scale gene duplications. For example, Blanc and Wolfe (2004) detected a Ks peak from Arabidopsis thaliana genomic data caused by tandemly duplicated genes. More recently, the Octopus bimaculoides genome indicated episodic expansion of many genome families from tandemly arrayed duplications (Albertin et al. 2015). It is also plausible that increased retention of adaptively significant genes, such as the over-representation of silk-gland specific transcripts in a Ks peak predating spiders, could produce peaks in Ks plots resembling those created by WGDs (Clark et al. 2015), although recent genomic evidence suggests the house spider Ks peak represents an ancient WGD (Schwager et al. 2017). There are no known empirical cases where non-WGD events have been mistaken for ancient WGDs from Ks plot analyses; however, these non-WGD bursts of duplicate genes may confound statistical analyses of Ks plots.
WGDs often are identified from a visual inspection of \( K_s \) plots from individual species (e.g. Vanneste et al. 2015). Yet an accurate, scalable, and ideally less subjective, model-based approach is desirable. Univariate mixture models can be used to identify and estimate the timing of WGDs (e.g. Schlueter et al. 2004; Cui et al. 2006; Barker et al. 2008; Vanneste et al. 2014; Barker et al. 2016). The most commonly used approach is to maximize a likelihood function that fits one or more \( (k) \) normal probability distributions to a \( K_s \) plot. Each normal distribution for \( k \) greater than one represents a putative WGD. However, this approach can overfit distributions (e.g. Johnson et al. 2016), leading to overestimates of the number of ancient WGDs, and consequently, the results are commonly ignored (e.g. Vanneste et al. 2015).

In this study, we used gene family simulation experiments to explore the effects of WGD age and gene retention rate following WGDs on our ability to infer WGDs from \( K_s \) plots and estimate ages of WGDs with mixture models. We varied the age of WGDs, in \( K_s \), to better understand the evolutionary distance range at which \( K_s \) plot analyses are effective. In spite of the many alternative hypotheses that could explain peaks in \( K_s \) plots aside from WGDs, there has been no attempt, that we are aware of, to quantify how much gene retention is necessary to produce a peak in a \( K_s \) plot analysis. Duplicate gene retention following WGD in plants is highly heterogeneous with as little as 7% of genes retained in the columbine-specific WGD and over 75% of genes retained from most recent soybean WGD (Tiley et al. 2016). Thus, we also varied duplicate gene retention from 0% to 100%, complete retention of a WGD. We also evaluated different methods to construct \( K_s \) distributions and infer WGDs from \( K_s \) plots, including using pairwise or tree-based (i.e. node-averaged) estimates of \( K_s \), different
mixture models, and limiting the $K_s$ plots to genes involved in the WGD. Our simulations account for variation in gene duplication and loss rates across gene families, which was not included in previous $K_s$ plot simulation studies (Cui et al. 2006; Vanneste et al. 2013), as well as heterogeneity in substitution rate across branches in each gene family.

**Results**

**Detecting WGDs Using Node-Averaged $K_s$**

Accurately estimating the number of WGD events using mixture models is challenging under many conditions. We found that mixture model analyses had difficulty identifying both the correct number of distributions that should have been evident in a $K_s$ plot as well as the age of the WGD distribution, regardless of whether we used the known simulated values of $K_s$ or re-estimated $K_s$ with ML methods. By analyzing both the true $K_s$ values from simulations and estimated $K_s$ values from simulated sequence data, we showed that errors made by mixture models are inherent to the mixture models themselves, and not because of errors in the estimation of $K_s$. In our simulations, if there was a WGD (i.e. retention rate > 0), we expected to detect two peaks in the distribution of $K_s$ for all paralogs, one representing the WGD and one representing the background duplications. However, we frequently detected more than two peaks, even in the absence of a WGD (i.e. retention rate = 0), especially if we were using the normal mixture model (Fig. 4-1; Object 4-1, Table S1). Using mixtures of normal distributions always results in more components than the exponential + normal model (Fig. 4-1b; Object 4-1, Table S1). When estimating the $K_s$ values, rather than using the actual $K_s$ values, we were only able to estimate the number of WGDs accurately in at least 50% of the simulations when $K_s = 0.5$, the retention rate is ≤ 0.1, and we were using the
exponential + normal model (Table 4-1). When the retention rates were 0.5 or 1, we generally overestimated the number of components (Fig. 4-1; Table 1). However, when the Ks value for the WGDs was large (i.e. 1.0, 3.0, or 5.0), we often did not detect evidence of a WGD if we were estimating the Ks, using an exponential + normal model, and the retention rate was ≤ 0.1 (Fig. 4-1; Table 4-1).

Mixture models will often fit the tail of the Ks distribution with one or more extra components, and the number of components inferred from Ks plot analyses can be affected by the maximum cut-off used for Ks (e.g. Barker et al. 2008; Vanneste et al. 2015). Thus, rather than strictly interpret any identified component as evidence for a WGD, we also examined only the component whose mean Ks was closest to the actual age of the WGD. We compared the mean Ks values of gene pairs comprising these components, which were detected using the exponential + normal model, to the known Ks value of the WGD that was used to simulate the data. This comparison indicates that a WGD often is only detectable when the gene retention rate is high (i.e. > 0.1; fig. 2). When retention rates are ≤ 0.1, the closest mean Ks for the putative WGD component generally far exceeds that Ks of the WGD, and it overlaps with the component you observe when there is no WGD (i.e., retention rate = 0; Object 4-1, Figs. S1-S4). While the mean Ks of gene pairs from the putative WGD component from the normal mixture model analyses overlaps with the true age of the WGD at the lowest retention rate of 0.01, a peak is also found at this Ks when there is no WGD (Object 4-1, Figs. S1-S4).

In the simulations with a WGD peak mean of Ks = 0.5, the WGD is detectable using the exponential + normal model when the retention rate is 1.0 and the Ks values of the gene pairs are known. In the simulations with a WGD at Ks = 0.5 and the
retention rate is 0.5, we can distinguish a WGD in some resampled data sets, which represent a set of possible random gene loss scenarios following WGD, when we know the gene pair $K_s$ values, although there is much overall error (Fig. 4-2a; Object 4-1, Fig. S1). However, when we used estimated gene pair $K_s$ values, the WGD is detectable with little error (Fig. 4-2c; Object 4-1, Fig. S1). When $K_s = 0.5$, the normal mixture model finds the WGD when optimizing the number of components, but this peak is also found at the same location in simulations without a WGD (i.e. retention rate = 0; Figs. 4-2b and 4-2d; Object 4-1, Fig. S1).

For simulations with a WGD at $K_s = 1.0$, the WGD is detectable when optimizing the number of mixing components using the exponential + normal model when retention rates are 0.5 and 1.0 (Figs. 4-2e and 4-2g; Object 4-1, Fig. S2). Again, the mixture of normal distributions identified the WGD peaks at high retention rates when the WGD is at $K_s = 1.0$, but they are also present when there is no WGD (Figs. 4-2f and 4-2h; Object 4-1, Fig. S2). For the WGDS at $K_s = 3.0$ and $K_s = 5.0$, the exponential + normal and normal mixture model performed similarly. A peak was detectable at the time of the WGD, but a peak was also found in the complete absence of a WGD (Object 4-1, Figs. S3 and S4).

Instead of optimizing the number of components in the mixture models, we also constrained the number of components to 2 (i.e. the correct number of components when the retention rate is > 0) and optimized the mixing distribution parameters to reveal scenarios where WGDs were most consistently detectable (Fig. 4-3). When a WGD occurs at $K_s = 0.5$, the second component mean is outside of the 95% confidence interval when the retention rate is 0.5 (Fig. 4-3a). With a retention rate of 1.0, only the
exponential + normal model discriminates the WGD peak from the background distribution (Fig. 4-3a). It detects the WGD peaks for $K_s = 1.0$ at retention rates of 0.5 and 1.0 when using the normal mixture model the WGD peak still lies outside of the 95% confidence interval for the second component means (Fig. 4-3b). For WGDs at $K_s = 3.0$ and $K_s = 5.0$, the exponential + normal and normal mixture models perform similarly, at least for the $K_s$ maximum likelihood estimates (MLEs; Figs. 4-3c and 4-3d). When a WGD was at $K_s = 3.0$ with retention of 0.5 or 1.0, the second component mean was slightly older than the true value of 3.0 (Object 4-1, Table S2). The true WGD age is within the distribution of second component means for $K_s = 5.0$; however, a second peak was also detected at $K_s = 5.0$ when there was no retention of duplicates from the WGD (Fig. 4-3d; Object 4-1, Table S2).

**Detecting WGDs Using Pairwise $K_s$**

Inferring the presence and age of a WGD using pairwise estimates of $K_s$ was even less accurate than using node-averaged $K_s$. When applying the exponential + normal model, we generally inferred more components than when we applied the node-averaged analyses (Fig. 4-1a). In the case of normal mixture models, there is both overfitting and uncertainty in the optimal number of components (Fig. 4-1b). When analyzing the means of the components closest to the true ages of the WGDs across replicates, analyses using pairwise distances detected a peak in the distribution that corresponds to the WGD event when retention rate = 0.5 or 1.0, but they overestimated the WGD age (Object 4-1, Figs. S1 and S2). For older WGDs at $K_s = 3.0$ and 5.0, a peak in the $K_s$ distribution was detected at the same evolutionary distance even when there was no WGD (i.e., retention rate = 0; Object 4-1, Figs. S3 and S4). For the simulations when the WGD was at $K_s = 3.0$, there was less variation in the distribution of component
means when using the estimated $K_s$ values than when using the known pairwise $K_s$ value (Object 4-1, Fig. S3), but the converse was observed when $K_s = 5.0$ (Object 4-1, Fig. S4). This effect is likely due to saturation, which is mitigated by gene trees in the node-averaged distributions, especially when $K_s \geq 3.0$ (Fig. 4-3; Object 4-1, Table S2). Across all simulation conditions, two components alone could not correctly identify a WGD from the pairwise distances of all paralogs, particularly when using $K_s$ estimated from gene pairs. There was little difference in the performance of the exponential + normal mixture and the normal mixture models when using pairwise $K_s$ estimates (Object 4-1, Table S2).

**Analyzing “Syntenic” Data Instead of All Paralogs**

We examined whether using $K_s$ plots built from only paralogs resulting from a WGD can be used to detect and date ancient WGDs more accurately than using $K_s$ plots built from all paralogs. In these experiments we removed all data points representing paralogs that were not from WGDs to represent paralogs from large syntentic regions within a genome, and we only used node-averaged estimates of $K_s$. We did not use simulations with retention rates of 0 and 0.01, as there are no genes resulting from the WGD when the retention rate is 0 and too few genes resulting from the WGD when the retention rate is 0.01 (i.e. less than 30 nodes across 1000 gene trees).

Over-fitting distributions remained an issue with the syntenic node-averaged data (Fig. 4-4). Only a single component was expected in these analyses, since all the duplications resulted from the WGD. However, when the retention rate is 0.05, a single component was chosen for estimated distributions of $K_s$ 17%, 28%, 62%, and 76% of the time for a WGD at $K_s=0.5, 1.0, 3.0,$ and $5.0$ respectively. The number of resampled
data sets with a single component decreased when the retention rate was 0.1; a single component was chosen only 5%, 13%, 35%, and 65% of the time when $Ks=0.5$, 1.0, 3.0, and 5.0 respectively. A single normal component was never chosen using the difference in the Bayesian information criterion ($\Delta BIC$) when retention rates = 0.5 or 1.0. Analyses using known $Ks$ values estimated from gene pairs followed a similar pattern (Object 4-1, Table S3).

