2017

Evolution of duplicated pathways and networks in polyploid cotton

Joseph Patrick Gallagher
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Botany Commons, Developmental Biology Commons, Evolution Commons, and the Genetics Commons

Recommended Citation
https://lib.dr.iastate.edu/etd/16515

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Evolution of duplicated pathways and networks in polyploid cotton

by

Joseph Patrick Gallagher

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Genetics and Genomics

Program of Study Committee
Jonathan F. Wendel, Major Professor
Steven Cannon
John Nason
Basil Nikolau
Amy Toth

Iowa State University

Ames, Iowa

2017

Copyright © Joseph Patrick Gallagher. All rights reserved.
I dedicate this work to

the friends and family who have supported me for the past six years,

my girlfriend Nikki,

my parents, Diane and Chuck,

my sisters and brothers-in-law, Katie, Kelly, Dave, and Fred,

and my niece, Frances,

and my dog, Rosie.
TABLE OF CONTENTS

ABSTRACT ..............................................................................................................v

CHAPTER 1. INTRODUCTION ..................................................................................1
Dissertation Organization .........................................................................................1
Background and Research Objectives ....................................................................3
References .............................................................................................................8

CHAPTER 2. INSIGHTS INTO THE ECOLOGY AND EVOLUTION OF
POLYPLOID PLANTS THROUGH NETWORK ANALYSIS .....................................14
Abstract .............................................................................................................14
Introduction ..........................................................................................................15
Polyploidy and the Prospects of Gene Network Analysis .......................................19
Adding Power to Polyploid Networks with Prior Knowledge and Other Data Types ..........................................................................................................................25
Conclusions .........................................................................................................35
Acknowledgements ...............................................................................................36
Author Contributions ...........................................................................................36
Data Accessibility Statement ..................................................................................36
Methods ...............................................................................................................36
References .............................................................................................................37
Tables and Figures .................................................................................................56

CHAPTER 3. A NEW SPECIES OF COTTON FROM WAKE ATOLL,
GOSSEYUM STEPHENSII (MALVACEAE) .................................................................64
Abstract .............................................................................................................64
Introduction ..........................................................................................................64
Materials and Methods .........................................................................................66
Results ..................................................................................................................69
Taxonomic Treatment ............................................................................................70
Acknowledgments .................................................................................................73
References .............................................................................................................74
Tables and Figures .................................................................................................77

CHAPTER 4. EFFECTS OF GENOME DUPLICATION AND DOMESTICATION ON THE FIBER GENE CO-EXPRESSION NETWORK OF
GOSSEYUM HIRSUTUM L ............................................................................................88
Abstract .............................................................................................................88
Introduction ..........................................................................................................89
Materials and Methods .........................................................................................92
Results ..................................................................................................................95
Discussion ............................................................................................................105
Conclusions .........................................................................................................113
Acknowledgements ..............................................................................................114
References ...........................................................................................................114
CHAPTER 5. EVOLUTION AND GENE EXPRESSION DYNAMICS OF THE ANTHOXYANIN BIOSYNTHESIS PATHWAY IN *GOSSYPIUM*..........................133
   Abstract.................................................................133
   Introduction.........................................................134
   Materials and Methods............................................139
   Results.................................................................143
   Discussion...........................................................157
   Conclusions.........................................................164
   Acknowledgements................................................164
   References..........................................................165
   Tables and Figures................................................175
   Supplementary Files..............................................188

CHAPTER 6. CONCLUSIONS..............................................189
Polyploidy is an evolutionary phenomenon resulting in the duplication of the whole genome. The merger of two diverged genomes, in the case of allopolyploidy, had multiple effects on the evolution of genes, gene expression, and the structure of the newly doubled genome. Gene pathways and gene co-expression networks are also duplicated as a whole and must be reconciled in the nascent polyploid. A considerable body of work addresses how pathways and gene co-expression networks evolve in diploids. Here I extend this approach to address how duplicated pathways and networks evolve in polyploids.

The effects of polyploidy on gene pathways and gene co-expression network duplication are investigated in the natural cotton allopolyploids *Gossypium hirsutum* and *G. barbadense*. I used targeted sequence capture to look at evolutionary rates in the genes of the anthocyanin biosynthesis pathway across in over 40 different polyploid accessions. The evolutionary rates of these genes do not correlate with position or branching as would be expected. I further investigated the expression of the genes in the anthocyanin biosynthesis pathway, and of the whole genome, by generating RNA-seq libraries across several tissues in the polyploid cotton species and their diploid progenitors. Evolutionary rates were correlated with expression levels, module assignment, and connectivity. More generally, the gene co-expression network as a whole generally is preserved between diploids and polyploids, but certain modules exhibit specific homoeolog biases and non-additive expression.

I also studied expression in fibers from wild and domesticated *G. hirsutum*, adding a layer of complexity, in the form of the strong selective pressure of domestication, to the gene co-expression analyses. While I still find general preservation of the gene co-expression network,
some modules show extreme homoeolog bias in the fiber as well. Comparisons between the wild and domesticated fiber co-expression networks show that there are drastic alterations in topology of the network due to strong, human-mediated selection. In addition, gene co-expression relationships have been strongly rewired as a consequence of the process of domestication.
CHAPTER 1. INTRODUCTION

Dissertation Organization

This dissertation is organized into six chapters. The first chapter introduces polyploidy and the evolution of pathways and networks. I describe the Gossypium species, colloquially known as cotton, and describe the objectives of this work.

Chapter 2 contains a perspective on the possibilities gene co-expression networks present for expanding our knowledge regarding polyploidy. Previous research regarding the ecological and evolutionary effects of polyploidy are discussed, and recent research that has integrated network analysis into its workflow is reviewed. Finally, an example of gene co-expression network analysis involving cotton fiber from the polyploid Gossypium hirsutum is provided, with a focus on the profilin family of genes. This chapter was published in Molecular Ecology in 2016. I undertook this work with Corrinne E. Grover, who is co-first author, Guanjing Hu, and Jonathan F. Wendel. All four of us conceived of the paper. Corrinne, Guanjing, and I performed the bioinformatic analysis. Corrinne, Guanjing, and I drafted the manuscript with help from Jonathan. All authors participated in the editing process.

Chapter 3 provides evidence for a new species of polyploid cotton, Gossypium stephensii J. Gallagher, C. Grover, & Wendel; this paper became possible due to the sequence capture data generated for the study in Chapter 5. Using morphological data and phylogenetics of chloroplast and nuclear genomes, we identified the accessions of cotton from the Wake Atoll as novel species. We provide the species description within the manuscript. This work was published in Plant Systematics in 2017. I undertook this research with Kristin Rex, Matthew Moran, Corrinne E. Grover, and Jonathan Wendel. Jonathan and I conceived of this research. Kristin and Matthew
provided samples of cotton from the Wake Atoll. I performed morphological measurements and generated descriptions. Corrinne performed genome and gene assembly. I performed phylogenetic and principal coordinates analysis with aid from Corrinne. I constructed herbarium specimens. I drafted the manuscript with assistance from Jonathan, and all authors provided edits and comments.

Chapter 4 is an analysis of the effects of domestication on the cotton fibers of *G. hirsutum*, a polyploid species of cotton. Here, we performed RNA-sequencing on accessions of both wild and domesticated cotton during part of the fiber development process. We combined differential expression, weighted gene co-expression, and differential co-expression analysis to gain a network viewpoint of the effects of domestication on the fibers of this species. By examining these data both in the aggregate, with gene pairs counted as a single gene, and as separate homoeologs, we studied the interaction of domestication with polyploidy. This work is being prepared for submission to the journal *Genome Biology and Evolution*. This research was performed with Guanjing Hu, Corrinne E. Grover, Josef J. Jareczek, and Jonathan F. Wendel. Jonathan, Corrinne, Guanjing, and I planned the research. I performed the data analysis, with assistance from Guanjing, Corrinne, and Josef. I drafted the manuscript with assistance from Jonathan and Guanjing.

Chapter 5 is a study of the anthocyanin biosynthesis pathway in allopolyploid cotton and its diploid progenitors. Anthocyanin biosynthesis gene candidates were identified in the D5 cotton reference genome and used as baits to perform targeted sequence capture across a panel of cotton species, including the diploid progenitors and two allopolyploid cottons *G. hirsutum* and *G. barbadense*. These genes were assembled and used to analyze the genic evolution of the pathway. At the same time, RNA-seq was performed on five tissues in the diploids and several
polyploid cotton accessions. These libraries were used to perform weighted gene co-expression network analysis to discern how the expression of the pathway genes was evolving due to polyploidy. This work is being prepared for submission to *Molecular Biology and Evolution*. This research was undertaken with Corrinne E. Grover, Guanjing Hu, Emma Miller, and Jonathan Wendel. Corrinne, Jonathan, and I planned the project. I performed DNA extraction. Corrinne and I performed targeted sequence capture. Corrinne and I assembled the genes, and I performed the evolutionary analysis. Emma and I performed RNA extractions. I performed gene-expression analysis with assistance from Guanjing Hu. I drafted the manuscript with the assistance of Jonathan and Guanjing.

Chapter 6 provides a brief discussion of the general conclusions of the thesis.

Background and Research Objectives

**Polyploidy in evolution**

Polyploidy is the state of having three or more complete sets of chromosomes. It is particularly prevalent in land plants, as all flowering plants are considered to be paleopolyploids – polyploids that have undergone genome fractionation so that they no longer contain two sets of each chromosome – while more than 15% of angiosperms and 31% of ferns are considered to be neopolyploids – polyploids that still retain complete copies of the doubled genomes (Wood et al. 2009; Jiao et al. 2011). Polyploids come in two types: autopolyploids, in which the two progenitors are so closely related that their chromosomes can pair and possibly form tetrads during cell division, and allopolyploids, in which the two progenitors are different enough that their chromosomes cannot pair during cell division (Birchler and Veitia 2012). Intermediate
between these two forms of polyploids are segmental allopolyploids, which can pair at certain regions between homoeologous chromosomes, but not all regions (Stebbins 1950).

The effects of polyploidy on genes and the genome across multiple plants species have become well studied, particularly with the advent of genomics (Wendel et al. 2016). Following polyploidy, chromosome-level changes such as aneuploidy, compensated aneuploidy, and intergenomic transfer can occur rapidly (Xiong et al. 2011; Chester et al. 2012). Polyploidy also leads to dramatic changes at the genic and genomic level: differential evolutionary rates, loss of homoeologous genes, gene conversion, fractionation, and transposable element proliferation (Senchina et al. 2003; Buggs et al. 2009; Hu et al. 2010; Schnable et al. 2011; Buggs et al. 2012). Genes can also change in function either through subfunctionalization, which is when the gene function is split between two homoeologs by developmental stage or tissue type, or neofunctionalization, which is when one homoeolog takes on a new function (biochemically, or in expression domain), while the other homoeolog maintains function (Stephens 1951; Ohno 1970; Conant and Wolfe 2008). These changes are not limited to changes at the sequence level, but also affect transcription and ultimately translation. Changes in expression can lead to homoeolog silencing, homoeolog expression bias, expression level dominance, and partitioned expression (Adams et al. 2003; Buggs et al. 2009; Rapp et al. 2009; Buggs et al. 2010; Flagel and Wendel 2010; Yoo et al. 2013). These changes in expression can then propagate forward into protein levels, although these changes are not consistently conserved (Hu et al. 2011). While many of these phenomena have been examined for whole genomes, they only very recently have been considered in a whole pathway context (Bekaert et al. 2011; Pfeifer et al. 2014; Hu et al. 2016).
Evolution of pathways and networks

The effect of pathway interactions on gene evolution has been studied in several different pathways and systems. Using the anthocyanin biosynthesis pathway as a model, Rausher et al. (1999) showed that, when comparing maize, snapdragon, and morning glory, several phylogenetically disparate angiosperms, pathway position had an effect on the evolutionary rate of genes in a biosynthetic pathway: the genes at the beginning of the pathway had a lower rate of nonsynonymous substitution than those at the end of the pathway (Rausher et al. 1999). This has been extended to other pathways and pathway features, such as flux, connectivity, or branch points (Ramsay et al. 2009; Yang et al. 2009; Olson-Manning et al. 2012). While these findings show several trends for diploid plants, how these constraints act on the evolution of genes in the duplicated pathways of polyploid plants is yet to be explored. It seems likely that genes in polyploid pathways will evolve according to some combination of these features and those evolutionary forces known to affect polyploids (e.g. fractionation). We hypothesize that genes from pathways duplicated via polyploidy would evolve at faster rates than seen in the studies here, due to the reduced constraint caused by the multiple gene copies.

Additionally, genes can be considered as parts of co-expression networks. This is a network of gene relationship based on the similarity of their expression patterns; genes share a higher degree of expression similarity are grouped together in the network, often into a subnetwork called a co-expression module (Langfelder and Horvath 2008). These co-expression modules provide another means of grouping and studying genes within pathways, as well as for the whole genome. Several studies have used gene co-expression networks to identify genes and gene expression patterns responsible for various phenotypes (e.g. Swanson-Wagner et al. 2012;
Ichihashi et al. 2014), proving their likely usefulness for understanding the evolution of a pathway.

More recently, several studies have used gene co-expression networks to examine polyploidy (Pfeifer et al. 2014; Hu et al. 2016). Pfeifer et al. (2014) examined co-expression within the wheat genome. They built a network using total expression data from the three subgenomes in wheat, and then described homoeolog bias within each module. They demonstrate that these biased modules are enriched for certain GO terms, suggesting that certain subgenomes are responsible for specific functions in wheat. In polyploid cotton, Hu et al. (2016) used co-expression networks to describe the changes in seeds and the oil and fatty acid biosynthesis pathways following domestication. They showed that the transcriptional network changed as a result of domestication, and that fatty acid biosynthesis pathway specifically increased in module density and in clustering coefficient. They also found that certain features of polyploidy – such as expression level dominance and transgressive expression – could be extended from the level of individual genes to modules. Further examination of polyploid gene co-expression networks would be beneficial; similar to what was shown for pathways, selective constraint has become associated with connectivity in the gene co-expression network in diploid (Mahler et al. 2017); testing if this holds for polyploids would help further elucidate the evolution of their expression dynamics.