Again, the true WGD still may be detectable even if the optimal number of components does not reflect the number of WGDs. For both the known and estimated node-averaged $Ks$, the WGD peak was detectable when gene retention rates were 0.05, 0.1, 0.5, and 1 across all WGD ages (Fig. 4-5). Still, the mean of the component generally underestimated the true age of the WGD, except when analyzing the MLEs of $Ks$ for retention rates of 0.05 and 0.1 when the WGD was at $Ks=5.0$. Although there was some uncertainty in the timing of the WGDs for retention rates of 0.05 and 0.1 (Fig. 4-5), the WGD was discernable using normal mixture models on our simulations. However, there was a slight bias towards more recent $Ks$ values, and a single WGD using only syntenic data did not fit the expectation of a single normal distribution (Fig. 4-6).

**Performance of Mixture Models on Empirical Transcriptomic Data**

Our simulations explicitly tested the effects of using node-averaged $Ks$ compared to pairwise $Ks$ estimates and the effects of whether and exponential distribution was accounted for in a mixture model on our ability to accurately infer the presence and age of ancient WGDs. To show our simulation results were not simply an artifact of our simulation process itself, we reanalyzed $Ks$ plot evidence for previously identified ancient WGDs (Shi et al. 2010; Barker et al. 2016). Results for the empirical
transcriptome data was generally consistent with the simulation results. We found that the normal mixture model fit just as many or more components than the exponential + normal mixture model (Object 4-1, Table S4). In spite of over-fitting of mixture models, the node-averaged $Ks$ estimates always identified a peak that was consistent in both the exponential + normal and normal mixture models, while there were no consistently identifiable peaks in two out of the five empirical $Ks$ plots examined here when using pairwise $Ks$ estimates (Fig. 4-7; Object 4-1, Table S4). The consistent peaks identified in the node-averaged data here correspond to WGDs characterized in previous studies (Shi et al. 2010; Barker et al. 2016); however, the exponential + normal model did fail to identify a prominent peak near $Ks = 0.15$ in both *Actinidia chinensis* and *Actinidia deliciosa* (Fig. 4-7; Object 4-1, Table S4). It is important to note that peaks consistently identified in pairwise data for *Barnadesia spinosa* and *Acicarpha spathulata* are very close to their node-averaged estimates, implying that pairwise estimates may not always be as misleading as inferred from our simulations (Fig. 4-7; Object 4-1, Table S4).

**Discussion**

Our simulation experiments demonstrate that mixture model analyses of gene age distributions often over-fit the number of components, potentially producing false evidence of ancient WGDs when strictly interpreted. $Ks$ plot analyses with reasonable discretion are capable of detecting WGDs when at least 10% of duplicate genes are retained from the WGD, but only during a limited range of evolutionary time. Although, $Ks$ from 0.5 to 1.0 is a biologically realistic range that can cover large expanses of time in millions of years. For example, many $Ks$ plot analyses of angiosperms in this range correspond to approximately 65 MYA (Vanneste et al. 2014), but differences in
generation time and substitution rates, such as in *Pinus*, can cause a *Ks* peak of 0.25 to correspond with greater than 200 MYA (Li et al. 2015). It is difficult to detect WGDs when the gene retention rate following the WGD is low (≤10% in our simulations), especially when the WGD is relatively recent (*Ks* = 0.5 in our simulations). Thus *Ks* plot analyses have likely worked well for many plant WGDs, because at least 10% duplicate gene retention from a WGD near *Ks* = 0.5 is not uncommon, such as 16% the most recent *Arabidopsis thaliana* WGD (Maere et al. 2005), 15% and higher for many Asteraceae (Barker et al. 2009), and 24% within Amaranthaceae (Yang et al. 2015).

Together these results indicate that putative ancient WGDs should be observable in *Ks* plots and detectable with mixture models; however, analyses with mixture models introduce many other complications.

There are a number of reasons to be critical of the mixture model results from *Ks* plots, but perhaps the most troubling is the tendency to falsely detect WGDs. When analyzing WGDs with low (≤ 10%) retention rates with the exponential + normal mixture model, not only was the estimated number of components and ages of the components inaccurate, but also, the results were indistinguishable from the simulations with no WGD (i.e. 0% genes retention; Fig. 4-2; Object 4-1, Figs. S1-S4). Clearly a strict interpretation of optimal mixture model components will lead to many false positives for the signal of ancient WGDs, even when limiting the *Ks* plot analyses to include only paralogs involved in the WGD (Fig. 4-4). The extra components that do not correspond to either the WGD or background distribution are generally fitting the tails of *Ks* plots and accounting for a very small proportion of the data, as demonstrated with our empirical analyses (Object 4-1, Table S4). Normal mixture models sometimes require
more than one component to fit the background distribution alone, which explains the overall increased number of components estimated for normal mixture models compared to the exponential + normal mixture models (Object 4-1, Table S4). These results indicate that the ΔBIC score should be interpreted with great caution when estimating the number of components (i.e. ancient WGDs). Alternative model selection approaches, such as generating null distributions of LRTs with nonparametric bootstrapping, may be more appropriate for detecting ancient WGDs from Ks plots (McLachlan 1987); however, they can be intensely computationally demanding.

Although our results suggest that extremely ancient WGDs (i.e., Ks = 5.0) can be identified with mixture models when using node-based estimates of Ks (Fig. 4-3d; Tables 4-1 and 4-2; Object 4-1, Fig. S4), we find that a second peak would manifest at Ks = 5.0 regardless of whether a WGD occurred or not, similar to Vanneste et al. (2013). Ks = 5.0 is outside of the range of values considered for most Ks plot analyses, but the location of the erroneous second peak depends on where a user truncates the Ks plot. For example, random peaks manifested in nearly all of the Ks plots in our experiments; Ks plots were truncated at 5.0 for WGDs at Ks = 0.5 and Ks = 1.0, which both showed a second peak near Ks = 3.0 in the absence of a WGD (Figs. 4-3a and 4-3b). Truncating Ks plots to lower values will likely push these arbitrary peaks into biologically plausible ranges. Thus, no matter how the Ks plot is constructed, it is important to distinguish between arbitrary model fitting and lines of evidence for a WGD.

Our simulations confirm findings from previous research (Cui et al. 2006; Vanneste et al. 2013) that Ks plots perform best at a limited range of ages and levels of gene retention (Figs. 4-2 and 4-3; Object 4-1, Figs. S1-S4). Thus, we might expect Ks
analyses to produce a biased view that ancient WGDs are clustered in time. Indeed, many WGDs in plants appear to coincide with the Cretaceous-Paleogene (K-Pg) boundary based on a second Ks peak generally falling between 0.5 and 1.0 (e.g., Fawcett et al. 2009; Vanneste et al. 2014; Lohaus and Van de Peer 2016), the range in which Ks plot analyses work best in our simulations. Therefore, it seems possible that this clustering around the K-Pg boundary is a result of biases in the Ks plot analyses rather than the importance of WGDs for surviving and diversifying following the K-Pg extinction event. Although ancient WGDs near the K-Pg boundary may be associated with evolutionary adaptations, such as salt tolerance in angiosperm seagrasses (Olsen et al. 2016), there is little evidence that genes surviving from ancient WGDs contribute to adaptive traits. Methods that characterize ancient WGDs in a phylogenetic context may be better able to test hypotheses on both the clustering of WGDs near the K-Pg boundary and their association with evolutionary innovation as well as the survival and diversification of major angiosperm groups (Tank et al. 2015; Kellogg 2016).

Although our simulations provide many reasons to question the strict interpretation of mixture model analyses of Ks plots, they also offer some guidance on ways to optimize the inference of WGDs. For example, it is always better to use node-based estimates of Ks distance than pairwise distance estimates. Criticism of pairwise Ks estimates arose early in the ancient WGD literature (e.g. Blanc and Wolfe 2004), and consequently, many studies have corrected for the redundancy of pairwise estimators with neighbor-joining trees made from pairwise Ks (Blanc and Wolfe 2004; Maere et al. 2005; Cui et al. 2006; Barker et al. 2008; Vanneste et al. 2013; Devos et al. 2016) with few actually utilizing phylogenetic estimates of evolutionary distances (Rensing et al.
2007; Olsen et al., 2016). However, the use of pairwise estimates in Ks plot analyses still persists (e.g. Ming et al. 2013; Nossa et al. 2014; Johnson et al. 2016). Our simulations indicate that pairwise Ks distances are limiting compared to node-averaged estimates; a WGD with 100% gene retention cannot be distinguished from the absence of a WGD (i.e., 0% gene retention) when the WGD occurred at Ks = 3.0 or 5.0 (Fig. 4-3; Object 4-1, Figs. S3 and S4). Pairwise estimators can perform well in some empirical cases though. The differences between node-based and pairwise Ks distributions are far greater in our simulations compared to the empirical cases shown here (Figs. 4-3 and 4-7), but given that the computation cost for node-averaged estimates is low, our experiments suggest there is little reason to use pairwise distances in Ks plot analyses.

Next, using the exponential + normal mixture model typically fit fewer components, and thus resulted in fewer false positives, than using the normal mixture model (Fig. 4-1). As observed in many empirical studies (e.g., Szövényi et al. 2015; Johnson et al. 2016), when using node-averaged Ks among all paralogs, mixture models generally over-fit components. However, when the gene retention rate was high, the age of a WGD often could be accurately characterized with only 2 components, regardless of the mixture model (Fig. 4-3). Thus, while more components may be preferred by the ΔBIC, the extra components are likely fitting the tails of the distributions. One possible strategy to improve detection of WGDs is to look for consistency in a component mean, that is, cases when the estimates of a component mean are similar, regardless of how many mixing distributions are incorporated into the model. We demonstrated this for analyses of all paralogs by comparing the exponential + normal mixture model where k was constrained to 2 with the exponential + normal
mixture model where \( k \) was freely optimized. A peak was detectable at the age of the WGD with only two components (Fig. 4-3), and the approximate age of this peak remained unchanged, even with the addition of more components when \( k \) was freely optimized (Figs. 4-1 and 4-2). Additionally, peaks corresponding to ancient WGDs in empirical \( Ks \) plots were detectable in both the exponential + normal and the normal mixture model while other component means were clearly not associated with a visible peak in the \( Ks \) distributions (Fig. 4-7; Object 4-1, Table S4).

The best performance in our simulations results from using \( Ks \) plots built only from paralogs that diverged at an ancient WGD (e.g. paralogs on large syntenic regions within a genome). Unfortunately, syntenic data are available from relatively few taxa with near-complete genome assemblies, and therefore, these simulations are applicable to a limited number of taxa. Still, in contrast to analyses that used all paralogs, when using only paralogs from a WGD, the WGD peak was consistently prominent across all ages, even when there was only a 5% gene retention rate following the WGD (Figs. 4-5 and 4-6). Syntenic data with node-averaged estimations could detect ancient WGDs even beyond \( Ks = 3.0 \). Also, analyses of the syntentic gene pairs could detect a distinct WGD component at \( Ks = 0.5 \) (Fig. 4-6), where the WGD peak is absorbed into the background duplication distribution in analyses of all paralogs. Although a number of empirical studies have detected peaks lower than \( Ks = 0.5 \), from ranges of \( Ks = 0.1 \) to \( Ks = 0.3 \), such as in *Zea mays* (Schlueter et al. 2004; Vanneste et al. 2014), *Glycine max* (Cui et al. 2006; Vanneste et al. 2014), and *Helianthus* (Barker et. al. 2008). This implies that background duplicate gene loss, especially of recent duplicates, is much faster than we can model with a simple stochastic birth and death process and that our
simulations may have retained too many genes from background duplications. For instance, Li et al. (2016) showed that putative single-copy orthologous groups tend to revert to a single-copy state rapidly following gene duplication across angiosperms. Nevertheless, our simulations still suggest that the availability of syntenic data can aid in the detection and timing of WGD peaks in $K_s$ plots. The use of $K_s$ plots built from syntenic data has helped resolve the absolute timing of notable WGD events such as two ancient WGDs shared by all Brassicaceae (e.g. Bowers et al. 2003), the papilionoid legume WGD (e.g. Schmutz et al. 2010), the ancestral eudicot triplication (Tang et al. 2008), a WGD predating angiosperms (Amborella genome project 2013), and at least two WGDs shared by most grasses (e.g. McKain et al. 2016). Syntenic data alone can be taken as evidence of a WGD (Kellis et al. 2004; Aury et al. 2006; Tang et al. 2008), and in these cases, the $K_s$ plot analyses may be viewed as corroborating evidence for an ancient WGD. However, large syntenic regions within a genome could also be caused by segmental duplications, which may be difficult to distinguish from a WGD.