Cotton as a model for evolution of pathways and networks

As stated above, there are many stabilized neopolyploids in the angiosperms, including many of the crop species that humans rely on for basic needs. One particularly good polyploid model is cotton. Between 5 and 10 mya, the genus *Gossypium* split into two branches, the New World cottons, or the D-genome cottons, and the Old World cottons (Wendel and Grover 2015).
Approximately 1-2 mya, one of the Old World species, a member of the A-genome group, made a transoceanic voyage to the New World, where it formed an allopolyploid species following hybridization with a D-genome species (Wendel 1989; Senchina et al. 2003). This new allopolyploid lineage then diversified into as many as seven species (Grover et al. 2012a), two of which are Upland cotton and Pima cotton, the primary species that we raise as crops today.

Many of the known effects of polyploidy listed above were discovered in or already have been observed in cotton. For example, Adams et al. (2003) found tissue-specific homoeolog silencing for the AdhA gene; this subfunctionalization of gene expression was later extended to the whole genome level (Hovav et al. 2008). Rapp et al. (2009) found that gene expression between homoeologs exhibited genomic dominance in which the total expression level of a particular gene mimicked the expression level of one of the progenitors (now termed expression level dominance) (Grover et al. 2012b). They also discovered that homoeologs exhibited expression bias, i.e. more of the transcripts were derived from one of the subgenomes. Recently, Yoo et al. (2013) discovered that the expression level dominance uncovered by Rapp et al. was due to changes in expression level of the homoeolog that was derived from the non-dominant progenitor. Gene conversion, or non-reciprocal homoeologous recombination, in which the copy of the gene from one subgenome replaces the copy of the gene from the other subgenome, has also been observed (e.g. Salmon et al. 2010; Guo et al. 2014). As mentioned above, the gene co-expression dynamics of polyploid cotton have also been examined (Hu et al. 2016). These are just a handful of the effects of polyploidy that have been observed in cotton. Because the effects of polyploidy in this model are so well-characterized, Gossypium provides an excellent system in which to extend our understanding to the analysis of the evolution of duplicated pathways and gene co-expression networks.
Dissertation objectives

My doctoral research aims to develop our understanding of gene pathway and gene co-expression network changes as a result of polyploidy. The data sets analyzed and provided here should elucidate the evolution and expression changes in interconnected genes and modules, as well as how these have been affected by domestication. My specific research objectives are:

1. To understand how genes in a pathway evolve following polyploidy, especially compared to the current understanding of pathway features and their effect on evolution
2. To understand how polyploidy affects the evolution of gene expression, both for genes in a pathway, and for the whole gene co-expression network
3. To understand the effects of selection on the duplicated gene co-expression network, using domestication as an exemplary selective force

References


CHAPTER 2. INSIGHTS INTO THE ECOLOGY AND EVOLUTION OF POLYPLOID PLANTS THROUGH NETWORK ANALYSIS

A paper published in Molecular Ecology in 2016 (Molecular Ecology 25: 2644–2660)

Joseph P. Gallagher, Corrinne E. Grover, Guanjing Hu, and Jonathan F. Wendel

Abstract

Polyploidy is a widespread phenomenon throughout eukaryotes, with important ecological and evolutionary consequences. Although genes operate as components of complex pathways and networks, polyploid changes in genes and gene expression have typically been evaluated as either individual genes or as a part of broad-scale analyses. Network analysis has been fruitful in associating genomic and other “-omic”-based changes with phenotype for many systems. In polyploid species, network analysis has the potential to not only facilitate a better understanding of the complex “omic” underpinnings of phenotypic and ecological traits common to polyploidy, but also to provide novel insight into the interaction among duplicated genes and genomes. This adds perspective to the global patterns of expression (and other “omic”) change that accompany polyploidy and to the patterns of recruitment and/or loss of genes following polyploidization. While network analysis in polyploid species faces challenges common to other analyses of duplicated genomes, present technologies combined with thoughtful experimental design provide a powerful system to explore polyploid evolution. Here we demonstrate the utility and potential of network analysis to questions pertaining to polyploidy with an example involving evolution of the transgressively superior cotton fibers found in polyploid Gossypium hirsutum. By combining network analysis with prior knowledge, we provide further insights into
the role of profilins in fiber domestication and exemplify the potential for network analysis in polyploid species.

Introduction

Polyploidy is a widespread phenomenon throughout eukaryotes, with important ecological and evolutionary consequences (Leitch & Leitch 2008; Levin 1983; Matsushita et al. 2012; Ramsey & Schemske 1998, 2002; Soltis et al. 2014; Stebbins 1940; Van de Peer et al. 2009). Although both recent and ancient polyploidy events have been identified in animals and fungi, modern polyploidy is not a particularly active process in these major clades (Albertin & Marullo 2012; Muller 1925; Orr 1990). Conversely, the importance of polyploidy to modern plant species is difficult to overstate. It now is recognized that all flowering plants have experienced multiple rounds of polyploidy at some point in their ancestry (Bowers et al. 2003; Jiao et al. 2011; Wood et al. 2009), and that it remains an active evolutionary and ecological process in many lineages (McAllister et al. 2015; Ramsey 2011; Ramsey & Ramsey 2014; Soltis & Soltis 2012; Wendel 2015).

From a genomic perspective, the myriad consequences of polyploidy for the various “omes” have been broadly evaluated for a number of model angiosperm genera (Bao et al. 2011; Chalhoub et al. 2014; Chelaifa et al. 2010; Chester et al. 2012; Coate et al. 2014; Dong & Adams 2011; Flagel et al. 2012; Guan et al. 2014; Ha et al. 2011; Hu et al. 2015; Hu et al. 2013; International Wheat Genome Sequencing Consortium 2014; Koh et al. 2012; Li et al. 2014; Li et al. 2015; Ng et al. 2012; Page et al. 2013b; Pang et al. 2009; Paterson et al. 2012; Sehrish et al. 2014; Wang et al. 2004; Xiong et al. 2011; Xu et al. 2014; Yoo et al. 2013; Zhang et al. 2015b). Among the notable phenomena are inter-genomic exchanges between genomes that once were
isolated in divergent progenitor diploids (Flagel et al. 2012; Guo et al. 2014; Salmon et al. 2010; Wang & Paterson 2011); biased loss and/or fractionation of duplicated genes (hereafter “homoeologs”) (Ozkan et al. 2001; Tang et al. 2012; Thomas et al. 2006; Woodhouse et al. 2010); and a variety of types of changes in gene expression arising with and following polyploidy (Adams et al. 2003; Buggs et al. 2009; Buggs et al. 2010; Chelaifa et al. 2010; Coate et al. 2014; Comai et al. 2000; Flagel & Wendel 2010; Flagel et al. 2012; Rapp et al. 2009; Shi et al. 2012; Wang et al. 2004). There also is deep interest in the relationship between these varied phenomena and biological diversification (Ramsey 2011; Ramsey & Schemske 1998; Soltis & Soltis 1999; Soltis & Soltis 2012).

To date, alterations in genes and gene expression arising from polyploidy have been evaluated most commonly on a gene-by-gene basis (e.g. Adams et al. 2003), even when these are aggregated on a genome-wide scale (e.g. Yoo et al. 2013) for purposes of generalization. That is, comparisons have often been made between levels of gene expression between diploids and their polyploid derivatives, for any number of genes within the genomes in question. Because genes do not function in isolation, but as components of complex biological networks, it also is important to study gene expression of networks and pathways, as exemplified for plants in recent studies (Bekaert et al. 2011; Chang et al. 2010; Coate et al. 2013; Ni et al. 2009; Pfeifer et al. 2014). These analyses follow on the pioneering consideration of the network responses to polyploidy in yeast (Conant & Wolfe 2006; Makino & McLysaght 2012; Qian et al. 2011; van Hoek & Hogeweg 2009; Wagner 2001).

Until recently, the analysis of biological networks was unrealistic for most non-model organisms; however, technological advances such as RNA-seq and other molecular genetic methodologies, combined with conceptual and computational advances in gene co-expression
network reconstruction, have made network analysis an increasingly attractive tool for biological discovery. A gene co-expression network is a map of gene expression correlation among samples; nodes in the network represent genes or transcripts, while connections represent expression relationships (Albert 2005; Langfelder & Horvath 2008). An important advantage of gene co-expression networks is the statistical assignment of genes to clusters, called modules. These co-expression modules become more robust with a greater number and higher diversity of samples across tissues and time points within an organism. With appropriate depth of sampling, comparisons of networks between species, ecotypes, or genotypes may reveal changes in modules that explain interesting phenotypic differences. Consequently, we now have the potential to better understand the complex “omic” underpinnings of phenotypic and ecological traits, as exemplified by recent studies of maize domestication (Swanson-Wagner et al. 2012), tomato leaf variability (Ichihashi et al. 2014), and environmental regulation of gene expression in pines (Cañas et al. 2015).

Modern high-throughput techniques also facilitate large-scale identification of other molecular components (e.g. genes, RNAs, proteins, metabolites, epigenetic marks), as well as characterization of their expression patterns and interactions. Consequently, various complex datasets are frequently generated to answer ecological, evolutionary, and functional questions. The challenge then becomes how to distill these large multidimensional data sets into biologically informative conclusions. Network analysis is designed to capture interactions and dependencies among components, often independent of prior knowledge and the interactions of individual components. For non-model species, this is particularly attractive in that it permits genome-scale analysis of ecological or evolutionary traits among conditions, taxa, or developmental stages, as well as the identification of gene modules that are likely to be
functionally related. Although these correlations are based on indirect evidence of relationship, strong expression correlation over multiple tissues/stages/etc. is often indicative of functional relationships (e.g. genes activated by promoters with similar regulatory elements). Furthermore, conservation of co-expression patterns among species could indicate conservation of functional relationships (van Noort et al. 2003).

Analyses of biological networks hold promise for providing insights into the dynamics and resolution of polyploid genomes (Bekaert et al. 2011; Conant & Wolfe 2006; Pfeifer et al. 2014), particularly when network analyses are integrated with other “omics” data sets. Network-based analyses in non-plant models have already demonstrated the utility of gene co-expression networks to reveal functional changes in genes and gene modules (Conant & Wolfe 2006; van Noort et al. 2003). Similar analyses can be used to provide further insight into the age-old question of how two (or more) diverged genomes function in a common nucleus subsequent to merger (at the time of hybridization) and following genome doubling, ultimately enhancing our understanding of polyploid ecology and evolution. We introduce this perspective here, by reviewing the application of network analysis to polyploid research and the layering of additional data types to yield novel biological insights. We provide a brief empirical example of network analysis from our ongoing work in the cotton genus (Gossypium), drawing distinctions between the kinds of insights derived from traditional approaches to gene expression analysis and those derived from network tools.

Polyploidy and the prospects of gene network analysis

Network analysis in polyploid species has the capability of providing new perspectives into the interaction among duplicated genes and genomes and the changes that accompany
polyploidy. These include patterns of genome-wide gene loss/retention and gene expression changes across the transcriptome. Ultimately, this analysis could disentangle some of the underlying forces that govern polyploid evolution and ecology. Here we illustrate how network analysis could expand our current knowledge on polyploidy, focusing on gene retention, gene expression alterations, and phenotypic changes.

**Gene retention and loss following polyploidy**

One outstanding question concerns the dynamics of duplicate gene loss versus retention. This question has been considered from several perspectives, including selection related to broad functional categories (De Smet et al. 2013), gene dosage demands (Birchler & Veitia 2012), and preferential retention of homoeologs from the less fractionated of two genomes (Schnable et al. 2011). Each of these possibilities could be enlightened by a deeper consideration of genic interactions in biological pathways and networks, which necessarily entail a number of related functional or mechanistic constraints. Broad characterizations of the types of genes preferentially retained in duplicate, such as transcription and signaling-related genes in *Arabidopsis* (Blanc & Wolfe 2004; Maere et al. 2005; Seoighe & Gehring 2004), or structural genes in the Compositae (Barker et al. 2008), or those returned to singleton status, such as photosynthesis-related genes (De Smet et al. 2013), provide evolutionary clues into the dynamics of duplicate gene expression and retention following polyploidy, but of necessity “paint with a broad brush”. It seems likely that the compression of complex information into generalized categories conceals interactions among genes that influence retention and loss. We note that some explanations for patterns of duplicate gene loss invoke interaction among genes as important for gene retention or loss. The gene balance hypothesis, for example, posits that proteins that contribute to a multiunit complex are selectively maintained in the correct ratios, as imbalance among components of multiprotein
complexes could be deleterious (Birchler et al. 2005; Birchler & Veitia 2007, 2012; Conant et al. 2014; Papp et al. 2003). This concept has been extended to polyploid species, where the entire genic complement has been duplicated; that is, duplicate gene retention may be influenced by stoichiometric or “balance” considerations for the duplicated members of multiprotein complexes. The application of the gene balance hypothesis, however, requires knowledge of interactions among the encoded proteins. For example, Makino and McLysaght (2012) showed that, in human, yeast, and Arabidopsis, there is an enrichment of genes that have protein interactions with other genes in the same conserved block, as opposed to protein interactions across blocks. Their proximity and interaction suggest that these enriched genes might be co-regulated, and are therefore in keeping with the gene balance hypothesis.