In spite of our assessment of mixture model analyses of $K_s$ plots, they have played a major role in the characterization of many well-accepted ancient WGDs across plants (e.g. Schlueter et al. 2004; Cui et al. 2006; Rensing 2007; Barker 2008; Vanneste et al. 2013; Szövényi et al. 2015; Johnson et al. 2016). Thus, the question remains whether $K_s$ plot analyses may perform better on empirical rather than simulated data. In our simulations, we assumed random gene loss following a WGD. While most gene copies go extinct following a WGD, those that survive typically maintain balances of gene products uniquely associated with WGDs compared to other gene duplication mechanisms (Freeling 2009; Edger and Pires 2009; Conant et al. 2014). The biases in
gene retention following a WGD may contribute to evidence for ancient WGDs from $K_s$ plots, especially when a relatively small number of additional gene copies were retained from the WGD, such as *Physcomitrella patens* (Rensing et al. 2007) and *Nelumbo nucifera* (Ming et al. 2013). Additionally, transcriptomic data may be more appropriate for $K_s$ plot analyses than whole-genome data. Yang et al. (2015) estimated duplicate gene retention following WGD to be approximately 30% from ancient WGDs in the Caryophyllales across multiple transcriptome sequences. This estimate is much higher than retention rate estimates for many putative land plant WGDs based on genomic data (e.g. Tiley et al. 2016) with the exception of the most recent WGD events in the history of *Glycine max* (Schmutz et al. 2010) and *Populous trichocarpa* (Tuskan et al. 2006). Increased retention of duplicates from WGDs should be expected for transcriptomes though, if WGD duplicates are more frequently generally expressed than duplicates generated from other small-scale duplication events. It is unclear how the polyploidy mechanism affects duplicate gene retention though. Our simulation experiments and analyses here implicitly assumed autopolyploidy. Allopolyploidy, where the $K_s$ peak represents the speciation of parents rather than the actually hybridization and WGD event (Thomas et al. 2016), may bias duplicate gene retention towards a single parent species (Garsmeur et al. 2014), causing ancient WGDs from autopolyploidy to be more prominent in $K_s$ plots than allopolyploidy. Thus, our results on minimal gene retention rates may not be applicable to allopolyploid cases, but we expect general conclusions on mixture model analyses and the use of node-averaged $K_s$ to be applicable to allopolyploid $K_s$ plots as well.
Based on our results, mixture model analyses of $K_s$ plots should be considered, at most, a hypothesis-generating tool for ancient WGDs. Evidence of an ancient WGD from $K_s$ plot analyses should not be considered proof of an ancient WGD, nor should the absence of evidence of a WGD from a $K_s$ plot be considered proof of the absence of an ancient WGD. Multiple lines of evidence, ideally including syntenic evidence and phylogenetic tests for a WGD, should be utilized for identifying and characterizing ancient WGDs. When $K_s$ plot analyses perform well, they can help bracket the phylogenetic placement of ancient WGDs (Barker et al. 2008; Barker et al. 2009; Barker et al. 2016), although in some cases $K_s$ plots from multiple taxa produce conflicting signals regarding the phylogenetic placement of the ancient WGDs (e.g. Devos et al. 2016). A more comparative genomic and transcriptomic data becomes available, combined analyses of $K_s$ plots and gene tree reconciliation can improve the phylogenetic placement of WGDs (Barker et al. 2009; Jiao et al. 2011; Yang et al. 2015; Li et al. 2015; Barker et al. 2016) and probabilistic models of gene gain and loss (Rabier et al. 2014; Tiley et al. 2016) can provide much more rigorous statistical tests for ancient WGDs compared to mixture model analyses. However, $K_s$ plots provide the benefits of being computationally inexpensive, not requiring comparative genomic data, and not requiring ultrametric phylogeny estimates. Simple methodological choices such as using node-averaged estimates of $K_s$ and discerning consistency in mixture model results can help evolutionary biologists maximize these benefits of $K_s$ plots.

**Methods**

The goals of this study are to characterize and evaluate the performance of several approaches for detecting WGDs from $K_s$ plots under different evolutionary scenarios. We simulated gene family evolution, including WGDs, while varying the age
of the WGDs, the gene retention rate following WGDs, and the background rates of
gene duplication and loss. We used node-averaged and pairwise estimators to generate
the \( K_s \) distributions from the simulated gene families and applied two different mixture
models to estimate the number of WGDs and their age. We also performed analyses
using all paralogous gene pairs and only those resulting from WGDs.

**Simulating Gene Family Evolution**

Gene trees were simulated using GENPHYLODATA (Sjöstand et al. 2013). We
simulated gene families with WGDs within a single species by allowing gene trees to
evolve under a gene duplication and loss process alone within a species tree, but
interpreting speciation events as WGDs. Like a speciation event, a WGD (or at least an
autopolyploidy event) will result in a split in each gene lineage. We generated four sets
of gene trees with a single WGD at different ages in millions of years (50 mya, 100 mya,
300 mya, and 500 mya), which corresponds to \( K_s = 0.5, 1.0, 3.0, \) and 5.0. For each set,
the gene trees had a gene duplication rate (\( \lambda \)) and a loss rate (\( \mu \)) for all branches. We
incorporated variation in \( \lambda \) and \( \mu \) among gene trees by allowing \( \lambda \) and \( \mu \) to be
independently distributed as \( \beta(4, 2460) \) and \( \beta(4, 2053) \) respectively. The distributions of \( \lambda \)
and \( \mu \) have means of 0.00162 and 0.00194, which were estimates for land plants based
on a model of gene copy number evolution in Tiley et al. (2016), with assumed equal
variances of \( 1e10^{-6} \). Additionally, we allowed for variation in the number of genes at the
root of each gene tree, and thus the size of the gene families, by allowing the root age
to be exponentially distributed, such that the root age had a mean of 600 million years
ago. One thousand gene trees were simulated for each of the four WGD ages.

For each simulated gene tree, we simulated codon sequences using the
EVOLVER program within PAML v4.8a (Yang 2007). First, to introduce among branch
variation in substitution rates, branch lengths were relaxed using BRANCHRELAXER (Sjöstrand et al. 2013). This made rates consistent with the autocorrelated lognormal model of Rannala and Yang (2007) with mean of 1 and drift distributed as $\Gamma(1,1000)$, such that drift had a mean of 0.001 and variance of $1e10^-6$. Codon sequence data was simulated on the gene trees with relaxed branch lengths using a Goldman-Yang (GY94) model of codon evolution (Goldman and Yang 1994; Nielsen and Yang 1998) with equal equilibrium codon frequencies, a transition/transversion rate ratio ($\kappa$) of 2, a global $dN/dS$ ($\omega$) of 0.2, and an alignment length of 1000 codons. The per site evolutionary distance ($t$) for simulations was set to 0.01268182 so that $dS = 0.01$/million years. This $t$ was selected to ease translating from a WGD at a known age in the ultrametric species tree to a known $dS$. For example, for a given GY94 instantaneous rate matrix $Q$, the $dS$ can be solved for given some $t$, $\omega$, and $\kappa$. For all $q_{ij}$ in $Q$ such that $i = \{1,2,\ldots,61\}$ and $j = \{1,2,\ldots,61\}$ where $\pi$ is the set of 61 equilibrium codon frequencies, the number of synonymous ($S_d$) and nonsynonymous ($N_d$) substitutions per codon are $S_d = tp_S = \sum_{i\neq j, a_1\neq aa_j} \pi_i q_{ij} t$ and $N_d = tp_N = \sum_{i\neq j, a_1\neq aa_j} \pi_i q_{ij} t$, respectively. The number of synonymous and nonsynonymous sites per codon can be calculated as $S = 3p_S t$ and $N = 3p_N t$. For $p_S t$ and $p_N t$, $\omega$ is constrained to 1 instead of applying the MLE of $\omega$. Thus $p_S t = \frac{1}{C} \sum_{i\neq j, a_1\neq aa_j} \pi_i q_{ij} t$ and $p_N t = \frac{1}{C} \sum_{i\neq j, a_1\neq aa_j} \pi_i q_{ij} t$, where $C = \sum_{i\neq j} \pi_i q_{ij}$ is a scaling constant. Finally, $dS$ is calculated by $S_d/S$.

To assess the performance of mixture models under varying degrees of gene loss following WGDs, we pruned gene trees such that both gene copies following a WGD had probability of surviving in 0%, 1%, 10%, 25%, 50%, and 100% of the gene
trees. A 0% survival probability, or retention rate, means there is no evidence of the WGD (i.e. all new gene copies are immediately lost), and a 100% survival probability, or retention rate, indicates that there is no instantaneous gene loss following the WGD. We created 100 resampled data sets of $K_s$ by allowing random instantaneous loss on the 1000 gene trees to proceed 100 times. If a gene tree was randomly chosen to lose a gene from the WGD, then the left and right subtree had equal probability to be removed from the gene tree. The simulation process is graphically depicted in Figure 4-8.

**Estimating Synonymous Substitution Rates**

We estimated substitution model parameters using the GY94 model (Goldman and Yang 1994; Nielsen and Yang 1998), the same model used to simulate the nucleotide alignments, but using the empirical (i.e., observed) codon frequencies (F3x4). The GY94-F3x4 estimator was implemented using codeml within PAML v4.8a (Yang 2007). Node-averaged $K_s$ values were obtained by optimizing model parameters on the gene tree used to simulate the data using the simulated sequence data. For each internal node, from the tips to the root, node $K_s = (((\text{distance to left child} + \text{left child node } K_s) + (\text{distance to right child} + \text{right child node } K_s)) / 2$. The node-averaged $K_s$ was extracted using Perl scripts from newick trees with branch length as $K_s$. Pairwise $K_s$ estimates were obtained by optimizing the ML estimator parameters for each pair of sequences from a gene tree with PAML.

We also constructed $K_s$ plots only from nodes in gene trees that resulted from the WGD. For example, consider that following a WGD there is both an A and B subgenome. If a node is the most recent common ancestor of a gene from the A subgenome and a gene from the B subgenome, then that node must have been generated by a WGD (Object 4-1, Fig. S5). $K_s$ plots were then constructed using only
these node-averaged \( Ks \) values that were generated by the WGD event. These genes could represent sequences located on large syntenic blocks within a genome resulting from ancient WGDs, and using only nodes from WGDs allowed us to assess the value of \( Ks \) plot analyses using only gene pairs from syntenic regions for identifying WGDs.