An added layer of complexity arises when the duplicated genes are non-identical, as is the case for many or most genes in allopolyploids, i.e., polyploids derived from two divergent genomes. In this case there may be suboptimal interactions among the protein products from the newly combined genomes, leading to fitness differentials among multi-subunit proteins derived from homo-genomic vs. hetero-genomic protein assembly. In principle this may result in preferential removal or silencing of specific homoeologs or even a cascade of gene loss sufficient to reduce maladaptive interactions and maintain stoichiometric (gene) balance (Birchler & Veitia 2012; Chang et al. 2010). Insights into these possibilities are best facilitated when genes and their protein products are viewed in the context of their functional connections, or networks. An example of such an approach is provided by Bekaert et al. (2011), who used the Arabidopsis thaliana metabolic network, a map of functional interactions, to look at gene retention and loss following ancient polyploidy events; one of their findings was that genes retained in duplicate
from the most recent whole genome duplication were clustered in the network, i.e. interacting with each other, as would be expected according to the gene balance hypothesis.

**Duplicate gene expression after polyploidy**

One of the important realizations about polyploids is that gene expression is often massively altered relative to diploid progenitors (see Introduction). While gene retention is fundamentally binary (retained or not, the latter including the full spectrum of silencing mutations), gene expression is quantitative. Moreover, the aggregate expression of both homoeologs is of interest, not just the expression of individual members of each duplicate gene pair. In this respect duplicate gene expression may variously be transgressive relative to the progenitor diploids, additive (or average), or in some sense mimic only one of the two parents (Yoo et al. 2014). Broad patterns of expression alteration may hint at mechanistic causes of these changes, given predicted functional information of up- or down-regulated genes, homoeolog usage, etc.; however, here again information regarding interactions among genes holds promise for providing insight into mechanisms of expression alteration.

Classically, the causes of duplicate gene retention have entailed some form of “subfunctionalization” or “neofunctionalization” (Force et al. 1999; Lynch & Force 2000), although these terms are not mutually exclusive nor wholly sufficient (Conant et al. 2014). Neofunctionalization traditionally invokes a novel and presumably adaptive function for one homoeolog arising post-duplication while the other homoeolog(s) maintains its ancestral function (Ohno 1970; Stephens 1951), whereas subfunctionalization traditionally occurs as regulatory divergence, such that the ancestral aggregate expression becomes partitioned among homoeologs in the relevant tissues and/or stages (Force et al. 1999; Lynch & Force 2000). In the context of polyploidy, an important early observation was that of Adams et al. (2003), who discovered
reciprocal silencing of alternate homoeologs in various tissues from tetraploid cotton, demonstrating that subfunctionalization may be rapidly established following polyploidization. Similar work has since been conducted on a genome-wide scale for a number of species (Duarte et al. 2006; Hughes et al. 2014; Renny-Byfield et al. 2014; Roulin et al. 2013); in some cases, the data are more suggestive of neofunctionalization, although sub- and neo-functionalization may be hard to disentangle in many cases (Conant et al. 2014; Conant & Wolfe 2008). In the context of the present paper, the connectivity of individual homoeologs of different gene pairs may be an important consideration, one that only becomes evident using network approaches. For example, Pfeifer et al. (2014) generated a co-expression network for bread wheat, grouping all homoeologs into a single gene node to construct the network, and subsequently evaluating the contribution of each homoeolog to the expression pool for various cell types and developmental stages. Interestingly, they found certain network modules exhibit unbalanced expression bias, i.e. an overall favoring of expression of homoeologs derived from one parent (Grover et al. 2012), which could be associated with function (based on Gene Ontology categorization) and tissue type. This observation, that functional modules may become biased toward a specific homoeologous genome, has broad implications for understanding the evolution of polyploid species. For example, there may be phenotypic consequences that arise from, for example, preferential utilization of functionally related genes derived from only one homoeologous genome (e.g., phenotypic similarity to one parent for a given trait). This preferential utilization or functional differentiation of individual homoeologous networks may be revealed by conducting separate network analyses for each constituent set of homoeologs, rather than on data where gene expression for each duplicate gene pair has been lumped prior to analysis. One can envision that such analyses might reveal several phenomena, including: (1) an overall bias toward one parental...
genome in a reconstructed gene network; (2) subfunctionalization of networks, such that the alternate homoeologous networks are used in different tissues/stages, in full or in part; and (3) network neofunctionalization (all, or some) whereby one set of homoeologs from the same progenitor genome participates in a separate, novel, pathway. Phenomena such as these have been described in *Saccharomyces cerevisiae*, where co-expression networks appear to have partitioned for ancient sets of paralogs (Conant & Wolfe 2006), and in the neoallopolyploid plant *Arabidopsis suecica*, where genes with more highly correlated expression were more often derived from the same parent of origin (Chang et al. 2010).

Yet another mechanism that could modulate polyploid gene expression is epigenetic regulation (Madlung & Wendel 2013). Several high-throughput methods (e.g. bisulfite sequencing, ChIP-seq, DNase I hypersensitivity assays) extend the convenience of next-generation sequencing to the epigenetic arena and permit genome-wide surveys of DNA methylation, histone modifications, and chromatin state, all of which have the potential to affect gene expression (Deal & Henikoff 2011; Furey 2012; Meyer 2011; Tsompana & Buck 2014; Zhang et al. 2014). Similarly, small RNAs (e.g., siRNAs and miRNAs), which also function in expression regulation, are becoming targets of next-generation sequencing as well (Abrouk et al. 2012; Gong et al. 2013; Li et al. 2014). When these features differ between parental genomes, expression of each homoeolog in the polyploid can diverge from that of their parents (Buggs et al. 2011; Chelaifa et al. 2010; Xu et al. 2014). Gene co-expression networks that show altered expression patterns between parental species and polyploids may be caused by these epigenetic effects.

Each of the foregoing aspects of duplicate gene expression feed into the short- and long-term patterns of gene loss and retention following the origin of a polyploid species or clade. A
somewhat mysterious observation in this respect is the phenomenon of “biased fractionation” (Woodhouse et al. 2010), whereby a polyploid lineage exhibits biased loss of genes from one progenitor genome as it returns to a diploid state. First demonstrated in Arabidopsis (Thomas et al. 2006), biased fractionation has been observed for a number of other taxa (Renny-Byfield et al. 2015; Schnable et al. 2011; Tang et al. 2012), suggesting that this may be a common outcome of polyploidy. The causes underlying biased fractionation are not well-understood, but one can imagine how the interplay of phenomena such as gene dosage, maladaptive interactions, biased or novel expression patterns, and others introduced above may affect long-term gene retention bias. One may envision how an enhanced understanding of “genes in context”, i.e., of regulatory interactions and of biological pathways and networks, may help elucidate the neutral and selective forces that govern loss and retention, and thereby affect phenotypes of relevance to the ecology and evolution of polyploids.

**Evolutionary and ecological changes following polyploidy**

While the genetic changes listed above are interesting in their own right, the ultimate goal of these studies is to discover the underlying changes that lead to ecologically or evolutionarily important phenotypic properties. A number of traits have been associated with polyploidy in plants, such as enlarged cell size, larger overall organism size, and delayed development (i.e. the gigas effect); exploitation of new ecological niches; and physiological or biochemical novelty (Flagel & Wendel 2009; Levin 1983; Ramsey & Ramsey 2014; Soltis et al. 2014). For example, McIntyre (2012) showed that, for a set of Claytonia species of different ploidy levels, polyploids occupied a different niche compared to diploids. In Achillea borealis, Ramsey (2011) performed field transplant experiments showing that hexaploids had a distinct advantage over tetraploids in the xeric dune habitats in which they occur. Coate et al. (2013) showed an increased
photoprotective capacity in *Glycine dolichocarpa* compared to its diploid progenitors, as well as identifying the genetic components that made that increased photoprotection possible. Importantly, these polyploidy-derived traits that allow for changes in habitat are the emergent consequences of the induced changes to different “omes” and how they interact with the environment. These traits are not likely due to changes in the sequence or expression of a single gene or pair of homoeologs, but due to changes in a suite of genes and their interactions. The real goal of applying network analysis is to discover how all of these pieces interact in order to produce the interesting and unique phenotypes we see in polyploid plants.

Adding power to polyploid networks with prior knowledge and other data types

Polyploidization is accompanied by substantial rewiring of biological networks (De Smet & Van de Peer 2012). While gene co-expression networks can illuminate these changes, the quality of the reconstructed network and the resulting inferences depends both upon the data used to generate the network and the specifics of the biological questions being asked. When considering the underlying transcriptional correlates of a particular phenotypic difference (e.g. fruit size, drought tolerance) between two genotypes, ecotypes, or species, it is a common practice to conduct RNA-seq expression profiling experiments, and then subject the transcriptional data to differential expression analysis. Extending these analyses to co-expression networks can allow for further inference (see above; Figure 1); however, more extensive and thoughtful sampling and the integration of prior knowledge or other data types can increase the inferential power of these networks even further.

The nuances behind network analysis design have been extensively discussed (Albert 2005; Horvath 2009; Krouk *et al.* 2013; Mitra *et al.* 2013). In general terms, the power of co-
expression network analysis to provide insight depends largely on the question, on the extent of sampling, and on the amount of prior knowledge available (Krouk et al. 2013). A recent example from Helianthus exemplifies what can be learned from appropriately tailoring the experimental data to the biological question. Marchand et al. (2014) utilized gene expression data in H. annuus under nine hormonal treatments from seven time-points to build a gene regulatory network (GRN) for drought stress in sunflowers, with a focus on an informed set of candidate genes. From this analysis, they: (1) uncovered hub genes for the drought stress GRN; (2) discovered a role for nitrate transporters in regulating transpiration; and (3) connected the abscisic acid dependent and independent pathways. Through this targeted approach, Marchand et al. were able to infer causal relationships involved in transcriptional regulation, in addition to co-expression patterns. This perspective sets the stage for further analyses, such as evaluating gene differentiation among Helianthus species and cultivars with various adaptations to drought stress. This approach can be applied to a wide range of experiments utilizing gene expression networks.

Gene co-expression networks are relatively straightforward to generate, and with the addition of prior knowledge and proper sampling, have successfully identified gene-to-gene connections related to phenotype and function. These co-expression relationships, however, reflect the results of a series of direct molecular interactions, e.g., protein-protein interaction, protein-DNA interaction, membership in metabolic networks, and small RNA regulatory interactions. For example, consider a hypothetical protein-DNA interaction that results in joint up-regulation of a set of genes. This co-regulation may be due to the involvement of these genes in a multimeric complex, and hence direct co-regulation of these components by the protein-DNA interaction complex; alternatively, this interaction may only be made possible when the right multimeric complex is present, which itself promotes co-up-regulation of the whole
pathway. Supplementing gene co-expression networks with this kind of information leads to more informative network reconstructions better capable of discerning the underlying biological interactions (Li & Jackson 2015; however, see Bloom & Adami 2003 for an example of the caveats of integrating data sets). Protein-protein interaction networks have been used to consider the consequences of gene and genome duplications, most notably in *Arabidopsis thaliana*, where the effects of both small scale and whole genome duplications were evaluated during network construction (Arabidopsis Interactome Mapping Consortium 2011). The Arabidopsis Interactome Mapping Consortium found that shared protein-protein interactions between paralogs decreased rapidly for duplicates derived from both small-scale and genome-wide duplications, followed by a period of much slower decrease. This trend was mirrored by the divergence in protein sequence, suggesting that, following duplication, the rate of protein evolution and the maintenance of interacting partners are connected. Similarly, specific protein-DNA binding interactions may be assessed (e.g. ChIP-chip or ChIP-seq; Heyndrickx *et al.* 2014) and layered on gene co-expression networks (Angelini & Costa 2014). Other, less specific assays (e.g., DNase I hypersensitivity assay) can also provide information with respect to the protein-bound regions of the genome (e.g. Zhang *et al.* 2012), which can subsequently be used to inform more specific protein binding assays (Zhu *et al.* 2015). Biologically, these protein-DNA interactions may include different forms of transcriptional regulation, such as enrichment or depletion of histones or transcription factor binding sites; in the context of allopolyploidy, differences in protein-DNA interactions between duplicate genomes could contribute to novel phenotypes. Ha *et al.* (2011), for example, used ChIP-seq to show that homoeologous gene expression patterns were altered via histone modification differences in *Arabidopsis*. Similar experiments in other polyploid species could help determine the underpinnings of their unique phenotypes or adaptations.
Most sophisticated network analyses have been conducted in model organisms, such as yeast and human, where multiple, layered data types (e.g., RNA-seq plus epigenetic surveys, phenotypes, transcription factor binding sites, etc.) have led to increased understanding of gene regulatory network analyses (Bocklandt et al. 2011; Cookson et al. 2009; Gao et al. 2004; Langfelder et al. 2012; Madan Babu & Teichmann 2003; Mason et al. 2009; Zhu et al. 2008). In yeast, for example, integration of genotypic, expression, protein-protein interaction, and transcription factor binding data led to the development of an extremely well-refined and useful gene regulatory network (Zhu et al. 2008). Through integration of well-constructed gene co-expression networks with prior knowledge and these other forms of interaction data, we may best address questions central to understanding the ecological success of polyploid lineages and their specific adaptations.