**Discriminate Analyses and Fitting Mixture Models**

We fit univariate mixture models to simulated distributions of \( dS \) by expectation maximization using R (R Core Team 2015) with in-house source code (https://github.com/gtiley/Ks_plots) that uses the finite mixture expectation maximization algorithm implemented by Benaglia et al. (2009). For each \( Ks \) plot, we fit two models with \( k \) mixing components: 1) \( k \) normal components (e.g. McLachlan et al. 1999), as used in numerous previous studies (e.g. Schlueter et al. 2004; Cui et al. 2006; Barker et al. 2008) and 2) an exponential distribution with \( k-1 \) normal components, to better account for the contributions of background gene duplication and loss. Distributions of syntenic data (i.e. nodes resulting from WGDs) were only analyzed with a mixture of normal distributions, as the exponential distribution from background duplication and loss is no longer present. We used 100 random starts to find the optimal mixing components for each number of distributions. The number of components (\( k \)) was inferred using the \( \Delta BIC \). Due to the large number of simulations performed here, it was not possible to use nonparametric bootstrapping to develop an empirical null distribution of likelihood ratio test statistics (McLachlan, 1987). Although \( \Delta BIC \) does not provide a formal hypothesis test, it can guide model selection and is often used in \( Ks \) plot analyses (e.g. Barker et al. 2008). We assumed that the \( \Delta BIC \) between nested models provides an approximation of Bayes factors, and we used \( \Delta BIC < 3.2 \) as a stopping criterion (Kass and Raftery 1985). Functions for automating the selection of the optimal
Comparisons with Empirical Data

To better assess the generality of our mixture model results on simulated data, we reanalyzed previous Ks plots for previously characterized ancient WGDs that occurred before the diversification of Actinidiaceae (Shi et al. 2010), most Asteraceae (Barker et al. 2016), and the common ancestor of Asteraceae and Calyceraceae (Barker et al. 2016). For the Actinidiaceae WGD, we reanalyzed both node-averaged and pairwise Ks from Actinidia chinensis and Actinidia deliciosa EST datasets (Shi et al. 2010). We used node-averaged and pairwise Ks from an Artemisia annua transcriptome for the Asteraceae-specific ancient WGD as well as transcriptomes from Barnadesia spinosa and Acicarpha spathulata for the shared Asteraceae and Calyceraceae WGD (Barker et al. 2016). We estimated the optimal number of mixing components and the mean age in Ks of each component using the bic.test.wgd R function for the exponential + normal and normal mixture models using 100 random starts with a maximum of five components.
Figure 4-1. Distributions of the optimal numbers of components across WGDs of different ages for both known and estimated values of $K_s$. Models applied are a) a single exponential and $k-1$ normal components and b) a mixture of $k$ normal distributions. The color scale refers to the percentage of replicates with an optimal $k$ of 1 through 5. The gene retention rate is the proportion of gene trees retaining at least one node whose children are products of a WGD.
Figure 4-2. Distributions of the mean $K_s$ of gene pairs comprising the components closest to the true age of the WGD at $K_s = 0.5$ and $K_s = 1.0$, when the number of components for each mixture model is optimized by ML. Horizontal black lines represent the true age of a WGD. Panels a-d represent WGDs at $K_s = 0.5$, and panels e-h are WGD at $K_s = 1.0$. Results for known simulated node-averaged $K_s$ when a WGD is at $K_s = 0.5$ are shown in panels a and b for the exponential + normal and normal mixture model respectively. Results for the WGD at $K_s = 0.5$ for estimated $K_s$ are in panels c and d. The distributions of means of components closest to the true WGD age when the WGD is at $K_s = 1.0$ is given in panels e and f for known simulated $K_s$, and panels g and h for estimated $K_s$. 
Figure 4-3. Distributions of $K_s$ for node-averaged and pairwise estimation. A WGD has an age of a) $K_s = 0.5$, b) $K_s = 1.0$, c) $K_s = 3.0$, and d) $K_s = 5.0$ where all gene families have variable $\lambda$ and $\mu$. Smoothing splines are kernel density estimates. Grey lines are known $K_s$ values while blue lines are MLEs. Squares represent the mean when $k = 2$ for the exponential + normal model while triangles represent the normal mixture model. Horizontal lines on the squares and triangles are the 95% confidence intervals.
Figure 4-4. Distributions of the optimal numbers of components for normal mixture models fit to syntenic node-averaged $K_s$ data. Retention rates of 0.05, 0.1, 0.5, and 1 for WGD ages $K_s = 0.5, 1.0, 3.0$, and 5.0 are shown for both known and estimated values of $K_s$ with among gene family variation in $\lambda$ and $\mu$. The color scale refers to the percentage of replicates with an optimal $k$ of 1 through 5. The gene retention rate is the proportion of gene trees retaining at least one node whose children are products of a WGD.
Figure 4-5. Distributions of the means of components closest to the true age of the WGD at $K_s = 0.5$, 1.0, 3.0, and 5.0, when the number of components for a normal mixture model is optimized by ML. Distributions are shown for known and estimated node-averaged $K_s$ of syntenic data. Horizontal black lines represent the true age of a WGD.
Figure 4-6. Distributions of $K_s$ for node-averaged a pairwise estimation when a WGD has an age of $K_s = 0.5$, $1.0$, $3.0$, and $5.0$ with among gene family variation in $\lambda$ and $\mu$. Smoothing splines are kernel density estimates to aid visualization. Grey lines are known $K_s$ values while blue lines are MLEs.
Figure 4-7. Ks plots for five previously published transcriptome sequences or EST datasets. Node-averaged ML estimates are shown on the left while pairwise ML estimates are shown on the right. Plots are truncated at $Ks = 2$ and arbitrarily on the y-axis to ease visualization. Blue asterisks indicated a component mean that was identified in both the exponential + normal and normal mixture model analyses. Complete mixture model results are given in Table S4 (Object 4-1).
Figure 4-8. Stylized workflow for simulating gene trees with a known WGD under a known retention rate. Gene families with an explicit WGD are simulated similar to gene trees evolving within a species tree. The branch length of the root of the species tree is drawn from a distribution so not all gene families are single copy before the WGD. The branch lengths following the WGD either are 50, 100, 300, or 500 mya, which correspond to $K_s$ of 0.5, 1.0, 3.0, and 5.0 respectively. Gene trees then evolve under variable rates of gene duplication and loss, followed by relaxation of branches to account for substitution rate variation. Sequence data was then simulated on these relaxed gene trees. We implemented a non-standard data resampling procedure here that allowed for random gene loss. Individual gene trees were not resampled, but the probability that both lineages following a WGD survived was resampled. One hundred $K_s$ plots were then generated from the one hundred resampled data sets of one thousand trees.
Table 4-1. Detection and age of WGD peaks for distributed duplication and loss rates with node Ks. The proportion of bootstrap replicates with an optimal number of 2 components (k) and the mean Ks when k is constrained to 2 is displayed as: Exponential + Normal model with simulated node Ks, Exponential + Normal model with estimated node Ks / Normal mixture with simulated node Ks, Normal mixture with estimated node Ks. 95% confidence intervals are given for WGD peak means.

<table>
<thead>
<tr>
<th>WGD age</th>
<th>Retention</th>
<th>Replicates with optimal k=2</th>
<th>Mean for k=2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0</td>
<td>0.01, 0.51 / 0, 0</td>
<td>3.16 ± 0.2, 3.12 ± 0.2 / 2.63 ± 0.21, 2.63 ± 0.21</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.04, 0.72 / 0, 0</td>
<td>3.18 ± 0.2, 3.17 ± 0.2 / 2.65 ± 0.21, 2.67 ± 0.21</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.18, 0.79 / 0, 0</td>
<td>3.22 ± 0.2, 3.28 ± 0.19 / 2.67 ± 0.21, 2.73 ± 0.21</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.24, 0.86 / 0, 0</td>
<td>3.26 ± 0.19, 3.33 ± 0.19 / 2.68 ± 0.21, 2.74 ± 0.2</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.36, 0.02 / 0, 0</td>
<td>3.13 ± 0.18, 2.4 ± 0.15 / 2.58 ± 0.2, 2.24 ± 0.19</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.92, 0 / 0, 0</td>
<td>0.46 ± 0.07, 0.49 ± 0.08 / 1.21 ± 0.16, 1.25 ± 0.16</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0.25, 0.1 / 0, 0</td>
<td>3.31 ± 0.18, 3.39 ± 0.19 / 2.70 ± 0.2, 2.76 ± 0.2</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.35, 0.13 / 0, 0</td>
<td>3.42 ± 0.18, 3.50 ± 0.18 / 2.80 ± 0.2, 2.82 ± 0.2</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.72, 0.34 / 0, 0</td>
<td>3.61 ± 0.16, 3.74 ± 0.16 / 2.99 ± 0.19, 3.06 ± 0.19</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.8, 0.41 / 0, 0</td>
<td>3.72 ± 0.16, 3.91 ± 0.15 / 3.07 ± 0.19, 3.23 ± 0.18</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0, 0 / 0, 0</td>
<td>0.99 ± 0.09, 1.21 ± 0.10 / 1.73 ± 0.16, 1.93 ± 0.16</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0, 0 / 0, 0</td>
<td>0.91 ± 0.09, 1.03 ± 0.10 / 1.68 ± 0.16, 1.82 ± 0.16</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0.23, 0.06 / 0, 0</td>
<td>5.13 ± 0.18, 4.91 ± 0.27 / 4.31 ± 0.25, 4.91 ± 0.27</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.29, 0.04 / 0, 0</td>
<td>5.06 ± 0.18, 4.97 ± 0.27 / 4.28 ± 0.25, 4.97 ± 0.27</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.5, 0.04 / 0, 0</td>
<td>4.97 ± 0.19, 4.98 ± 0.27 / 4.28 ± 0.25, 4.98 ± 0.27</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.44, 0.15 / 0, 0</td>
<td>4.95 ± 0.19, 5.0 ± 0.27 / 4.30 ± 0.25, 5.0 ± 0.27</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.02, 0.03 / 0, 0</td>
<td>4.87 ± 0.2, 5.08 ± 0.27 / 4.38 ± 0.25, 5.09 ± 0.27</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0, 0 / 0, 0</td>
<td>4.85 ± 0.2, 5.11 ± 0.27 / 4.43 ± 0.26, 5.11 ± 0.27</td>
</tr>
</tbody>
</table>
CHAPTER 5
CONSERVED RATES OF MOLECULAR EVOLUTION AMONG DUPLICATE GENE COPIES AFTER PALEOPOLYPLOIDY AND A COMPLEMENTARY HYPOTHESIS TO GENE DOSAGE BALANCE

Introduction

Gene duplication is the most important mechanism for the evolution of new genes and gene families (Ohno 1970). Duplicate genes are often the basis of traits associated with major groups such as feathers in avian reptiles (Wu et al. 2015), floral morphological diversity (Pelaz et al. 2000; Howarth and Donoghue 2006), and complexity of transcriptional regulation in eukaryotes (de Mendoza et al. 2013). Gene duplication has certainly contributed to variation in gene copy number and diversity at the coding sequence level between species, but also patterns of tissue-specific gene expression (Miyata et al. 1994; Gu et al. 2004; Roulin et al. 2012). One reason for tissue-specific expression across duplicate genes is when upstream cis-regulatory regions are duplicated along with the protein-coding gene sequence and diverge (e.g. Papp et al. 2003).

Differentiating between divergence of protein-coding gene sequence and regulatory sequence is important, as both mechanisms can be associated with different types of gene duplication. Specifically, we expect divergence at the protein-coding sequence level to be more important for gene retention following small-scale duplication (SSD) events, as some regulatory motifs may be absent following duplication (Ohno 1970). However, divergence at the cis-regulatory sequence level, may play a larger role in gene retention following whole-genome duplication (WGD) since all of the regulatory components would have duplicated as well (Ohno 1970). Understanding gene retention following SSD and WGD in plants is especially of interest to evolutionary
biologists, since plants have the highest observed rates of gene duplication and loss among eukaryotes (Barker et al. 2012) and all extant angiosperms have experienced one or more rounds of WGD (Jiao et al. 2011). However, plants show heterogeneity in the fraction of genes retained from independent WGDs (Tiley et al. 2016) and it is not clear what mechanisms are driving differential retention of duplicate genes following WGDs in plants.

Genes retained from WGDs and SSDs, which includes tandem duplication by unequal crossing over and dispersed duplication by retrotransposition, have previously been characterized with different rates of molecular evolution. For example, genes likely derived from ancient WGDs in the Brassicales have a lower nonsynonymous to synonymous substitution rate ratio ($dN/dS$) than genes derived from tandem duplication (Hofberger et al 2015) as well as duplicate genes putatively not of WGD origin (Yang and Gaut 2011). Duplicate genes that are physically linked on the same chromosome exhibit higher variation in rates of molecular evolution and a higher $dN/dS$ compared to duplicate genes that reside on separate chromosomes (Wang 2013). Understanding these differences in $dN/dS$ are relevant to the survival of duplicate genes, as larger differences in $dN/dS$ between paralogs are associated with increased divergence in expression in *Arabidopsis* and *rice* (Wang 2013). Gene expression ultimately affects the survival and evolution of duplicate genes (e.g. Casneuf et al 2006; Birchler and Veitia 2012), and at least in soybean, differences in $dN/dS$ can be a reasonable predictor of differential expression across tissues, where greater $dN/dS$ corresponds to greater differences in expression (Roulin et al 2012). We are interested if $dN/dS$ can be used to explain differential gene retention following WGD and SSDs among species.
Prevailing explanations for duplicate gene retention following WGDs is that duplicates related through the same regulatory networks are retained to maintain stoichiometric balances of protein products, frequently referred to a gene dosage balance (Papp et al. 2003; Birchler and Vieta 2012). Recent studies in cotton and soybean have shown evidence for biased retention of genes from WGDs that are interconnected through the same regulatory networks (Gallagher et al. 2016; Coate et al. 2016). Initial analyses of complete plant genomes inferred that transcription factors are over-represented among genes retained from WGDs (Cannon et al. 2004; Seoighe and Gehring 2004; Maere et al. 2005) and recent analyses of gene ontological data across species also support these apparent biological constraints on duplicate retention following WGD (e.g. Yang et al. 2015; Rody et al. 2017; Mandáková et al. 2017). Although, transcription factor retention may also be a feature of increasing gene duplication age (Rody et al. 2017). Adaptation has also been invoked to explain gene retention following WGDs (e.g. Guo et al. 2013; Geiser et al. 2015; Hofberger et al. 2015; Olsen et al. 2016), but population genetic processes have been less explored to explain patterns of biased gene retention across ancient WGDs (but see McGrath et al. 2014; Douglas et al. 2015).

Population genetic processes to be important for duplicate gene retention because of positional effects and chromatin structure. Following WGD, genomes experience a number of architectural rearrangements that facilitate rapid gene loss, such as recombination between homeologous chromosomes, changes in DNA methylation, and activation of transposable elements (e.g. Wendel et al. 2016). Despite these complex biological processes generating variation in genome architecture, plant
genomes share many common features, including gene-rich euchromatic regions with elevated recombination rates (e.g. Bowers et al. 2005). This non-random distribution of genes may be important for duplicate gene retention, since high recombination increases the efficacy of selection, which can facilitate selection against even slightly deleterious rearrangements or mutations (e.g. Campos et al. 2014). For example, recombination rates are higher in regions of synteny between rice and *Sorghum* compared to non-syntenic regions (Bowers et al. 2005), and large proportions of genetic maps in legumes are syntenic between species (Hougaard et al. 2008). Additionally, the distributions of recombination rate are associated with nucleotide diversity in grass and legume genomes (e.g. Flowers et al. 2012; Paape et al. 2012), and there are genome-wide associations between increased recombination rate and more effective purifying selection across angiosperms (Tiley and Burleigh 2015).

Hence, we attempted to explain differential gene retention from WGDs and SSDs in plants at the protein-coding sequence level through integrated analysis of rates of molecular evolution and genome structure. We reconciled broad evolutionary differences between WGD and SSD duplicates across multiple plant species by addressing three evolutionary questions: 1) Do duplicate genes that arise by WGDs evolve, on average, at a different rate of molecular evolution, $dN/dS$, compared to SSDs? 2) Do different gene duplication mechanisms exhibit different levels of asymmetry in $dN/dS$ following duplication? 3) Does genome architecture explain differences in asymmetry of $dN/dS$ across gene duplication mechanisms?
Results

Rates of Molecular Evolution Following Gene Duplications in Grasses

For the hypothesis testing, we identified 123 gene trees that included at least one WGD, cis-SSD, and trans-SSD duplication node. These gene trees were fit to one of four models using nested likelihood ratio tests (LRTs). The 1-rate model assumes all gene tree branches have the same $dN/dS$, the 2-rate model allows branches following speciation and duplication nodes to have different $dN/dS$, the 3-rate model allows branches following speciation, WGD, and non-WGD to have a different $dN/dS$, and the 4-rate model additionally allows branches following the two types of non-WGD duplications (i.e., cis-SSD and trans-SSD) to have a different $dN/dS$. After correcting for multiple tests, 40 gene trees fit the 1-rate model, 32 fit the 2-rate model, 13 fit the 3-rate model, and 35 fit the 4-rate model. We excluded three gene trees because they had insufficient data to estimate $dN/dS$ accurately for some branches.

Branches descending from speciation nodes display statistically identical distributions of $dN/dS$ across 1-rate, 2-rate, 3-rate, and 4-rate gene trees (Fig. 5-1; one-way ANOVA; $p = 0.31$). This implies duplication events in the 1-rate gene trees evolve with $dN/dS$ consistent with speciation nodes across all gene trees, and that rate differences among 2-rate, 3-rate, and 4-rate gene trees may reveal interesting features of gene duplication for each rate category.

In the 2-rate model gene trees, branches following duplication nodes typically have a higher $dN/dS$ (median = 0.342; Fig. 5-1) than branches descending from speciation nodes (median = 0.2; Fig. 5-1) (one-way ANOVA; $p = 0.002$; Object 5-1, Table S1). Overall, 27 out of 32 2-rate gene trees showed higher $dN/dS$ among duplication branches compared to speciation branches. However, there are only minor
differences in the overall distributions of $dN/dS$ between branches following different types of gene duplications for the 3-rate and 4-rate models (Fig. 5-1; Object 5-1, Table S1). Thus, we derived simple tests based on the qualitative differences between the types of duplication branches on each gene tree by ranking them into three classes: 1) the highest $dN/dS$, 2) the intermediate $dN/dS$, and 3) the lowest $dN/dS$. This allowed us to perform simple binomial tests for over- or under-representation in one of the three classes (or two classes for the 3-rate gene trees), testing the null hypothesis that the different types of duplication occur in each of the three $dN/dS$ classes with equal frequency (Object 5-1, Table S2). For the 13 gene trees that fit the 3-rate model, we did not reject the null hypotheses that branches following WGD or non-WGD duplications are equally represented in the low and high $dN/dS$ rate classes (Object 5-1, Table S2). When observing only the 35 gene trees that fit the 4-rate $dN/dS$ model, trans-SSDs are significantly over-represented in the lowest $dN/dS$ rate class (binomial test; $p = 0.03$; Fig. 5-1) and are under-represented at the highest $dN/dS$ rate class (binomial test; $p = 0.006$). We did not reject the null hypothesis that either WGD or cis-dispersed duplicates occur at expected ratios in each of the three qualitative $dN/dS$ classes.

Our LRTs to compare $dN/dS$ across branches following different types of gene duplications do not inform us of whether there are rate differences between each pair of duplicate genes. Therefore, we tested for asymmetry in $dN/dS$ for each duplicate gene pair in the pool of significant hypothesis tests. For the 2-rate gene trees, we tested 294 duplication nodes for asymmetry in $dN/dS$. We discarded 10 duplication nodes because we could not obtain credible estimates of $dN/dS$ from the data (i.e., $dN/dS > 5$ for the pair or one copy). Twenty-four duplication nodes rejected the null hypothesis of equal
rates between the two branches following the duplication, and 260 did not (Fig. 5-2). Based on this, we estimated the expected probability of asymmetry in $dN/dS$ for a pair of duplicate genes or branches to be 0.085 (Object 5-1, Table S3). However, the average proportion of asymmetry among all 1-rate, 2-rate, 3-rate, and 4-rate gene trees is 0.085, indicating that asymmetry among 2-rate gene trees may be a reasonable estimator of expected asymmetry for duplication nodes.

We tested for asymmetry in $dN/dS$ in the branches following 19 WGD nodes and 71 non-WGD nodes in the 3-rate gene trees (Fig. 5-2). Only two of the 19 WGD duplication nodes rejected the null hypothesis of equal rates among the duplicates, which is consistent with the expected levels of asymmetry based on the 2-rate gene trees (binomial test; $p = 0.674$). Eleven of the 71 non-WGD nodes rejected the null hypothesis of equal rates asymmetry though. This is a higher rate of asymmetry than the WGD nodes, but again, it is consistent with the expected levels of asymmetry based on the 2-rate gene trees (binomial test; $p = 0.051$).

In the 35 4-rate gene trees, none of the 66 WGD nodes showed asymmetry among duplication branches (Fig. 5-2), and the under-representation of asymmetry was significant by a binomial test ($p = 0.006$). In contrast, 18 out of 172 the cis-duplication nodes rejected the null hypothesis of equal $dN/dS$ among duplicates branches, which is consistent with expected ratio from the 2-rate trees (binomial test; $p = 0.602$). For the trans-SSD duplication nodes, we inferred a higher than expected ratio of asymmetry (15 out of 84 branch pairs; binomial test; $p = 0.005$).

**Rates of Molecular Evolution among Gene Duplication Mechanisms in Legumes**

In the legumes, we identified 244 gene trees appropriate for our hypothesis testing: 101 followed the single-rate model, 45 followed the 2-rate model, 38 followed
the 3-rate model, and 58 followed the 4-rate model. Estimates of $dN/dS$ for the 1-rate gene trees as well as speciation branches for all other models are identically distributed (Fig. 4-3; Object 4-1, Table S1), consistent with the grass results and implying that duplication branches among 1-rate gene trees evolve similarly to speciation branches in 2-rate, 3-rate, and 4-rate gene trees. Among the 2-rate gene trees, duplication branches show an elevated $dN/dS$ (median = 0.29) compared to speciation branches (median = 0.16) (one-way ANOVA; $p=0.007$), with 35 out of 15 gene trees showing elevated $dN/dS$ across duplication branches. Similar to grasses, there were no detectable differences among duplication distributions and speciation distributions among 3-rate and 4-rate gene trees (Object 5-1, Table S1) and we ranked duplication branches from 3-rate and 4-rate gene trees into $dN/dS$ rate classes (Object 5-1, Table S2). In the 3-rate gene trees, non-WGD duplicates frequently had a higher $dN/dS$ than WGD duplicates (23 out of 38), but this difference was not significant (binomial test; $p = 0.128$). Among the 58 gene trees that fit the 4-rate $dN/dS$ model, trans-SSDs were over-represented in the lowest $dN/dS$ class (binomial test; $p = 0.024$), and cis-dispersed duplicates were enriched for the highest $dN/dS$ class (binomial test; $p < 0.001$). WGD duplicates were enriched in the intermediate $dN/dS$ class (binomial test; $p < 0.001$) and under-represented from the highest $dN/dS$ class (binomial test; $p < 0.001$).