**Empirical example from Gossypium**

Network analyses have clear advantages in model systems because of the availability of genomic and other “omic” resources and established experimental methods. In plants, the best developed models for most molecular biological experiments are *Arabidopsis*, rice, and several other species with well-developed transformation technologies. These species also have the distinct, bioinformatic advantage of diploidy; as with many other types of analyses, direct application of established protocols and pipelines to polyploid species presents complications arising from redundant genomes, particularly from a bioinformatics perspective. As these challenges become addressed (Duchemin et al. 2015; Mithani et al. 2013; Page et al. 2013a; Page et al. 2014; Page & Udall 2015), we can now envision experiments that will inform us about network processes in polyploid species.
As an example, we present here an application relevant to the question of the evolution of fiber development and morphology in polyploid *Gossypium hirsutum*, the species that provides most of the world’s cotton crop. Phenotypically, fibers from modern cultivated cotton are vastly improved relative to those of their wild progenitors (Fig. 2), having longer, stronger, and finer fiber as a consequence of several millennia of strong directional selection under domestication. Morphologically, “fibers” are single-celled epidermal trichomes arising from the surface of the ovule, already evident on the day the flower opens and thereafter growing rapidly over a period of a couple of months through the stages of primary and secondary wall synthesis, followed by maturation and programmed cell death (Haigler et al. 2012). In our earlier work, we showed that the transcriptome of these cells is highly dynamic, and that domestication has led to a massive transcriptomic and proteomic rewiring (Bao et al. 2011; Chaudhary et al. 2008; Hovav et al. 2008a; Hovav et al. 2008b; Hu et al. 2013; Hu et al. 2014; Rapp et al. 2010; Rapp et al. 2009). One insight emerging from this body of work is that the profilin gene family contains six members (PRF1-PRF6), five of which are upregulated in fibers early in development in modern cotton species relative to its wild progenitor (Bao et al. 2011). Moreover, this same upregulation was shown to have repeatedly been unknowingly selected under domestication in three independently domesticated cottons. Subsequent differential gene expression analysis of two fiber developmental stages in multiple wild and domesticated accessions suggested that PRF1, PRF3, and PRF4 may be key players in the difference between wild and domestication cotton fiber development (Yoo & Wendel 2014). As profilin plays an important role in actin polymerization, it is reasonable to propose that proteins relevant to cytoskeletal behavior were targeted by humans during domestication and crop improvement (Bao et al. 2011; Yoo & Wendel 2014). These observations raise many questions; for example, have the profilin genes
themselves been the targets of selection, or have their upstream regulators? Are all three important in specifying the domesticated fiber phenotype, or is it one gene that encodes the key player, with others passively co-upregulated? What other genes are up- or down-regulated, or coexpressed, during fiber development that may be related to PRF1, PRF3, and PRF4?

Both differential gene expression (DGE) and co-expression network analyses have the potential to increase our understanding of the role of profilin genes in fiber development, although from different angles. To illustrate this, we used RNA-seq data from an ongoing analysis of fiber development for both DGE and network reconstruction (Gallagher et al., unpublished). These data consist of four developmental stages (5, 10, 15, and 20 days post-anthesis, or flower opening, hereafter “dpa”) from three accessions each of wild and domesticated *G. hirsutum*. Standard DGE analysis (see Methods) of the data revealed 3,811 genes up-regulated under domestication during fiber development (~domestication + dpa, Domesticated>Wild, adjusted *P*-value < 0.05), including 167 transcription factors and 42 genes with known or suspected involvement in cytoskeleton and cell wall organization (Table 1). The numbers of up-regulated and down-regulated genes between developmental stages and between wild and domesticated cotton demonstrate several notable results, such as the observation of incredible transcriptomic stasis from 10 to 15 dpa in wild cotton and massive change during the transition to secondary wall synthesis in domesticated *G. hirsutum* between 15 and 20 dpa (Fig. 3; Bao *et al.* 2011; Rapp *et al.* 2010; Yoo & Wendel 2014). These results illustrate and confirm the massive scope of transcriptional rewiring that has accompanied the morphological transformations engendered by human selection on this single-celled structure. Of the profilins, we find PRF1 (Gorai.009G028500) is significantly up-regulated between domesticated and wild cottons at 10, 15, and 20 dpa (Fig. 4A), in contrast to the earlier quantitative PCR (QPCR) results where PRF1
to PRF5 were all up-regulated by domestication (Bao et al. 2011); however, this new result was similar to a previous RNA-seq study where only PRF1 was found to be significantly up-regulated by domestication at 10 and 20 dpa (Yoo & Wendel 2014), possibly indicative of the more sensitive nature of QPCR. Previous microarray results also indicate up-regulation of a profilin from 7 to 20 dpa (Rapp et al. 2010), which coincides with the up-regulation of one of the profilins (Gorai.003G061200) from 15 to 20 dpa in domesticated cotton (Fig. 4A). The differences in results of these several studies may be due to the greater variance in RNA-seq (versus QPCR or microarrays), as direct inspection of individual gene expression profiles based on normalized read counts suggests that at least half of the profilins are differentially expressed between wild and domesticated cotton while exhibiting diverse developmental programs (Fig. 4A). These expression patterns are highly consistent with Bao et al. (2011) and further our understanding of variation among profilins, both among gene family members and between wild and domesticated cotton fibers. Without prior interest in the profilin family, however, it is likely the differences in expression of these profilins, and their potential relevance to the cotton fiber developmental programme, would be lost in the morass of differentially expressed gene lists.

The same data were subjected to unsupervised network reconstruction via WGCNA, using default parameters (Langfelder & Horvath 2008). From this analysis, the whole fiber transcriptome was clustered into 48 co-expression modules, with the profilin genes located in five different modules. While similar results could be obtained through more complicated DGE analyses than employed above, network reconstruction (as used here) provides a relatively simple and readily accessible method for distilling complex information into partitioned sets of putatively functionally related genes. Although no expression changes of PRF2 were suggested by DGE results, PRF2 was clustered into the same module (ME) with PRF1 (ME5), which
represents genes continuously up-regulated during fiber development in domesticated cotton (Fig. 4B). While other modules containing profilin genes (ME1 with PRF5 and Gorai.001G025300, ME2 with PRF4, ME3 with PRF6 and ME12 with PRF3 and Gorai.003G061200; Fig 4B) may indicate coordinated expression changes by domestication limited to a certain developmental stage (such as ME1 and ME2 at 10dpa), we chose to focus on the module containing PRF1 and PRF2 to exemplify what can be gained through network analysis. The genes comprising this module number 1,508, less than half the number of up-regulated genes in domesticated versus wild cotton from the DGE gene set. Furthermore, only 1,062 are common between the two gene lists, winnowing the DGE candidate gene list by ~2800, but identifying an additional ~500 genes that may be related to PRF1 and its role in fiber development that are overlooked by standard DGE analysis. Included in these ~1500 genes are 71 transcription factors, as well as 16 genes with known or suspected involvement in fiber development (J. Jareczek, unpublished), including 11 that also are present in the DGE gene list.

In addition, we extracted putative protein-protein interactions of profilins from a cotton bioinformatics database GraP (Zhang et al. 2015a), and identified a phosphatidylinositol phosphate kinase 10 gene (Gorai.004G153600) that was both bioinformatically predicted to interact with PRF1 and was also present in the DGE and co-expression module gene lists. Given that protein kinases are key players in intracellular signal transduction, further investigation into this gene of interest could reveal other functional genes or pathways that act in coordination with profilin at important time points in fiber development.

We note that in the foregoing we considered the aggregate expression of both duplicated copies (homoeologs) of each profilin gene, in that G. hirsutum is an allopolyploid species. In the context of the present paper, we are especially interested in how network approaches can
illuminate processes in *polyploids*, which requires analyses wherein each duplicated copy of each gene is bioinformatically distinguished and transcriptionally measured. From the standpoint of understanding polyploidy, we might then ask if instances of homoeolog expression bias are randomly distributed among modules, or if instead they are coordinated at this level, which would be suggestive of selection for network coordination and optimization (Blanc & Wolfe 2004; Chang *et al.* 2010; Conant & Wolfe 2006). This phenomenon, termed “concerted divergence”, was seen by Blanc and Wolfe (2004) in *Arabidopsis thaliana*, where paralogs from an ancient genome duplication event were found to diverge in gene expression and form two parallel networks with other paralogs from the same event. To the extent that concerted divergence of homoeologs exists, it may indicate ecologically or evolutionarily interesting sub-or neo-functionalization of gene regulatory networks following genome merger and doubling. Therefore, evaluating homoeolog expression bias in the context of networks can provide insight into the underlying architecture of relevant phenotypes. For example, by overlaying relative homoeolog expression on our profilin example (Figure 2), we illustrate patterns of homoeolog expression bias for 5 dpa fiber from wild and domesticated cotton. A first observation is that there exists no detectable overall or module-wide bias in homoeolog usage for either wild or domesticated cotton in 5 dpa fiber. A second observation is that PRF1 homoeolog usage switches from A-homoeolog biased to D-homoeolog biased in wild versus domesticated cottons, respectively. Third, a similar trend is observed for several genes when looking at the top 10% of nearest neighbors (most closely connected genes) to profilin, i.e., homoeolog usage bias that was previously near-equal becomes biased toward the D-homoeolog under domestication. This is not true for all nearest-neighbors, as one gene coding for a putative LETM1-like protein (Gorai.002G121400) remained slightly A-biased and a second, a putative hydroxyacylglutathione
hydrolase gene (Gorai.005G073200), went from near-equal homoeolog usage to being slightly A-biased. At a broader scale, for those genes in the top 25% of nearest-neighbors to PRF1, the shift under domestication is more evenly spread between A-bias and D-bias. These observations are notable, as long, spinnable fibers are found only in the A-genome parent of the polyploidy species, indicating possible recruitment of the alternative homoeolog for certain pathway segments during the phenotypic transition from the coarse brown fiber of wild cotton to the fine white fiber of domesticated cotton; however, further data are required to evaluate this veracity of this speculation.

The above example is intended to illustrate some possibilities for simple co-expression network analyses in polyploid species, how they can reveal genomic interactions within a polyploid nucleus that would not be evident from more traditional DGE analyses, and in a manner potentially connected to phenotypes of ecological or evolutionary relevance. Additional analytical possibilities will emerge in the future, as the resources and tools are developed. For example, in the above we constructed co-expression networks based on few samples and aggregate expression of homoeologs. As such, we are limited to what essentially are aggregate co-expression relationships of both homoeologs simultaneously, and for only a single cell type (fibers). Extension to multiple tissues may reveal novel network interactions among loci/homoeologs that may be indicative of sub- or neofunctionalization. Furthermore, evidence from other types of analyses (e.g., ChIP-seq, methyl-seq, protein-protein interactions) can be integrated and layered in a network context (Gomez-Cabrero et al. 2014; Hawkins et al. 2010) to provide further understanding of the interactions among genes, their regulation, and their evolution in a duplicated context. Such integrative network analysis holds promise for increasing our understanding of the complex foundations of the novel and/or adaptive phenotypes and
ecological traits of polyploids (Bocklandt et al. 2011; Cookson et al. 2009; Gao et al. 2004; Langfelder et al. 2012; Madan Babu & Teichmann 2003; Mason et al. 2009; Zhu et al. 2008). The power of network-based analyses will become more fully realized as additional molecular relationships and interdependencies become layered on the expression data, such that a more complete understanding emerges into omics relationships and their connections to ecological and evolutionary traits of interest.

Conclusions

Despite substantial gains in understanding the immediate and long-term consequences of polyploidy, there remain many unanswered questions. Recent research in multiple polyploid systems has described patterns of genomic and transcriptomic change (as well as other measurable –omics) on different timescales. These patterns have illuminated many of the phenomena associated with genomic merger (hybridization) and doubling, as well as subsequent cladogenesis and diversification. An exciting prospect is that we are able to reveal the underpinnings of complex phenotypes and ecological traits. As the generation of multiple data types becomes more accessible, we have the opportunity to reveal multiple sub-cellular connections that previously were hidden from view, thereby facilitating our understanding of how omics changes manifest into evolutionarily and ecologically important traits. While massive, integrated datasets (e.g. Gerstein et al. 2012; Zhu et al. 2008) are necessary to truly understand the nuances of interplay amongst polyploid “omes”, the integration of multiple datasets from multiple perspectives will ultimately increase our understanding of the formation of novel and adaptive traits in polyploids, and their evolutionary and ecological significance.
Acknowledgements

The authors thank Josef Jareczek for assistance with identifying cell wall and cytoskeleton genes. The authors are supported by grants from the National Science Foundation and Cotton Incorporated. J. P. Gallagher is supported by National Science Foundation Graduate Research Fellowship DGE1247194.

Author Contributions

JPG, CEG, GH, and JFW conceived of the concept of the paper. GH and CEG performed example analysis. JPG, CEG, GH, and JFW wrote the paper.

Data Accessibility Statement

RNA-seq data are archived at Dryad DOI http://dx.doi.org/10.5061/dryad.256hn and will be made publicly available upon final analysis. Scripts for performing the example analysis will be made available at https://github.com/Wendellab/MolEcol2016.

Methods

RNA was collected from fibers as described previously (Yoo & Wendel 2014). RNA-sequencing was performed on Illumina Hi-Seq2500 at Iowa State University DNA Facility (http://www.dna.iastate.edu/). Reads were mapped to the Gossypium raimondii genome (Paterson et al. 2012) using GSNAP with SNP tolerant mapping (Page & Udall 2015; Wu & Nacu 2010). Differential gene expression analysis was conducted in R software v.3.2.0 (R Core Team 2015) with package DESeq2 (Love et al. 2014). For weighted gene co-expression network analysis, raw read count data were rlog (regularized logarithm built in DESeq2) transformed and then analyzed.
using the WGCNA package in R with default parameters (Langfelder & Horvath 2008, 2012). The resulting networks were visualized using Cytoscape v3.2.0 (Shannon et al. 2003).

References


Ha M, Ng DWK, Li WH, Chen ZJ (2011) Coordinated histone modifications are associated with gene expression variation within and between species. *Genome Research* **21**, 590-598.


**Tables and Figures**

**Table 1.** Gene families related to cytoskeleton and cell wall organization.

<table>
<thead>
<tr>
<th>Gene family</th>
<th>Exemplary References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actin</td>
<td>Bouget <em>et al.</em> (1996); Cleary and Smith (1998); Gallagher and Smith (1997); Vidali and Hepler (2001)</td>
</tr>
<tr>
<td>Actin-related proteins (ARPs)</td>
<td>Schafer and Schroer (1999)</td>
</tr>
<tr>
<td>Actin capping</td>
<td>Cooper and Schafer (2000); Hart <em>et al.</em> (2000)</td>
</tr>
<tr>
<td>Annexin</td>
<td>Morgan and Fernandez (1997); Morgan and Pilar Fernandez (1997)</td>
</tr>
<tr>
<td>Callose synthase</td>
<td>Irshad <em>et al.</em> (2008)</td>
</tr>
<tr>
<td>CESA</td>
<td>Arioli <em>et al.</em> (1998); Betancur <em>et al.</em> (2010)</td>
</tr>
<tr>
<td>CESA-like</td>
<td>Cutler and Somerville (1997)</td>
</tr>
<tr>
<td>COBRA-like</td>
<td>Roudier <em>et al.</em> (2002)</td>
</tr>
<tr>
<td>Cofilin/ADF</td>
<td>Carlier <em>et al.</em> (1997)</td>
</tr>
<tr>
<td>eEF1α</td>
<td>Sun <em>et al.</em> (1997)</td>
</tr>
<tr>
<td>Protein Type</td>
<td>Refs</td>
</tr>
<tr>
<td>------------------------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>Fimbrin/Plastin</td>
<td>Kovar et al. (2000); McCurdy and Kim (1998)</td>
</tr>
<tr>
<td>Formin</td>
<td>Deeks et al. (2010)</td>
</tr>
<tr>
<td>Fragmin</td>
<td>Furuhashi and Hatano (1989); Huang et al. (2011)</td>
</tr>
<tr>
<td>Gelsolin</td>
<td>Huang et al. (2011); Sun et al. (1999)</td>
</tr>
<tr>
<td>Glycosyl hydrolase</td>
<td>Irshad et al. (2008)</td>
</tr>
<tr>
<td>Kinesin</td>
<td>Oppenheimer et al. (1997)</td>
</tr>
<tr>
<td>Methylesterase inhibitor/pectinesterase</td>
<td>Irshad et al. (2008)</td>
</tr>
<tr>
<td>MS4A/Peroxidase</td>
<td>Irshad et al. (2008); Shigeto et al. (2013)</td>
</tr>
<tr>
<td>Myosin</td>
<td>Kinkema and Schiefelbein (1994); Kinkema et al. (1994); Yamamoto et al. (1995)</td>
</tr>
<tr>
<td>Profilin</td>
<td>Christensen et al. (1996); Huang et al. (1996)</td>
</tr>
<tr>
<td>Severin</td>
<td>Brown et al. (1982); Huang et al. (2011)</td>
</tr>
<tr>
<td>Tubulin</td>
<td>Amos (2000); Schroer (2001); Wick (2000)</td>
</tr>
<tr>
<td>Villin</td>
<td>Khurana et al. (2010); Tang et al. (2012)</td>
</tr>
<tr>
<td>Xyloglucan transferase</td>
<td>Irshad et al. (2008)</td>
</tr>
</tbody>
</table>
Figure 1. Differences between traditional gene expression analysis and a network-focused approach. Blue box A and yellow box B represent organisms with divergent phenotypes. From these two organisms, high-throughput gene expression data are generated, processed, mapped, and converted into counts per gene model (grey box). These data typically are subjected to differential expression analyses (left) and/or co-expression network construction (right). These two forms of analysis lead to complementary but different results: a list of up- and down-
regulated genes or a set of co-expression connections between genes, respectively (green boxes). Importantly, part of co-expression network construction statistically analyzes these connections and groups them into modules of significantly connected genes. From these two sets of results, further analyses often are performed. Differentially expressed genes often are subjected to cluster analysis, similar to co-expression analysis, or are used for gene set enrichment, such as analyzing the list for Gene Ontology term and KEGG pathway enrichment. The gene co-expression connection results typically are queried for gene set enrichment, specifically looking at enrichment within modules. In addition, other layers of connection information, such as protein-protein interaction data, may be stacked on top of the gene co-expression networks, allowing for further inferences of genes important for divergence of phenotypes. Of note, the differential expression data set may be one of these layers, providing information on expression modulation within the connected network.
**Figure 2.** Homoeolog expression bias for genes possibly co-regulated with profilin 1 in fiber from wild (left) and domesticated (right) cotton, *G. hirsutum*. Unsupervised network reconstruction via WGCNA of fiber RNA from four time-points (5, 10, 15, and 20 dpa) produces several modules, including one ("module 5") that demonstrates up-regulation in domesticated *G. hirsutum* (versus wild) for all four stages. This module contains a total of 1,508 genes, including two profilin genes, one of which (PRF1) has been implicated in fiber morphology differences between multiple wild and domesticated cotton species/accessions (Bao *et al.* 2011; Yoo & Wendel 2014). Homoeolog expression biases at 5 dpa were mapped (as an example) to a reduction of module 5 which contains only the top 10% (top) or top 25% (bottom) nearest-neighbors to profilin 1 (shown as a yellow circle) for wild (bottom left) and domesticated (bottom right) *G. hirsutum*. Blue circle borders indicate bias toward A-derived homoeologs, red circle borders indicate bias toward D-homoeologs, and the intensity of the color indicates the intensity of homoeolog expression bias for that gene. Identities of the genes in the top 10% of nearest-
neighbors and their putative annotations (Paterson et al. 2012) are shown. Representative images of cotton fiber are shown for the wild (far left) and domesticated (far right) cottons.

**Figure 3.** Results of differential gene expression analysis for four developmental stages (in days post-anthesis, “dpa”) of cotton fiber in representatives of wild and domesticated *Gossypium hirsutum*. Green and blue boxes represent wild (top) and domesticated (bottom) cotton gene expression at a given developmental stage, respectively. Shown above horizontal lines are the numbers of statistically significantly up-regulated genes in the developmental stages to their left and right, respectively. To the left and right of the vertical lines are shown, respectively, the number of statistically significantly up-regulated genes in the comparisons between the two cotton phenotypes. Comparisons including at least one significantly up-regulated profilin homolog are highlighted in red.
**Figure 4.** Patterns of expression for (A) profilin family members and (B) profilin-containing co-expression modules during cotton fiber development. (A) Individual gene expression profiles for eight profilin family members between wild (blue) and domesticated (orange) *G. hirsutum* during fiber development. Expression is displayed using rlog transformed read counts for each profilin gene, with the standard error among replicates shown. Each profilin is associated with an annotated gene (i.e., Gorai) from the published D-genome reference (Paterson *et al.* 2012). (B) Expression patterns for network modules (ME) containing one or more profilin family members (PRF): PRF1 and PRF2 in module ME5; PRF3 and Gorai.003G061200 in ME12; PRF5 and Gorai.001G025300 in ME1; PRF4 in ME2; and PRF6 in ME3. Within-module relative
expression, summarized and standardized as per Langfelder and Horvath (2007), is shown for each accession and fiber stage in all profilin-containing modules. For each module, the expression of all member genes was summarized into one representative expression profile, which is zero-centered and relative between accessions and among developmental stages.
CHAPTER 3. A NEW SPECIES OF COTTON FROM WAKE ATOLL, GOSSYPUM STEPHENSI (MALVACEAE)

Joseph P. Gallagher, Corrine E. Grover, Kristen Rex, Matthew Moran, and Jonathan F. Wendel

Abstract

Wake Atoll is an isolated chain of three islets located in the Western Pacific. Included in its endemic flora is a representative of the genus Gossypium colloquially referred to as Wake Island cotton. Stanley G. Stephens pointed out that “Wake Island cotton does not resemble closely either the Caribbean or other Pacific forms”. Taking into consideration morphological distinctions, the geographic isolation of Wake Atoll, and newly generated molecular data presented here, we conclude that the cottons of Wake Atoll do in fact represent a new species of Gossypium, here named Gossypium stephensii. This name is chosen to commemorate the eminent natural historian, evolutionary geneticist, and cotton biologist, S. G. Stephens.

Introduction

Wake Atoll is an isolated chain of three islets (Peale, Wake, and Wilkes) located in the Western Pacific approximately 3,500 kilometers west of Hawaii and 2,500 kilometers east of Guam (Fig. 1; Levenson 2008). The remoteness of this island group in conjunction with its diminutive land area constrains its floristic and faunistic diversity, but also provides the opportunity for evolutionary endemism. The floristic communities of Wake Atoll have been cataloged by several naturalists, including E. H. Bryan Jr. and F. R. Fosberg (Bryan 1942;
Fosberg 1959). Within the *Tournefortia* L. (Boraginaceae)-dominated forest found on the islets of the atoll, there is a representative of the genus *Gossypium* L. that has often been identified as *G. hirsutum* L., a member of the allopolyploid clade of cottons, and is colloquially referred to as Wake Island cotton (Fosberg 1959; Levenson 2008).

The Wake Atoll cotton specimens collected by Bryan and Fosberg were labeled *Gossypium hirsutum var. religiosum* (L.) G. Watt (Bryan 1942; Fosberg 1959). As Fryxell (1968) pointed out, the name *Gossypium religiosum* L. is “the most tortured of all the Linnean names in *Gossypium*” due to its complex history. While Linnaeus originally named *G. religiosum* in 1767, Watt (1907) reduced it to synonymy with *G. hirsutum*, after his examination of the original holotype. Fryxell (1968) notes that the holotype itself bears the annotation “in Indiis”, suggesting a specimen origin in either the West Indies or India. Thus, we conclude that this name has no bearing on the naming of the new entity described here.

Wake Atoll cotton was included in early biogeographic research on the genus. Stephens (1966) considered Wake Island cotton in his work on oceanic dispersal, treating it as a wild form of *G. hirsutum*. Notably, he pointed out that “Wake Island cotton does not resemble closely either the Caribbean or other Pacific forms,” noting its “sprawling shrub” growth habit, ”densely hairy” pubescence, and larger than average petal spot in comparison to the other Pacific cottons (Stephens 1966). With consideration of prevailing ocean currents in conjunction with seed buoyancy and salt-water survival tests, Stephens suggested that Wake Island cotton may have originated following dispersal from the western coast of Mexico (Stephens 1966).

In our studies of these older, as well as more recent collections made by members and support contractors of the United States Air Force, we validated the distinguishing characteristics of the Wake Atoll forms relative to *G. hirsutum* (Fryxell 1992). We further assessed its
distinctiveness and taxonomic status through the use of extensive DNA sequence data from both the nuclear and chloroplast genomes, which reveal a clear differentiation of the Wake Atoll cottons from a large sampling of *G. hirsutum* accessions. Taking into consideration morphological distinctions, the geographic isolation of Wake Atoll, and newly generated molecular data presented here, we conclude that the cottons of Wake Atoll do in fact represent a new species of *Gossypium*, here named *Gossypium stephensii* J. Gallagher, C. Grover, & Wendel. This commemorative name honors the late S. G. Stephens (1911 – 1986), an underappreciated and remarkably insightful natural historian, evolutionary geneticist, and cotton biologist (Wendel and Goodman 2011).

**Materials and Methods**

**Plant materials**

Seeds collected on Wake Atoll by Ray Fosberg (formerly USGS, Washington, D.C.) were obtained from Paul A. Fryxell (formerly University of Texas, Austin) by J. Wendel. Additional collections were made by M. Moran and K. Rex. Plants were grown in the R.W. Pohl Conservatory at Iowa State University from these original seed collections. For comparative studies, we also sampled widely from the six other allopolyploid (AD genome) species of *Gossypium* (Grover et al. 2015), as well as from representative diploids that serve as models of the ancestral A- and D-genome donors (Table 2; see Wendel and Grover 2015). DNA was extracted from leaves of each accession using the Qiagen DNeasy Plant Mini Kit (Qiagen, Valencia, California), as per manufacturer instructions.
**Sequence capture and sequencing**

Sequencing libraries were constructed at West Virginia University Genomics Core Facility (http://genomics.as.wvu.edu/) using the Illumina TruSeq kit. Following library construction, DNA libraries underwent targeted sequence capture with custom bait sequence using the Mycroarray MYbaits kit (http://www.mycroarray.com/), as per the manufacturer’s protocol. Our custom bait sequence pool was designed to enrich 267 genes, as detailed in Table 3. Target-enriched libraries were then sent to either the Iowa State DNA Facility (http://www.dna.iastate.edu/) or the Beijing Genomics Institute (http://www.genomics.cn/en/index), where they were sequenced (150-bp paired end) on an Illumina HiSeq 2500.

**Sequence assembly**

Reads were quality trimmed using sickle (Joshi and Fass 2011) and were mapped to the *G. raimondii* D-genome reference sequence (Paterson et al. 2012) for nuclear genes, or to the *G. hirsutum* chloroplast genome (Lee et al. 2006) for chloroplast genes, using GSNAP (Wu and Nacu 2010). Nuclear genes were mapped in combination with a diagnostic SNP-index (Page et al. 2013) and partitioned into A-genome and D-genome derived reads via PolyCat (Page et al. 2013). This process partitions reads into the two homoeologs expected for each gene in allopolyploid genomes. We treated *G. stephensii* as an allopolypoid due to its recognized morphological similarity, as well as previous molecular treatments of the plants (e.g. Grover et al. 2012). Both the chloroplast reads and the partitioned, mapped nuclear gene reads were assembled into individual sequences using bam2consensus from the BamBam suite of tools (Page et al. 2014). The resulting gene alignments were iteratively processed to remove uncaptured sequences/gene regions (process_alignments;
https://github.com/Wendellab/phylogenetics), using a threshold of >80% ambiguous nucleotides (N) per sequence or >20% Ns per position for the first round and >50% Ns per sequence (30% for chloroplast) or >10% Ns per position for the second round. Any final gene or chloroplast alignment that did not contain at least one representative of each polyploid species, as well as a minimum of four Wake Atoll representatives, was excluded from the dataset. This quality filtering process resulted in the final inclusion of: (1) 102,227 bases of sequence from the chloroplast genome, roughly 65% of the total chloroplast genome [as compared to Lee et al. (2006)]; and (2) 474 genic regions representing 234 homoeologs from the A-genome and 240 homoeologs from the D-genome.