For the 2-rate gene trees, 32 out of 457 (7%) duplication nodes showed significant asymmetry based on the LRTs (Fig. 5-4; Object 5-1, Table S3). Even when considering all 1-rate, 2-rate, 3-rate, and 4-rate gene trees, we find asymmetry in 197 out of 2274 (7.85%) of duplication nodes, which is very close to the estimate from 2-rate gene trees alone. Among 3-rate gene trees, we observed asymmetry in 16 out of 106
WGD duplication nodes (binomial test; p = 0.003) and 21 out of 263 non-WGD nodes (binomial test; p = 0.544), implying high asymmetry than expected from the 2-rate gene trees in \( dN/dS \) following WGDs but not non-WGD duplications. However, WGD duplications showed asymmetry at an expected level for the 4-rate gene trees (10 out of 142 pairs; binomial test; p = 1.0). Also in the 4-rate gene trees, both the cis-SSDs (33 out of 319 gene pairs; binomial test; p = 0.027) and trans-SSDs (15 out of 109 gene pairs; binomial test; p = 0.013) have higher levels of asymmetry than expected given the null hypothesis from the 2-rate genes (Fig. 4-4; Object 5-1, Table S3).

**Biased Retention of Genes Following WGD**

Across grass 4-rate gene trees, the GO function protein binding (GO:0005515) was over-represented among symmetric WGD nodes (Table 5-1). The evidence was weaker in our grass 3-rate gene trees and legume analyses, but we consistently observed a higher proportion of symmetric \( dN/dS \) in protein binding gene trees among WGD nodes compared to non-WGD nodes (Table 5-1; Object 5-1, Table S7). There was no evidence for enrichment of any other GO functions across the gene trees tested in our 3-rate and 4-rate analyses (Table 5-1; Object 5-1, Table S7).

We then attempted to associate our \( dN/dS \) results and the enrichment of protein binding functions with genome architecture. We used a least common ancestor mapping approach to identify gene tree nodes as evolutionarily syntenic or non-syntenic (see Methods: *Associating synteny across species with rates of molecular evolution*), and performed \( \chi^2 \) tests for over-representation of symmetric nodes in regions of evolutionarily conserved synteny across duplication node classes. For grass and legume 4-rate gene trees, we found more symmetric WGD duplication nodes on
evolutionarily conserved syntenic blocks than expected, but cis-SSD nodes and trans-SSD nodes occurred at expected frequencies (Table 5-2). Legume WGD nodes in 3-rate gene trees were consistent with the 4-rate gene tree results, showing strong deviations from expectations under the 2-rate gene trees due to increased symmetric $dN/dS$ in evolutionarily conserved regions, but grass WGD nodes from 3-rate gene trees were not (Table 5-2). There was still an extreme $X^2$ statistic for grass 3-rate WGD nodes, but this was simply because of small sample size (Table 5-2). However, we analyzed a larger data set that does not attempt to differentiate between cis-SSDs and trans-SSDs (Object 5-1), and the grass 3-rate WGD nodes became consistent with the rest of our results and show an enrichment of symmetric WGD nodes in evolutionarily conserved syntenic regions (Object 5-1, Table S8). We found that approximately 30% of WGD nodes with symmetric $dN/dS$ occurred in evolutionarily conserved syntenic regions, while non-WGD nodes occurred in syntenic regions at expected levels of 7%-8.5% (Table 5-2; Object 5-1, Table S8). Although the majority of gene duplication nodes in this study lie outside of evolutionarily conserved syntenic blocks, the majority of duplications maintained in regions of between-species synteny are putative WGD duplicates (Figure 5-5; Table 5-2).

To test if the over-representation of symmetric WGD nodes in syntenic regions was due to protein binding alone, we reran the Fisher exact test with nodes corresponding to protein binding function removed. Twelve out of the 20 symmetric and syntenic duplication nodes among WGD duplication events in 4-rate grass gene trees corresponded to the protein binding GO function. Excluding these twelve nodes from GO enrichment tests reduced, but did not completely erode evidence for the over-
representation of the protein binding function among symmetric WGD duplication nodes (Fisher exact test; p = 0.032). In legumes, there were 45 gene duplication nodes with symmetric \( dN/dS \) that mapped to evolutionarily conserved syntenic regions among 4-rate WGD nodes, including thirteen contributing to protein binding. While the protein binding function was not strongly enriched among WGD duplication nodes in legumes as it was in grasses, removing these symmetric and syntenic nodes further diminished their retention (Fisher exact test; p = 1). Thus, there is likely an association between synteny and protein binding enrichment following WGDs, but evolutionarily conserved synteny does not completely explain high retention of protein binding genes among WGD nodes.

We found no evidence that asymmetry in rates of molecular evolution, and thus detectable patterns of retention in previous analyses, are dependent on the age of the duplication. Logistic regression model results on node age and asymmetry suggests that the log of odds of asymmetry is only differs from zero in one scenario (Object 5-1. Fig. S1 and Table S4). In trans-SSDs in legume 4-rate gene trees, we found a slightly positive relationship between asymmetry and the age of the duplication nodes (Object 5-1, Fig. S1 and Table S4); however, we did not observe a negative relationship between asymmetry and duplication age under any condition.

**Discussion**

Previous studies have identified the genes of putative WGD origin evolve, on average, at a lower \( dN/dS \) compared to non-WGD duplicates (Yang and Gaut 2011; Wang 2013; Hofberger et al. 2015). Our results partially agree with these findings in two multi-species contexts. A simple observation of the distribution of \( dN/dS \) for WGD duplicates and non-WGD duplicates in our 3-rate gene trees implies a generally lower
$dN/dS$ for WGD duplicates (Fig. 5-1; Fig. 5-3). However, our nonparametric testing procedure, which we argue is more appropriate for understanding relative rates of $dN/dS$ across gene families, implies equal levels of increased or decreased $dN/dS$ between WGD and non-WGD duplicates, except in our expanded grass dataset where $dN/dS$ among WGD nodes are consistently lower than non-WGD nodes (Object 5-1). Surprisingly, analyses of 4-rate gene trees in grasses and legumes revealed consistent patterns of decreased $dN/dS$ among trans-SSDs (Fig. 5-1; Fig. 5-3). Thus, we should not expect WGD duplicates to necessarily be associated with lower $dN/dS$ compared to other gene duplication mechanisms.

Our analyses of asymmetry in $dN/dS$ following individual duplication events suggested that the differences in global $dN/dS$ are explained, in part, by differences in fates of duplicate genes and genome architecture. The 4-rate gene trees in grasses and legumes revealed increased levels of asymmetry in $dN/dS$ between gene tree branches following cis-SSDs and trans-SSDs compared to WGD nodes (Fig. 5-2; Fig. 5-4). The 3-rate gene trees in grasses also support increased asymmetry among non-WGD nodes compared to WGD nodes (Fig. 5-2; Object 5-1); however, 3-rate legume gene trees imply higher asymmetry among WGD nodes or are inconclusive in the expanded dataset (Fig. 5-4; Object 5-1). The conflicting results between 4-rate and 3-rate gene trees in grasses may be due to the more recent soybean-specific WGD, but more importantly these results imply that the 3-rate and 4-rate gene trees may also reflect biological differences where WGD duplicates differentiate at the coding-sequence level more frequently than SSDs.
The pattern of increased asymmetry in $dN/dS$ in the 3-rate non-WGDs and 4-rate cis-SSDs and trans-SSDs may be associated with the fates of duplicate genes, such as neofunctionalization at the coding-sequence level. We found the highest proportion of neofunctionalization, in context of Ohno (1970) where one duplicate copy has $dN/dS > 1$, among asymmetrically evolving duplicates in the 4-rate cis-SSDs in both grasses and legumes (Fig. 5-2; Fig. 5-4; Object 5-1, Table S5). The majority of neofunctionalization cases in non-WGDs consist of leucine-rich repeat resistance genes while kinases and regulatory proteins are found among WGD neofunctionalization cases (Object 5-1, Table S6), which may contribute to previously observed biases in duplicate gene retention following WGD (e.g. Cannon et al. 2004; Seoighe and Gehring 2004; Maere et al. 2005; Yang et al. 2015; Rody et al. 2017; Mandáková et al. 2017). However, the total numbers of neofunctionalization cases among duplication nodes are very low (Fig. 5-2; Fig. 5-4; Object 5-1, Table S5), implying that relaxed purifying selection following duplication is most common, likely followed by pseudogenization or subfunctionalization.

Among WGD duplication nodes, we more frequently observed symmetric $dN/dS$, which has correlated with equal levels of gene expression in previous studies (Wang 2013; Roulin et al. 2012). Thus, WGD duplicates may be for frequently conserved at the protein-coding sequence level than SSDs, and subfunctionalization at the expression level (Duarte et al. 2006; Hughes et al. 2014; Pophaly and Tellier 2015), such that one gene copy following WGD can show dominant expression levels in specific tissues, may be more important for their survival. However, the globally increased $dN/dS$ among WGD duplicates with respect to trans-SSDs in grass and legume 4-rate gene trees
implies some degree of relaxed purifying selection in both duplicate gene copies following WGD, which would be consistent with the duplication-degeneration-complementation model of subfunctionalization at the coding sequence level (Force et al. 1999). It is difficult to attribute differences in $dN/dS$ between WGD duplicated and SSDs to any one fate determining mechanism, but our results support notable differences in rates of molecular evolution between WGD duplicates and SSDs, which are likely tied to different propensities for fates of duplicate genes.

In addition to biases in rates of molecular evolution between WGDs and SSDs, there are potential functional biases too. Our results agreed with many previous studies on the retention of genes annotated as protein binding (GO:0005515), which likely play some role in the regulation of transcription following WGDs (e.g. Maere et al. 2005; Mandáková et al. 2017; Table 5-1). It should be noted though that the protein binding GO annotations are highly abundant in our data (Table 5-1; Object 5-1, Table S7) as well as other genomic studies. For example, the GO function protein binding (GO:0005515) has been over-represented following ancient WGDs in vertebrates too (Pérez-Bercoff et al. 2010; Montoya-Burgos 2011), suggesting that some patterns of gene retention following WGD should be expected and may lack any adaptive significance for any single plant WGD. Thus, biased patterns of gene retention at the ontological level may not be due to gene balance and selection acting on polyploidy phenotypes, but greatly affected by other forces such as genome architecture.

Effects of genome architecture and recombination would explain, in part, the observed patterns of molecular evolution and gene retention. For instance, some WGD duplicates could be preserved because they exist in an evolutionarily conserved
syntenic block. A consistently unexpected 30% of symmetrically evolving WGD duplication nodes mapped to genomic regions where synteny was present between species (Fig. 5-5; Table 5-2; Object 5-1, Table S8). The biased gene retention in evolutionarily conserved regions was not observed for non-WGD duplicates in either the 3-rate or 4-rate gene trees for grasses or legumes (Fig. 5-5; Table 5-2; Object 5-1, Table S8). Additionally, there is evidence that recombination occurs more frequently in regions that are syntenic between species (e.g. Bowers et al. 2005; Hougaard et al. 2008), which also happens to be gene-rich euchromatin. Thus, genes in these evolutionarily conserved syntenic regions should be subject to effective purifying selection or at least sufficient levels of recombination to prevent fixation of deleterious mutations through hitchhiking (Muller 1964; Hill and Robertson 1966) or purging of beneficial mutations through background selection (Charlesworth et al. 1993). It is possible that a larger proportion of non-WGD duplicates, including cis-SSDs and trans-SSDs observed in this study reside in low-recombination regions, which could lead to increased \( dN/dS \) and increased asymmetry relative to WGD duplicates (Fig. 5-2; Fig. 5-4; Table 5-2; Object 5-1, Table S8).