**Phylogenetic analysis**

A maternal phylogeny was generated from the cpDNA sequences from a maximum likelihood analysis using RAxML (Stamatakis 2014; 2015). Parameters selected included rapid bootstrap analysis, GTR + Γ model of evolution, and 1,000 alternative starting trees. Trees were visualized using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

For the nuclear data, alignments were filtered for PCoA analysis by first removing genes that did not contain all accessions sequenced, concatenated using SequenceMatrix (Vaidya et al. 2011), and subsequently removing all ambiguous sites. These concatenated sequences were then input into R for principal coordinates analysis (PCoA) using the ape package (Paradis et al. 2004; R Core Team 2015). Scripts for implementing these analyses are available on https://github.com/Wendellab/phylogenetics.
Results

Chloroplast phylogenetics

Chloroplast DNA sequences were assembled for representatives of all polyploid species of *Gossypium* (Table 2), including five representatives of *G. stephensii* and 34 representatives of *G. hirsutum*, the species to which *G. stephensii* was formerly thought to belong. A total of 102,227 aligned positions, representing ~65% of the *Gossypium* chloroplast genome, relative to *G. hirsutum* (Lee et al. 2006), were recovered and analyzed with maximum likelihood via RAxML and a GTR + γ model (Stamatakis 2014; 2015). All accessions of *G. stephensii* were recovered as a monophyletic group that is distinct from both *G. hirsutum* and *G. ekmanianum* Wittm. (Fig. 2). The overall topology of the tree is consistent with previous reports (Grover et al. 2012; Grover et al. 2015), with the *G. stephensii* clade phylogenetically sister to the *G. ekmanianum* clade. Strong bootstrap values (>90%) support almost all the species clades. The one exception is the divergence between *G. ekmanianum* and *G. stephensii*; each clade has 44% and 16% bootstrap support, respectively (Fig. 2).

Analysis of nuclear gene sequences

Because of the relatively recent diversification of the allopolyploid *Gossypium* clade and the attendant lack of species-level coalescence for many nuclear genes, phylogenetic analysis of different nuclear gene sequences yields multiple topologies (as described in Grover et al. 2015). Hence, to visually depict overall genomic relationships relevant to the taxonomic status of *G. stephensii*, we used a phenetic approach. Multiple accessions of *G. hirsutum* and *G. barbadense* L., as well as representatives from all other polyploid species (*G. tomentosum* Nutt. ex Seem., *G. mustelinum* Miers ex G. Watt, *G. darwinii* G. Watt, *G. ekmanianum*, *G. stephensii*; Table 2), were evaluated via PCoA of multilocus data (Fig. 3). Specifically, we included 474 genic
regions, partitioned into 234 homoeologs from the A-genome (Fig. 3A) and 240 homoeologs from the D-genome (Fig. 3B), representing the two co-resident genomes of allopolyploid cotton (reviewed in Wendel and Grover 2015). Both analyses revealed similar overall depictions of genomic distinctions. Notably, the primary distinction for both sets of homoeologs along the first axis is the separation of the *G. hirsutum-G. ekmanianum-G. stephensii* grouping from the *G. barbadense-G. darwinii* species pair, with the representatives of *G. mustelinum* and *G. tomentosum* between the two. Along the second axis, which shows differentiation within the groups specified above, *G. stephensii* forms a tight cluster that is distinct from other species, and most importantly, it is clearly distinguished from a broad spectrum of *G. hirsutum* accessions spanning the full wild to domesticated continuum.

Taxonomic Treatment


*G. stephensii* differs from *G. ekmanianum* in having dense leaf pubescence, 8-11 bracteal teeth, subglobose capsules, and 3-4 locules in the capsule (Table 1). Although many characteristics of *G. stephensii* overlap with the species concept of *G. hirsutum*, it differs from *G. hirsutum* in lacking extrafloral nectaries at anthesis; these appear during capsule development (Table 1).

Sprawling woody shrubs, branching just above the base, with branches growing at nodes along woody stems; stems pubescent with stellate hairs; punctae black (lysigenous cavities;
colloquially “gossypol glands”), abundant, more apparent in younger than in older stems. Leaves alternate, ascending, densely pubescent; stipules 6-10 mm; lamina cordate, weakly 3-lobed, rarely 5-lobed, broadly triangular to ovate, 65-75 mm long, 78-95 mm wide; apex acute to acuminate; margin entire, with stellate trichomes; adaxial and abaxial surfaces both with mostly stellate trichomes, pubescence denser on abaxial surface; 5 major veins abaxially raised; foliar nectaries on central midvein, 1.5 mm long, ~3.5 mm from base; petioles 50-53 mm long on mature leaves. Epicalyx 26-32 mm long, 26-28 mm wide, reflexed, dissected with 8-11 teeth, with numerous punctae throughout, diminishing toward the distal edge, hairs sparse, nectaries absent. Sepals 5-7 mm long, basally connate into a cup, lobes 1-2 mm long, acuminate, sinuses rounded, trichomes sparse to wanting, black punctae diminishing distally. Petals white to cream, with faint red to fuchsia basal spots, 39-42 mm long, petal spots 8-11 mm long. Staminal column white; filaments 6-10 mm; pollen white to cream-colored. Style 18-20 mm long, exserted 11-12 mm beyond staminal column, clavate, three ridged, punctate throughout; stigma decurrent on style. Capsules subglobose, apiculate, 18-20 mm in diam., with subtending epicalyx nectaries, locules 3-4, with 2-4 seeds per locule, capsule walls punctate. Seeds green, lanate, seed fibers tan, 10-12 mm long. (Fig. 4).

**Comments**

Because of the extensive and partially overlapping morphological variability between *G. stephensii*, *G. hirsutum*, and *G. ekmanianum*, DNA sequence data may be helpful for species confirmation. Within the chloroplast alignment, nucleotide changes specific to *G. stephensii*, as compared to the other allopolyploid species, appear at site 5,562 (A to T), within the intron of *rps16*; site 33,766 (T to A), in the intergenic space between *trnT-GGU* and *psbD*; and site
130,280 (A to C), in the intergenic space between trnL-UAG and rpl32 (sites based on G. hirsutum chloroplast sequence; Lee et al. 2006).

**Distribution and habitat**

*Gossypium stephensii* is found exclusively on the islets that make up Wake Atoll. Wake Atoll has a tropical maritime climate and is subject to trade winds and occasional typhoons. These plants occur in what has been described as a sparse *Tournefortia argentea* L. f.-dominated forest. Cotton plants make up one of the more prevalent floristic elements, along with *Ipomoea pes-caprae* (L.) R. Br. These forests are found from the beaches up to the inland portion of the islands (Levenson 2008).

**Etymology**

The specific epithet honors the late Stanley George Stephens, paying tribute to his numerous contributions to our understanding of *Gossypium* diversity and evolution (Wendel and Goodman 2011). Stephens was a researcher at the Cotton Research Station in Trinidad and at North Carolina State University, where his many insightful studies of *Gossypium* evolution, systematics, and genetics resulted in his election into the National Academy of Sciences of the U. S. A. We name this species *Gossypium stephensii* in his memory.

**Paratypes**

U. S. A., Iowa: R. W. Pohl Conservatory, Bessey Hall, Iowa State University, grown from seed collected on Wilkes Island, received by J. F. Wendel from P. A. Fryxell in 1988. 5 May 2015, *Gallagher 1* (ISC); R. W. Pohl Conservatory, Bessey Hall, Iowa State University, grown from seed collected on Wilkes Island, received by J. F. Wendel from P. A. Fryxell in 1988. 31 October 2015, *Gallagher 3* (ISC); R. W. Pohl Conservatory, Bessey Hall, Iowa State University, grown from seed collected on Peale Island, Wake Atoll, N 19° 18’ 33.2” E 166° 37’
43.6”, collected by Matt Moran in August 2012. 31 October 2015, *Gallagher 4* (ISC); R. W. Pohl Conservatory, Bessey Hall, Iowa State University, grown from seed collected on Peale Island, Wake Atoll, by Kristen Rex in March or April 2013. 31 October 2015, *Gallagher 5* (ISC); R. W. Pohl Conservatory, Bessey Hall, Iowa State University, grown from seed collected on Peale Island, Wake Atoll, by Kristen Rex in March or April 2013. 31 October 2015, *Gallagher 6* (ISC); R. W. Pohl Conservatory, Bessey Hall, Iowa State University, grown from seed collected on Wake Island, Wake Atoll, near Wake Marina Causeway on way to Wilkes Island, by Kristen Rex in March or April. 31 October 2015, *Gallagher 7* (ISC).

Acknowledgements

We thank Kara Grupp, Anna Tuchin, and Lori Schmidt for assistance in collecting plant materials for this project, and Paul A Fryxell for providing seeds. This work is funded by the National Science Foundation Plant Genome Program and by Cotton Incorporated. Joseph P. Gallagher is supported by NSF Graduate Research Fellowship DGE1247194.
References


Watt, G. 1907. The wild and cultivated cotton plants of the world.


Table 1. Species diagnostic traits for *G. stephensii* and its most closely related species, *G. hirsutum* L. (based on Fryxell 1992) and *G. ekmanianum* Wittm. (based on Krapovickas and Seijo 2008).

<table>
<thead>
<tr>
<th>Feature</th>
<th><em>G. stephensii</em></th>
<th><em>G. hirsutum</em></th>
<th><em>G. ekmanianum</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Growth habit</td>
<td>Sprawling shrub</td>
<td>Wide branching shrub</td>
<td>Sprawling shrub</td>
</tr>
<tr>
<td>Leaf pubscense</td>
<td>Dense</td>
<td>Dense to glabrous</td>
<td>Scattered</td>
</tr>
<tr>
<td>Floral nectaries</td>
<td>Absent</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>Bracteal Teeth</td>
<td>8-11</td>
<td>3-19</td>
<td>3-9</td>
</tr>
<tr>
<td>Petal size</td>
<td>40 mm</td>
<td>20-50 mm</td>
<td>40 mm</td>
</tr>
<tr>
<td>Capsule shape</td>
<td>Subglobose</td>
<td>Broadly ovoid to subglobose</td>
<td>Broadly ovoid</td>
</tr>
<tr>
<td>Number of locules</td>
<td>3-4</td>
<td>3-5</td>
<td>3</td>
</tr>
<tr>
<td>Locality</td>
<td>Wake Atoll</td>
<td>Caribbean, Central America, South Pacific</td>
<td>Dominican Republic</td>
</tr>
</tbody>
</table>
Table 2. Alphabetical list of species considered in chloroplast phylogeny and principal coordinates analysis: species name, alphabetical list of accessions examined, asterisk if included in chloroplast phylogeny, voucher number, SRA accession number. All herbarium specimens are maintained at ISC, with the exceptions of GB0262, GB0287, GB0369, GB0319, LKT511, and TAMCOT, which perished prior to vouchering.

<table>
<thead>
<tr>
<th>Species Name</th>
<th>Accession Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gossypium arboreum L.</td>
<td>A2-101*, Gallagher 14, SRR5001776</td>
</tr>
<tr>
<td>Gossypium barbadense L.</td>
<td>SRR5001784, GB0262*, no voucher, SRR5001789, GB0287*, no voucher, SRR5001786, GB0303*, Gallagher 19, SRR5001784, GB0369*, no voucher, SRR5001799, GPS52*, Gallagher 18, SRR5001801, K101*, Gallagher 17, SRR5001832, Phy76*, Gallagher 16, SRR5001793, Pima S6*, Gallagher 15, SRR5001811, Gossypium darwinii G. Watt, lab accession*, Gallagher 22, SRR5001813, Gossypium ekmanianum Wittm., TX2263*, Gallagher 23, SRR5001816, TX2265*, Gallagher 24, SRR5001825, TX2266*, Gallagher 25, SRR5001814, TX2271*, Gallagher 26, SRR5001794, TX2273*, Gallagher 27, SRR5001778, Gossypium herbaceum L., A1-Wagad*, SENCHINA 312, SRR5001775, Gossypium hirsutum L., ARK2402*, Gallagher 12, SRR5001770, Cascot L7*, Gallagher 9, SRR5001823, Coker 315*, Gallagher 13, SRR5001827, CRB 252*, Krush and Grupp s.n. (ISC450783), SRR5001826, FM 958* Krush and Grupp s.n. (ISC450786), SRR5001773, GB0319*, no voucher, SRR5001817, LKT511*, no voucher, SRR5001818, Maxxa*, Gallagher 8, SRR5001774, PM145*, Krush and Grupp s.n. (ISC448082), SRR5001783, TAMCOT*, no voucher, SRR5001785, TM1*, Gallagher 10, SRR5001831, TX44*, Schmidt s.n. (ISC446804), SRR5001802, TX480*, Schmidt s.n. (ISC446777), SRR5001788, TX665*, Schmidt s.n. (ISC446834), SRR5001822, TX672*, Schmidt s.n. (ISC446832), SRR5001792, TX786*, Schmidt s.n. (ISC446829), SRR5001803, TX1009, Schmidt s.n. (ISC446783), SRR5001790, TX1037*, Schmidt s.n. (ISC446782), TX1037*.</td>
</tr>
</tbody>
</table>
SRR5001815, TX1046*, Schmidt s.n. (ISC446820), SRR5001805, TX1055*, Krush and Grupp s.n. (ISC448110), SRR5001780, TX1107*, Krush and Grupp s.n. (ISC448120), SRR5001791, TX1110*, Krush and Grupp s.n. (ISC450784), SRR5001798, TX1120* Krush and Grupp s.n. (ISC448122), SRR5001812, TX1182*, Krush and Grupp s.n. (ISC448067), SRR5001809 and SRR5001769, TX1226*, Krush and Grupp s.n. (ISC448091), SRR5001771, TX1228*, Krush and Grupp s.n. (ISC448116), SRR5001796, TX1236*, Krush and Grupp s.n. (ISC448043), SRR5001810, TX1748*, Krush and Grupp s.n. (ISC447904), SRR5001795, TX1982*, Krush and Grupp s.n. (ISC447907), SRR5001829, TX1988*, Krush and Grupp s.n. (ISC448052), SRR5001797, TX1996*, Krush and Grupp s.n. (ISC448073), SRR5001807, TX2002*, Krush and Grupp s.n. (ISC450774), SRR5001830, TX2089*, Krush and Grupp s.n. (ISC447886), SRR5001800, TX2090*, Gallagher 11, SRR5001821, TX2091*, Krush and Grupp s.n. (ISC448053), SRR5001772, TX2092*, Krush and Grupp s.n. (ISC448055), SRR5001828, TX2094, Krush and Grupp s.n. (ISC448044), SRR5001804, TX2095*, Krush and Grupp s.n. (ISC447885), SRR5001787, Gossypium mustelinum Miers ex G. Watt, Lab accession*, Gallagher 21, SRR5001820, Gossypium raimondii Ulbr., JFW*, Gallagher 28, SRR5001779, Gossypium stephensii J. Gallagher, C. Grover, & Wendel, Peale1B*, Gallagher 4, SRR5001824, PealeKR1*, Gallagher 5, SRR5001782, PealeKR4*, Gallagher 6, SRR5001806, Wake Causeway 1*, Gallagher 7, SRR5001808, Wilkes A*, Gallagher 3, SRR5001781, Wilkes B*, Gallagher 1 and Gallagher 2, SRR5001777, Gossypium tomentosum Nutt. ex Seem., AD3-95*, Gallagher 20, SRR5001833
Table 3. List of target genes for sequence capture. Identifiers based on the *Gossypium raimondii* reference genome (Paterson et al. 2012).

<table>
<thead>
<tr>
<th>Gene Identifier</th>
<th>Gene Identifier</th>
<th>Gene Identifier</th>
<th>Gene Identifier</th>
<th>Gene Identifier</th>
<th>Gene Identifier</th>
<th>Gene Identifier</th>
<th>Gene Identifier</th>
<th>Gene Identifier</th>
<th>Gene Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gorai.001G014000</td>
<td>Gorai.001G015200</td>
<td>Gorai.001G015300</td>
<td>Gorai.001G020400</td>
<td>Gorai.001G020500</td>
<td>Gorai.001G020600</td>
<td>Gorai.001G053900</td>
<td>Gorai.001G059900</td>
<td>Gorai.001G075800</td>
<td>Gorai.001G077600</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gorai.001G087200</td>
<td>Gorai.001G087400</td>
<td>Gorai.001G089400</td>
<td>Gorai.001G090000</td>
<td>Gorai.001G126200</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gorai.001G134900</td>
<td>Gorai.001G172200</td>
<td>Gorai.001G173700</td>
<td>Gorai.001G173900</td>
<td>Gorai.001G174000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gorai.002G268600</td>
<td>Gorai.003G007000</td>
<td>Gorai.003G011500</td>
<td>Gorai.003G011500</td>
<td>Gorai.003G021100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gorai.003G051800</td>
<td>Gorai.003G061700</td>
<td>Gorai.003G098300</td>
<td>Gorai.003G103000</td>
<td>Gorai.003G105200</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gorai.003G112000</td>
<td>Gorai.003G118000</td>
<td>Gorai.003G139800</td>
<td>Gorai.003G163300</td>
<td>Gorai.003G175400</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gorai.003G186300</td>
<td>Gorai.004G002100</td>
<td>Gorai.004G026700</td>
<td>Gorai.004G027600</td>
<td>Gorai.004G044000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gorai.004G079500</td>
<td>Gorai.004G088000</td>
<td>Gorai.004G095400</td>
<td>Gorai.004G101100</td>
<td>Gorai.004G117700</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gorai.004G120400</td>
<td>Gorai.004G125200</td>
<td>Gorai.004G130900</td>
<td>Gorai.004G141400</td>
<td>Gorai.004G150800</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gorai.004G151000</td>
<td>Gorai.004G155700</td>
<td>Gorai.004G169000</td>
<td>Gorai.004G196800</td>
<td>Gorai.004G205900</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gorai.004G212200</td>
<td>Gorai.004G247700</td>
<td>Gorai.004G250800</td>
<td>Gorai.004G264200</td>
<td>Gorai.004G264600</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gorai.004G280500</td>
<td>Gorai.004G280800</td>
<td>Gorai.004G280500</td>
<td>Gorai.004G280800</td>
<td>Gorai.004G280800</td>
</tr>
</tbody>
</table>
Gorai.009G093700, Gorai.009G100600, Gorai.009G104200, Gorai.009G146500,
Gorai.009G167800, Gorai.009G174600, Gorai.009G175500, Gorai.009G181300,
Gorai.009G185200, Gorai.009G200600, Gorai.009G213800, Gorai.009G235000,
Gorai.009G236700, Gorai.009G237900, Gorai.009G243600, Gorai.009G250900,
Gorai.009G258800, Gorai.009G268600, Gorai.009G271100, Gorai.009G276600,
Gorai.009G280900, Gorai.009G302500, Gorai.009G313100, Gorai.009G339300,
Gorai.009G403800, Gorai.009G439200, Gorai.010G008900, Gorai.010G023800,
Gorai.010G035000, Gorai.010G063400, Gorai.010G081000, Gorai.010G085400,
Gorai.010G087200, Gorai.010G123400, Gorai.010G149300, Gorai.010G183600,
Gorai.010G230100, Gorai.010G230200, Gorai.010G230400, Gorai.010G231800,
Gorai.010G242800, Gorai.010G243100, Gorai.011G011200, Gorai.011G040900,
Gorai.011G083000, Gorai.011G085100, Gorai.011G106100, Gorai.011G106200,
Gorai.011G139200, Gorai.011G155800, Gorai.011G161200, Gorai.011G161300,
Gorai.011G188900, Gorai.011G189300, Gorai.011G190600, Gorai.011G200200,
Gorai.011G201900, Gorai.011G212900, Gorai.011G213900, Gorai.011G241700,
Gorai.011G279800, Gorai.012G001200, Gorai.012G009300, Gorai.012G009400,
Gorai.012G009500, Gorai.012G026100, Gorai.012G026200, Gorai.012G075600,
Gorai.012G112000, Gorai.012G149800, Gorai.012G150900, Gorai.012G175500,
Gorai.012G176200, Gorai.013G019500, Gorai.013G023400, Gorai.013G039800,
Gorai.013G045300, Gorai.013G055700, Gorai.013G080300, Gorai.013G096100,
Gorai.013G112000, Gorai.013G114800, Gorai.013G114900, Gorai.013G118300,
Gorai.013G121000, Gorai.013G141500, Gorai.013G163900, Gorai.013G202800,
Gorai.013G203900, Gorai.013G224300, Gorai.N007500
Figure 1. Map of Wake Atoll and its location within the Pacific Ocean, denoted by the black circle. A is Wilkes Island, B is Wake Island, and C is Peale Island.
Figure 2. Maximum likelihood tree based on analysis of maternally inherited chloroplast DNA sequences. The Gossypium stephensii clade is phylogenetically distinct from the other allopolyplloid species, including G. ekmanianum and G. hirsutum. Node values are bootstrap support percentages.
Figure 3. Principal coordinate analysis based on nuclear gene sequences from relevant allopolypld and diploid *Gossypium* species. Separate analyses were conducted for homoeologs from the two co-resident genomes (A-, D-) in allopolypld cottons. (A) Axes 1 and 2 for the 237 concatenated A-genome gene sequences. (B) Axes 1 and 2 for the 240 concatenated D-genome gene sequences. Key to species: AD1 = *G. hirsutum*, AD2 = *G. barbadense*, AD3 = *G. tomentosum*, AD4 = *G. mustelinum*, AD5 = *G. darwinii*, AD6 = *G. ekmanianum*, AD7 = *G. stephensii*, D5 = *G. raimondii*. 
Figure 4. Holotype specimen of *Gossypium stephensii* J. Gallagher, C. Grover & Wendel.
Figure 5. Illustrative morphological features of *Gossypium stephensii*: A. The sprawling growth habit of *G. stephensii* in its native habitat in Wake Atoll. B. Flower. C. Immature capsule. D. Dehisced capsule and seed, showing light reddish-brown fibers. E. Comparison of seed fibers from *G. hirsutum* cv. Maxxa, a domesticated line of Upland cotton (left), *G. stephensii* (center), and *G. hirsutum* TX2094, a wild form of *G. hirsutum* var. *yucatanense* (right).
CHAPTER 4. EFFECTS OF GENOME DUPLICATION AND DOMESTICATION ON THE FIBER GENE CO-EXPRESSION NETWORK OF *GOSSYPIUM HIRSUTUM* L.

A paper in preparation for *Genome Biology and Evolution*

Joseph P. Gallagher, Guanjing Hu, Corrinne E. Grover, Josef J. Jareczek and Jonathan F. Wendel

Abstract

*Gossypium hirsutum*, Upland cotton, has undergone hybridization, polyploidization, and subsequent domestication. Here we analyze the fiber from three wild and three domesticated accessions during fiber development. Using gene co-expression network analysis, we compare gene pair and homoeolog-specific networks. Modules from both networks are enriched for similar biological processes, including cell wall biosynthesis and elongation. The majority of modules generated in the aggregate data set are preserved in the partitioned data set. However, only one quarter of homoeolog pairs are found in the same module in the partitioned gene co-expression network. By constructing networks from the wild and domesticated accessions separately, we find that the modules in wild cotton are generally poorly preserved in domesticated cotton. Homoeolog bias generally favors the D-subgenome, but specific modules show A-subgenome bias. Using a combination of differential co-expression and differential expression, we identify a group of genes that has been highly affected by domestication. Overall, we conclude that gene co-expression dynamics are strongly altered by domestication.
Introduction

Cotton is the primary fiber crop grown throughout the globe. The species *Gossypium hirsutum* L., or Upland cotton, makes up more than 90% of production worldwide. It is a member of the polyploid clade of New World cottons, which formed *ca.* 1-2 million years ago when a member of the African, A-genome diploid cottons made a transoceanic voyage to the Americas and then hybridized with a native, D-genome diploid cotton (Wendel and Grover 2015b). This remarkable hybridization and genome doubling event led to the subsequent evolution of the polyploid (AD-genome) clade of *Gossypium*, which includes seven species (Krapovickas and Seijo 2008; Grover et al. 2015; Gallagher et al. 2017). *Gossypium hirsutum*, which in the wild is native to coastal Yucatan, Mexico and more sparsely elsewhere in nearby regions (extending as far north as the Florida Keys), was domesticated approximately 5000 years ago (Wendel et al. 1992; Brubaker and Wendel 1994; d'Eeckenbrugge and Lacape 2014). Following initial domestication, likely in the Yucatan Peninsula, cotton spread rapidly throughout central America, where semi-domesticated or “door-yard” forms are still found today. In the last several hundred years, strong directional selection by humans led to the modern annual forms of highly improved Upland cotton, resulting in the elite cultivars grown globally today (Figure 1).

Single-celled cotton fibers develop to maturity over approximately a fifty day period, beginning with fiber initials on the surface of the ovules, which already are differentiated to become fibers on the day of anthesis (Haigler et al. 2012). Fiber elongation and primary cell wall synthesis are followed by secondary cell wall synthesis (Applequist et al. 2001; Haigler et al. 2012). Between one half and one third of genes in the genome are expressed in cotton fiber during fiber development (Yoo and Wendel 2014; Tuttle et al. 2015). Previous studies have
shown that gene expression profiles during fiber development in *G. hirsutum* are drastically different between wild and domesticated accessions (Rapp et al. 2010; Yoo and Wendel 2014). Yoo and Wendel (2014), using RNA-seq from several wild and domesticated accessions, found that approximately 5000 genes were differentially expressed between wild and domesticated fiber while looking at just two time points in fiber development. They also confirmed that specific classes of genes are differentially expressed during fiber development, as described previously (Hovav et al. 2008a; Rapp et al. 2010; Bao et al. 2011; Yoo and Wendel 2014). In summary, changes in gene expression wrought by domestication, are remarkable in their genomic scope, perhaps even doubly so because they concern only a single cell type.

In addition to domestication, polyploidy itself likely has had pervasive effects on gene expression during fiber development. Following the hybridization and genome duplication event 1-2 Mya, the duplicated genes in the polyploids, which evolved in diploid ancestors over a 5-10 million year period on different continents (Wendel and Grover 2015a), became reunited in a common *trans* regulatory environment but with their more or less diverged *cis* controls. Prior to the formation of the polyploid clade of cotton, the A-genome cottons were the only ones to possess long spinnable fiber. As such, it is expected that the A-subgenome of *G. hirsutum* would carry all the information necessary for long fiber generation. However, previous studies have found that domestication has also acted on gene expression derived from the D genome; Hovav et al. (2008a) showed that the ratio of homoeolog expression between A- and D-subgenomes was skewed toward the D-subgenome following domestication in over a fifth of the genes expressed during fiber development. Other studies have found balanced genome-wide homoeolog expression bias, although they still reported gene-by-gene A-subgenome and D-subgenome biases (Yoo and Wendel 2014; Zhang et al. 2015; Fang et al. 2017). In general, these expression
alterations following polyploidy are thought to underlie the transgressive and enhanced properties of cultivated allopolyploid cotton relative to its diploid ancestors.