One factor complicating our comparison of WGD duplicates and SSDs is if \( dN/dS \) was dependent on the age of the duplication. Recent comparisons of duplication in mouse and rat have shown elevated \( dN/dS \) and asymmetry among young duplicate genes (Pegueroles et al. 2013). However, we find no evidence that our results were affected by duplicate gene age. None of our logistic regression analyses show increasing asymmetry with decreasing duplicate node age (Object 5-1, Fig. S1 and Table S4). Additionally, previous analyses of the Arabidopsis alpha and beta WGDs
found comparable distributions of $dN/dS$ between ancient WGDs with an approximate distance of 35 million years between them (Yang and Gaut 2011). Therefore, we have evidence that the differences among rates, fates, and gene retention between WGD and non-WGD duplications here reflect biological differences and not age disparities between ancient WGDs and more recent SSDs.

Altogether, our approach to hypothesis testing with gene trees supports different patterns in rates of molecular evolution and fates of duplicate genes among WGD duplicates and SSDs. Genome architecture may play a larger role than appreciated in the retention of duplicate genes following WGD. Many WGD duplicates are likely retained because they reside in gene-rich regions of high recombination while many non-WGD duplicates may not be subjected to nearly as efficacious purifying selection. Duplicate gene evolution following WGD is a complex process, and integrated analysis of sequences, expression, and genetic maps is needed to further our understanding on the biological importance of ancient WGDs to extant plant diversity. Additionally, non-WGDs are deserving of attention too, as our study reveals higher potential for neofunctionalization at the coding sequence level with respect to WGD duplicates, especially in important gene families such as plant resistance genes. There is no other study in plants, that we are aware of, explicitly testing differences in symmetry in rates of molecular evolution between gene duplication mechanisms, so it is difficult to determine the generality of these results. However, our results are statistically rigorous and the consistent results in grasses and legumes suggest these patterns may be common for angiosperms.
Methods

Data, Alignments, and Gene Trees

We downloaded annotated genes and their respective genome features from Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html#; Goodstein et al. 2012). For the grasses, we selected the genomes of *Brachipodium distachyon* (version 3.0; Vogel et al. 2010), *Oryza sativa* (version 7.0; Ouyang et al. 2007), and *Sorghum bicolor* (version 2.1; Paterson et al. 2009), because they are well assembled and include the necessary comparative syntenic data. We did not include *Zea mays* because it contains a more recent WGD not shared with the other grass taxa (Gaut and Doebley 1997). For the legumes, we selected the genomes of *Glycine max* (version Wm82.a2.v1; Schmutz et al. 2010), *Phaseolus vulgaris* (version 2.1, Schmutz et al. 2014), and *Medicago truncatula* (version Mt4.0v1; Tang et al. 2014). *Glycine max* shows evidence of a WGD approximately 13 mya that is not shared with *Phaseolus vulgaris* or *Medicago truncatula* (e.g. Cannon et al. 2015); however, we assume that it is sufficiently old such that homeologous genes have survived the initial genome fractionation process following the WGD. We analyzed genomes from these two plant families separately due to GC composition biases between grasses and other angiosperm groups that could bias our analyses (Alexandrov et al 2009; Tiley and Burleigh 2015).

Annotated protein sequences from the grass and legume genomes were clustered separately into orthologous groups with OrthoFinder (Emms and Kelly 2015). For all of the resulting clusters that included at least two sequences from each species with at least 100 codons per species, we constructed codon alignments with PRANK (Löytynoja and Goldman 2005) using the default codon options with insertion but allowing ten iterations. We estimated a maximum likelihood (ML) gene tree from each
codon alignment under the GTR + Γ model (Tavaré 1986, Yang 1994) with RAxML (Stamatakis 2014). We then attempted to reduce potential topological errors in gene tree estimation with TreeFix, which searches for equally likely alternate gene tree topologies that minimize the number of implied duplications and losses (Wu et al. 2013). For the TreeFix analysis, we used rooted species trees based on previous phylogenetic analyses (e.g., Ruhfel et al. 2014).

Finding Syntenic Genes within and between Genomes

We identified syntenic gene pairs within and between genomes for the grasses and legumes with DAGchainer (Haas et al. 2004). We defined syntenity based on relative gene order and we determined homology between gene pairs based on BLASTP (Altschul et al. 1990; e-value < 1e-5). In the DAGchainer analyses, we examined syntenity over a range of parameters and chose a minimum of 10 genes needed to seed a syntenic region and set the maximum number of interleaving genes between syntenic gene pairs at 20, as these parameters resulted in the longest syntenic blocks with the most conservative settings. We used an approach adapted from Wang et al. (2012) to classify duplicate genes. We classified genes on syntenic blocks within a genome as WGD duplicates. We considered duplicate gene pairs located on the same chromosome but not on self-syntenic regions as cis-SSDs. We considered all duplicate gene pairs located on separate chromosomes and not in self-syntenic regions as trans-SSDs. If a genomic region was syntenic across species at the root of the species tree (i.e., a syntenic block is found in *Sorghum bicolor* and at least one other grass taxon or *Medicago truncatula* and at least one other legume taxon), we called this an evolutionarily conserved syntenic region. We made circular plots showing the
distribution of evolutionarily conserved syntenic regions across species and duplicate gene pairs within species with Circos (Krzywinski et al. 2009).

**Hypothesis Testing with Rates of Molecular Evolution**

We used gene trees and information on pairwise gene duplication relationships to test 1) differences in average $dN/dS$ following different types of gene duplications (i.e., WGD, cis-SSD, trans-SSD) and 2) asymmetry in $dN/dS$ following different types of gene duplications. For each gene tree, we mapped pairwise gene duplication relationships onto nodes using least common ancestor mapping, which infers a most parsimonious gene duplication mapping on a species tree (e.g., Goodman et al. 1979; Page 1994; Guigó et al. 1996). We assumed a model of gene gain only, such that each non-speciation node was a duplication node, using the mapping algorithm from Goodman et al. (1979). To further classify gene duplication nodes as representing a WGD, cis-SSD, or trans-SSD, we used the pairwise self-synteny data among sequences to find the most parsimonious reconciliation of gene duplication types on gene trees (Object 5-1, Figure S2). First, we identified WGD nodes by finding the least common ancestors for all pairs of sequences that were syntenic within each species. Then we found the least common ancestors of all cis-SSD gene pairs on the gene trees, excluding any nodes that were already annotated as WGD nodes. Finally, we mapped trans-SSD relationships onto the gene trees, excluding nodes previously annotated as WGD or cis-SSD nodes. In our analyses, we only used gene trees in which all sequences were annotated on a chromosome; gene trees containing genes on unmapped scaffolds were not used. For all of these gene trees, we could annotate all gene duplication nodes as WGDs, cis-SSDs, or trans-SSDs. Secondly, in our main
analysis, we only used gene trees containing at least one duplication node for WGDs, cis-SSDs, and trans-SSD and at least two sequences per species.

After annotating the gene tree nodes with gene duplication mechanisms, we used the GY94 codon model (Goldman and Yang 1994) to test nested molecular evolution hypotheses using likelihood ratio tests (LRTs) with PAML (Yang 2007). We first addressed if there were different $dN/dS$ estimates on branches following the different gene duplication classes. To do this, we binned gene tree branches in successively more complex models, yielding a 1-rate, 2-rate, 3-rate, and 4-rate models. The 1-rate model assumes all branches in the gene tree have the same $dN/dS$. The 2-rate model assumes speciation and duplication branches have independent $dN/dS$. The 3-rate model allows speciation, WGD, and non-WGD branches to have separate $dN/dS$. Finally, the 4-rate model distinguishes among speciation, WGD, cis-SSD, and trans-SSD branches. Each model introduces only one additional free parameter, which allows a LRT that assumes a $X^2$ distribution with one degree of freedom. A graphical description of our hypothesis testing procedure is given in Figure S3 (Object 5-1). A Benjamini-Hochberg correction with a 5% FDR (Benjamini and Hochberg 1995) was applied across gene trees.

For the 2, 3, and 4-rate trees we tested for asymmetry among duplication branches (Object 5-1, Figure S4). We defined asymmetry as when two daughter branches following a gene duplication event have different $dN/dS$. For each gene tree, we assume branches follow their previous 2-rate, 3-rate, or 4-rate designations. Then, we visit each duplication node and allow both daughter branches to have an independent, but shared, $dN/dS$ from the rest of the gene tree. This is the null
hypothesis for a specific gene duplication node. The alternative hypothesis then allows both daughter branches to have their own $dN/dS$. Since there is only one more parameter in the alternative hypothesis compared to the null hypothesis, we assume the LRT statistic is $\sim \chi^2_1$ and apply a nominal $\alpha$ of 5%.

**Enrichment of Gene Ontological Categories**

We were interested if asymmetry in $dN/dS$ following WGD or non-WGD duplications was strongly associated with specific GO functions, and if symmetry of WGD duplicates might explain transcription factor retention following WGDs (Cannon et al. 2004; Seoighe and Gehring 2004; Maere et al. 2005). We used GO annotations from Phytozome gene annotations, which were predicted from interpro2go (Mitchell et al 2015). We assigned molecular function GO annotations to individual gene trees using a list of GO Slim plant categories downloaded from the Gene Ontology Consortium (http://geneontology.org/page/download-ontology; last accessed 12 December 2016).

We tested for both under-representation and over-representation of molecular functions based on the asymmetry of $dN/dS$ across gene duplications represented in the 3-rate, and 4-rate gene trees using a one-sided Fisher exact test in R (R Core Team 2015). We constructed a 2x2 contingency table for each GO function by comparing asymmetry and non-asymmetry among WGD and non-WGD duplications for both the 3-rate and 4-rate gene trees.

**Associating Synteny across Species with Rates of Molecular Evolution**

To test if patterns of symmetry and asymmetry of $dN/dS$ across different gene duplication classes could be explained by genome architecture, we mapped evolutionarily conserved regions onto nodes of our gene trees. We used the pairwise synteny data across species to identify gene sequences at the tips of all gene trees that
existed in evolutionarily conserved syntenic regions. We then traversed each gene tree from the tips to the root. If both child nodes were syntenic, then their parent node was also considered to be syntenic. We used this data to form 2x2 contingency tables based on whether gene tree nodes were syntenic or non-syntenic as well as symmetric or non-symmetric for each duplication class across our 2-rate, 3-rate, and 4-rate gene trees. This allowed us to perform $X^2$ tests for under- or over-representation of any category for the 3-rate and 4-rate gene trees, using results from the 2-rate gene trees as expected proportions.

**Evaluating Time-Dependence of Asymmetry after Gene Duplication**

To test whether differences in $dN/dS$ between pairs of branches after gene duplication was due to the timing of the duplication, we estimated ultrametric branch lengths for each gene tree using $dS$ branch lengths and assuming a global molecular clock. We made branch lengths on the gene tree ultrametric with r8s (Sanderson 2003) using the Langley-Fitch ML estimator (Langley and Fitch 1974). We fixed the timing of speciation nodes using the median diversification age estimates from timetree (Hedges et al. 2015). We then treated symmetry and asymmetry following gene duplication as a binary character and performed logistic regression analyses with node ages in R (R Core Team 2015). We excluded all nodes with ages greater than 150 million years, between 2% and 5% of nodes across all data sets, in logistic regression analyses due to uncertainty in estimating unconstrained root node ages.
Figure 5-1. Distributions of $dN/dS$ among speciation, and duplication branches across gene trees in grasses. The 1-rate model represents gene trees where a single $dN/dS$ parameter fits all branches. The 2-rate model partitions branches into duplication and speciation. The 3-rate model discriminates between WGD duplications and non-WGD duplications while the 4-rate model further parameterizes cis-SSDs and trans-SSDs with independent $dN/dS$. Dashed vertical lines represent median $dN/dS$ values from their respective distributions.
Figure 5-2. Each duplication node in the 2-rate, 3-rate, and 4-rate gene trees were tested for different $dN/dS$ following duplication in grasses. The Y-axis is ordered from highest to lowest $dN/dS$ estimates, first showing asymmetrical than symmetrical branches. Asymmetry in $dN/dS$ show grey circles for the lower $dN/dS$ estimate and blue circles for the higher $dN/dS$ estimate; black lines are to distinguish the distance between points. A single grey circle represents symmetrical $dN/dS$ estimates in both branches following duplication.
Figure 5-3. Distributions of $dN/dS$ among speciation, and duplication branches across gene trees in legumes. The 1-rate model represents gene trees where a single $dN/dS$ parameter fits all branches. The 2-rate model partitions branches into duplication and speciation. The 3-rate model discriminates between WGD duplications and non-WGD duplications while the 4-rate model further parameterizes cis-SSDs and trans-SSDs with independent $dN/dS$. Dashed vertical lines represent median $dN/dS$ values from their respective distributions.
Figure 5-4. Each duplication node in the 2-rate, 3-rate, and 4-rate gene trees were tested for different $dN/dS$ following duplication in legumes. The Y-axis is ordered from highest to lowest $dN/dS$ estimates, first showing asymmetrical than symmetrical branches. Asymmetry in $dN/dS$ show grey circles for the lower $dN/dS$ estimate and blue circles for the higher $dN/dS$ estimate; black lines are to distinguish the distance between points. A single grey circle represents symmetrical $dN/dS$ estimates in both branches following duplication.
Figure 5-5. Distributions of synteny between species and gene duplication mechanisms within species. Grey links represent evolutionarily conserved syntenic blocks while orange links are WGD duplicates, purple links are cis-SSDs, and pink links are trans-SSDs. The grass data chromosomes are shown in brown (*Sorghum bicolor*), blue (*Oryza sativa*), and gold (*Brachypodium distachyon*). The legume data chromosomes are shown in brown (*Medicago truncatula*), blue (*Glycine max*), and gold (*Phaseolus vulgaris*). Tick marks on chromosomes represent 10 megabase intervals.
Table 5-1. Probabilities for counts of symmetric and asymmetric WGD and non-WGD gene pairs from a two-sided Fisher exact test. Counts for each GO molecular function are 2x2 contingency tables.

<table>
<thead>
<tr>
<th>Family</th>
<th>Gene Trees</th>
<th>GO ID</th>
<th>molecular function</th>
<th>Asymmetric WGD</th>
<th>Symmetric WGD</th>
<th>Asymmetric non-WGD</th>
<th>Symmetric non-WGD</th>
<th># duplicate pairs</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grasses 3-rate trees</td>
<td>GO:0005198</td>
<td>structural molecule activity</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>12</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0016787</td>
<td>hydrolase activity</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>9</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0005615</td>
<td>protein binding</td>
<td>0</td>
<td>9</td>
<td>7</td>
<td>20</td>
<td>36</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0003824</td>
<td>catalytic activity</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0004871</td>
<td>enzyme regulator activity</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0030234</td>
<td></td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>9</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0005215</td>
<td>transporter activity</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0016740</td>
<td>transferase activity</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4-rate trees</td>
<td>GO:0005515</td>
<td>protein binding</td>
<td>0</td>
<td>43</td>
<td>27</td>
<td>194</td>
<td>264</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0016787</td>
<td>hydrolase activity</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>9</td>
<td>15</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0005198</td>
<td>structural molecule activity</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>10</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>transcription factor activity</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>12</td>
<td>20</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0003700</td>
<td>DNA binding</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>41</td>
<td>51</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0003824</td>
<td>catalytic activity</td>
<td>0</td>
<td>6</td>
<td>4</td>
<td>12</td>
<td>18</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0016740</td>
<td>transferase activity</td>
<td>0</td>
<td>5</td>
<td>7</td>
<td>28</td>
<td>40</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0003677</td>
<td>DNA binding</td>
<td>0</td>
<td>5</td>
<td>7</td>
<td>28</td>
<td>40</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0005215</td>
<td>transporter activity</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>7</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0004871</td>
<td>signal transducer activity</td>
<td>0</td>
<td>6</td>
<td>3</td>
<td>17</td>
<td>26</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0003700</td>
<td>DNA binding</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>12</td>
<td>20</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Legumes 3-rate trees</td>
<td>GO:0003677</td>
<td>DNA binding</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>15</td>
<td>17</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0016787</td>
<td>hydrolase activity</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>8</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0005515</td>
<td>protein binding</td>
<td>1</td>
<td>20</td>
<td>3</td>
<td>47</td>
<td>71</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4-rate trees</td>
<td>GO:0005515</td>
<td>protein binding</td>
<td>2</td>
<td>31</td>
<td>10</td>
<td>84</td>
<td>127</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0003677</td>
<td>DNA binding</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>25</td>
<td>32</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0003824</td>
<td>catalytic activity</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Table 5-2. Counts of symmetric and asymmetric gene pairs on syntenic blocks maintained across species and syntenic blocks not maintained across species. Expected numbers of genes in each category for the 3-rate and 4-rate hypothesis tests are based on observed counts in the 2-rate hypothesis tests.

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Hypothesis</th>
<th>Non-Syntenic/ Symmetric</th>
<th>Syntenic/ Symmetric</th>
<th>Non-Syntenic/ Assymetric</th>
<th>Syntenic/ Assymetric</th>
<th>X²</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grasses</td>
<td>2-rate</td>
<td>249</td>
<td>11</td>
<td>22</td>
<td>2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Grasses</td>
<td>3-rate WGD</td>
<td>16</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>262.1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Grasses</td>
<td>3-rate non-WGD</td>
<td>58</td>
<td>2</td>
<td>9</td>
<td>2</td>
<td>7.44</td>
<td>0.006</td>
</tr>
<tr>
<td>Grasses</td>
<td>4-rate WGD</td>
<td>46</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>126</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Grasses</td>
<td>4-rate SSD</td>
<td>161</td>
<td>11</td>
<td>15</td>
<td>3</td>
<td>4.04</td>
<td>0.044</td>
</tr>
<tr>
<td>Grasses</td>
<td>4-rate trans-SSD</td>
<td>84</td>
<td>0</td>
<td>14</td>
<td>1</td>
<td>9.3</td>
<td>0.002</td>
</tr>
<tr>
<td>Legumes</td>
<td>2-rate</td>
<td>391</td>
<td>34</td>
<td>30</td>
<td>2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Legumes</td>
<td>3-rate WGD</td>
<td>59</td>
<td>31</td>
<td>11</td>
<td>5</td>
<td>131.6</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Legumes</td>
<td>3-rate non-WGD</td>
<td>233</td>
<td>9</td>
<td>20</td>
<td>1</td>
<td>6.06</td>
<td>0.014</td>
</tr>
<tr>
<td>Legumes</td>
<td>4-rate WGD</td>
<td>87</td>
<td>45</td>
<td>8</td>
<td>2</td>
<td>126.9</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Legumes</td>
<td>4-rate cis-SSD</td>
<td>264</td>
<td>22</td>
<td>31</td>
<td>2</td>
<td>5.52</td>
<td>0.019</td>
</tr>
<tr>
<td>Legumes</td>
<td>4-rate trans-SSD</td>
<td>88</td>
<td>6</td>
<td>14</td>
<td>1</td>
<td>8.03</td>
<td>0.005</td>
</tr>
</tbody>
</table>
Object 5-1. Supplementary material for Chapter 5
Our understanding of how eukaryotic genomes evolve has advanced dramatically from investigations in plant genomics. While plant genomes are among the most difficult to sequence, assemble, and annotate, they provide unique opportunities to research gene family evolution for these very same reasons. Chromosome rearrangements, multiple rounds of WGD, and the expansion and contraction of transposable elements have led to high rates of gene gain and loss, providing new insights on the importance of gene duplication in biodiversity.

Here, we have aimed to show how recombination rate can be a factor in the survival of duplicate genes. Recombination acts to preserve duplicate gene copies largely by increasing the efficacy purifying selection. Duplicate gene copies in regions of low recombination may be subject to relaxed purifying selection and ultimately experience pseudogenization at a higher rate than duplicate copies in regions of high recombination. We demonstrated these effects with genome-wide comparisons of recombination rate and gene-wide $dN/dS$, and by showing that duplicate copies in regions of synteny conserved across species tend to have similar $dN/dS$ compared to duplicate genes not in syntenic regions. This effect was prominent in duplicate genes produced by WGDs. Our results partially explain gene retention biases following WGDs, but also give a non-adaptive perspective to the role of WGDs in plant evolution.

The adaptive significance of WGDs in plant evolution are further called into question by highlighting the weaknesses of methods currently used to detect WGDs. The $K_s$ plot approach is capable of accurately detecting WGDs in limited window of intermediate time, and only when a substantial amount of the duplicate genes have
been retained. Thus, it is not surprising that many WGDs can be dated to 65 MYA. We did present an alternative approach to statistically testing both the existence and timing of a WGD using a probabilistic model of gene gain and loss. The gene count approach appears more sensitive, with retention rate estimates near 1%. However, the oldest of plant WGDs, such as an event that preceeded the common ancestor of all angiosperms, will likely remain contentious as even the gene count method lacks power to detect these extremely old events near the root of major clades.

Our results do have a number of limitations though. While computational investigations, even with our rigorous hypothesis testing approaches, can leverage large amounts of data across many taxa, the most we can gleam from these studies are associations and future hypotheses. For instance, the effects of recombination rate variation on duplicate gene retention have yet to be shown experimentally. Our ability to do so has been limited, until recently, to produce high quality genome assemblies where we also have dense genetic maps for natural populations. Any such effort is costly and time-intensive for one genome, let alone multiple populations. To advance our understanding on how population genetic processes affect traits at the macroevolutionary scale, such gene duplication and loss, we must advance our ability to integrate syntenic data from well-assembled genomes with quantitative expression data and estimates from recombination with genetic maps. As biotechnology and computational methods make such tasks feasible, plants will remain ideal candidates for exploring eukaryotic evolution with these approaches.
LIST OF REFERENCES


Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc Natl Acad Sci USA. 106:5737-5742.


Fullerton SM, Carvalho AB, Clark AG. 2001. Local rates of recombination are positively correlated with GC content in the human genome. Mol Biol Evol. 18:1139-42.


Kellogg EA. 2016. Has the connection between polyploidy and diversification actually been tested? Curr Opin Plant Biol. 30:25-32.


Lyons E, Pedersen B, Kane J, Freeling M. 2008. The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. Tropical Plant Biology 1:181-190.


Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in does-sensitive genes. Genome Res. 16:934-946.


BIOGRAPHICAL SKETCH

George was born in Hammond, LA in 1990. He was largely interested in music and baking, but always cared deeply for nature and living things. In 2008, George enrolled in the Biology program at Southeastern Louisiana University. He conducted research with Dr. Rick Miller on a range of questions in plant systematics, population genetics of sweet potatoes, and molecular evolution in flower pigment pathways. During his sophomore year, George became more interested in mathematics and the theoretical aspects of population genetics and phylogeny estimation. He earned a B.Sc. in biology with a minor mathematics in 2012, and started his Ph.D. later that year in the lab of Dr. Gordon Burleigh. For his doctoral dissertation, George has worked on a number of empirical investigations in genome evolution of plants, birds, and insects, while developing numerical and computational approaches to address a broad range of questions in molecular evolution. Much of his Ph.D. research has highlighted the importance of population-level processes for explaining differences across the genome between species. Upon completion of his Ph.D. program, George will be moving to a postdoctoral research position in Dr. Anne Yoder's lab at Duke in Durham, NC. His postdoctoral research will include investigating lemur speciation and the evolution of grassland communities in Madagascar to promote conservation and the importance of biodiversity.