Differential gene expression has been a powerful tool to identify genes of interest in domestication, allowing for direct comparison between wild and domesticated plants, and for exploring polyploid transcriptomics. Gene co-expression network analysis methods can also be leveraged to learn new information about domestication and the effects of polyploidy on the transcriptome. These methods allow for the construction of gene co-expression networks and identification of modules of shared co-expression patterns; they also provide a framework for testing the preservation of these co-expression patterns between wild and domesticated accessions and between diploid and polyploid species. (Gallagher et al. 2016). These methods have already been useful in identifying domestication-related genes in tomato, wheat, maize, and cotton (Swanson-Wagner et al. 2012; Ichihashi et al. 2014; Pfeifer et al. 2014; Hu et al. 2016).

Here, we confirm that gene expression dynamics have been drastically altered by domestication with over 2,000 genes being differentially expressed between wild and domesticated accessions. Weighted gene co-expression analysis shows scale free topology that results in 26 modules in the duplicate gene network, and 52 modules when the network is partitioned into homoeologous contributions to aggregate expression. Genes from significant modules appear to be enriched for cell wall biosynthesis and elongation functions. Comparing wild and domesticated fiber, gene co-expression networks primarily show loss of preservation; some modules in the gene co-expression network have been rewired as a consequence of domestication. Focusing on those genes that are both differentially expressed and differentially coexpressed, we identify biologically relevant functions and interactions.
Materials and Methods

Plant materials, mRNA sequencing and mapping

Three wild and three domesticated accessions of *Gossypium hirsutum* were used in this study (Figure 1; Table 1). Developing fibers were collected at 5, 10, 15, and 20 days post anthesis (dpa), and subjected to RNA extraction. Fibers were vortexed with glass beads to break open the cells, and RNA was extracted using the Sigma-Aldrich Spectrum Plant Total RNA kit (Sigma-Aldrich, St. Louis, MO). The extracted RNAs were then purified using a phenol-chloroform purification protocol. 5M ammonium acetate and phenol:chloroform:isoamyl alcohol were added to the eluted RNA and vortexed for 10 seconds. The samples were then centrifuged for 10 minutes at 4° C for 10 minutes. The aqueous layer was removed and precipitated with isopropanol overnight, then washed with ethanol before resuspending in H₂O. The mRNA libraries were constructed and sequenced on the HiSeq 2500 system with single end 100 bp cycle at the DNA Facility of Iowa State University (http://dna.biotech.iastate.edu/). A total of twenty-four RNA-seq libraries were generated with an average of 30 million reads per library. The raw sequencing data are deposited in Dryad (DOI: http://dx.doi.org/10.5061/dryad.256hn).

RNA-seq data were mapped to the reference diploid *G. raimondii* (D5) genome (Paterson et al. 2012) using the SNP-tolerant mapping option of GSNAP with a cotton-specific SNP database that distinguishes A- and D- subgenomes in allopolyploid cotton (Wu and Nacu 2010; Page et al. 2014). Mapped reads containing diagnostic SNPs were partitioned into A- and D-subgenome by Polycat (Page et al. 2014) to estimate homoeolog-specific expression. Count data were generated using samtools and HTseq-count (Li et al. 2009a; Anders et al. 2015).
**Differential expression analysis**

The R package DEseq2 (Love et al. 2014) was used to evaluate differential expression of genes in both aggregate and partitioned data sets. Gene set enrichment analysis for the significantly differentially expressed genes was performed using the topGO package in R 3.3.1 (R Core Team 2015; Alexa and Rahnenfuhrer 2016) using Fisher’s exact test and gene ontology (GO) annotation from the *G. raimondii* reference genome via CottonGen (Paterson et al. 2012; Yu et al. 2014).

**Weighted gene co-expression network construction**

Gene co-expression networks were constructed using the R package WGCNA (Langfelder and Horvath 2008), for both the aggregated expression of homoeolog gene pairs and gene specific expression partitioned into individual homoeologs. Read counts across different RNA-seq libraries were normalized using the *rlog* function of DESeq2 (Love et al. 2014). The WGCNA function *blockwiseModules* was used to construct networks with default settings. Briefly, Pearson correlations were calculated between each pair of genes, and the correlation matrix was raised to a power of $\beta = 12$ to generate an adjacency matrix representing the connection strengths among genes. Based on the criterion to approximate scale-free network topology, a topological fit index of 0.80 was reached for each constructed network. Next, the adjacency matrix was used to calculate the topological overlap matrix (TOM), which measures network interconnectedness for each pair of genes in relation to all the other genes in network. By performing average linkage hierarchical clustering with a dynamic tree cutting algorithm on the topological overlap dissimilarity measure ($1 – \text{TOM}$), highly interconnected genes were grouped together to reveal the subnetwork organization in modules, for each network. Network hub genes were identified as those with node connectivity ($K$, the sum of connection strengths to
a given gene) over 0.9. As a representative expression profile for the gene members of a given module, module eigengene (ME) was calculated as the first principal component of the scaled expression profiles of all module gene members. Module eigengene-based connectivity (kME), also known as module membership, was calculated for each gene by the Pearson correlation between the gene expression pattern with its respective ME; kME was used to identify module hub genes of kME > 0.9. To determine whether a set of genes (e.g., a list of differentially expressed genes or genes belonging to a GO category) is significantly enriched in the hub of whole network or within a specific module, the ranked lists of network connectivity K or module kME, respectively, were subjected to gene set enrichment analysis (GSEA) with the Preranked function (Mootha et al. 2003; Subramanian et al. 2005).

**Conservation and divergence of gene co-expression networks**

The construction of consensus networks and network preservation tests were performed as previously described (Hu et al. 2016). Briefly, consensus networks were constructed using the minimum topological overlap similarity measure (TOM) derived from individually constructed and calibrated networks. Given the consensus network topology, the preservation measure of module eigengenes between individual networks using the setCorrelationPreservation function of WGCNA (Langfelder and Horvath 2008). To assess how well the modular structure of a reference network is preserved in a test network, the WGCNA function modulePreservation calculates two types of preservation statistics - $Z_{\text{summary}}$ and medianRank scores: Lower medianRank score of a module indicates higher preservation relative to other modules; modules with $Z_{\text{summary}} > 10$ are interpreted as strong preservation, whereas $Z_{\text{summary}}$ between 2 and 10 are weak to moderately preserved, and $Z_{\text{summary}} < 2$ indicates no preservation (Horvath 2011). The
Z\text{summary} is sensitive to module size, while medianRank is not; thus medianRank is used to confirm the Z\text{summary} results.

**Differential correlation analysis**

Rewired edges were calculated by comparing the adjacency matrices of the wild and domesticated data sets. The R package DiffCorr was used to call differentially coexpressed genes between the wild and domesticated aggregate data sets. This package generates correlation coefficients for gene pairs in each data set and compares them via Fisher’s Z test, followed by a local false discovery rate p-value adjustment, as implemented by R package fdrtool (Strimmer 2008; Fukushima 2013). R package DGCA was used to test specifically those genes called as differentially co-expressed in the aggregate data set, rather than using the partitioned data set (McKenzie et al. 2016).

**Results**

**Gene expression in fiber development**

We performed RNA-seq analysis for three domesticated and three wild accessions of *G. hirsutum* over four fiber developmental stages, which encompass the stages of primary fiber cell elongation and transition to secondary cell wall synthesis (Applequist et al. 2001; Rapp et al. 2010; Haigler et al. 2012). A total of 24 libraries, each consisting of 5.5-55.9 million reads were obtained (Table 1), with an average of 30.5% of reads mapped to the reference *G. raimondii* genome.

Principal Components Analysis (PCA; Fig. 2) of the fiber transcriptomic profiles revealed that the first principal component accounts for 76.1% of the total variance and clusters fiber samples by developmental stage. One 10 dpa (days post-anthesis) library of *G. hirsutum*
var. yucatanense may have been mislabeled, as it clusters with 20 dpa samples (Fig. 2; red arrow). This library was excluded from further analyses, and replaced with a previously sequenced RNA-seq library of G. hirsutum var. yucatanense 10 dpa fibers (Fig. 2, blue circle; SRX062250; Yoo and Wendel 2014); the replacement library clustered with the other 10 dpa samples, providing support to the possibility of the original sample’s mislabeling.

Among 37,505 protein-coding loci in the reference diploid G. raimondii genome, a total of 29,706 homoeologous gene pairs (pairs of A-subgenome and D-subgenome genes) were expressed in allopolyploid G. hirsutum during the period of fiber development studied here. When the total expression of homoeologous gene pairs was partitioned into A-subgenome and D-subgenome homoeologs, 25,474 and 25,522 genes were expressed from the A- and D-subgenome, respectively, accounting for 68% of allopolyploid cotton genes.

**Gene co-expression network analysis for G. hirsutum fiber**

In order to investigate the regulatory control of gene expression that underlies cotton fiber development, and how this is modified by domestication, we conducted gene co-expression network analyses to examine interrelationships among fiber genes and the dynamics of gene co-regulatory structures. Using the developing fiber transcriptomes from both wild and domesticated cotton, we employed two different strategies to construct weighted co-expression networks and explore network topologies.

First, we built gene co-expression networks including all fiber datasets from this study, which provide a global view of co-expression relationships shared between wild and domesticated fibers. This analysis was conducted using both the aggregated expression of homoeolog gene pairs and partitioned expression for each homoeolog gene. With each node representing a pair of homoeologous genes, the aggregated network inherently assumes co-
regulation of homoeologs and presents the aggregated topology of the A- and D- genome networks, thus representing a lack of evolution in regulatory regions and the maintenance of similar function for each gene pair. In contrast, the construction of the partitioned homoeolog gene network enables the analysis of regulatory divergence between a pair of homoeologs, and provides measures for both within and between subgenome connections in the duplicated gene network of tetraploid cotton. By comparing the aggregated and partitioned network construction, we aim to test the null hypothesis of co-regulation of homoeologs and investigate the evolution of homoeolog-specific expression and co-expression patterns underlying cotton fiber development.

Secondly, we constructed the wild and domesticated cotton fiber networks for topological comparison, again as aggregate or partitioned networks (as described above). Here we can compare the gene co-expression networks as they are generated in both the wild-only and the domesticated-only data sets, and for aggregated vs. homoeolog-specific data sets. We examine both preservation of module structure and differential correlation between genes at these stages.

**G. hirsutum fiber co-expression network**

A weighted gene co-expression network was generated with WGCNA using aggregate expression and partitioned expression for all samples (Fig. 3; Langfelder and Horvath 2008). In the aggregate network constructed with 29,706 gene pairs, 26 modules were generated, ranging from 72 to 7,659 genes within each co-expression module. Based on module eigengenes, or the first principal component of the scaled expression profiles of all module gene members, 18 modules were significantly associated with developmental stage and domestication status (ANOVA, p < 0.05). The module membership (kME), the correlation between expression of the genes within the module and the module eigengene, ranged from an average 0.645 to 0.798 in
each module. In the partitioned network constructed with 50,996 genes, 52 modules were generated that included between 38 and 9,314 genes. Again, based on module eigengene, 28 modules were significantly associated with developmental stage and domestication status (ANOVA, p < 0.05). Average module membership within each module ranged from 0.609 to 0.878. While the aggregate hub genes, those genes showing high connectivity within the co-expression network, were not enriched for a particular set of biological processes, the partitioned network hub genes were enriched for several categories, including primary and secondary cell wall biogenesis, xylan metabolic, catabolic, and biosynthetic processes, regulation of gibberellin biosynthetic processes, glucoside transmembrane transporter activity, and oxidoreductase activity (GSEAPreranked, P < 0.05, Q < 0.05).

Patterns of differential expression were examined between wild and domesticated accessions. Comparisons were made between consecutive developmental stages within each set and between members of each set at corresponding time points. In the wild accessions, 5,875 gene pairs are differentially expressed in any given developmental comparison, while 7,189 gene pairs are differentially expressed in developmental comparisons in domesticated accessions (Fig. 3, green rows). Thus, domestication increased transcriptional variance, as reported previously (Rapp et al. 2010; but see Yoo and Wendel 2014). In the partitioned data set, 7,962 and 10,076 genes are differentially expressed during development in the wild and domesticated accessions, respectively (Fig. 3, purple rows), demonstrating a similar level of developmental expression variation as a consequence of domestication. Of the genes differentially expressed during development in the aggregate data set, 4,707 (80%) and 5,978 (83%) are represented on the list of developmentally differentially expressed genes in the partitioned data set, and 1,840 (32%) and 2,367 (33%) homoeolog pairs show differential expression in both homoeologs, in wild and
domesticated cottons, respectively. This shows that the proportions of homoeolog bias in differentially expressed genes are approximately equal between the wild and domesticated accessions.

In comparisons between wild and domesticated accessions, 2,190 unique genes were differentially expressed between the wild and domesticated accessions at corresponding stages in the aggregate dataset, while 3,230 unique homoeologs were differentially expressed between these accession sets in the partitioned data set (Fig. 3A, green and purple rows); these numbers are notably lower than those differentially expressed for the developmental comparisons, but are of a similar magnitude to those reported previously (Yoo and Wendel 2014). One thousand five hundred fifty eight (71%) of the aggregate gene pairs were also represented on the partitioned differential expression list, and 309 (14%) homoeolog pairs showed differential expression of both members.

Using both the aggregate and partitioned networks built with the complete list of accessions, we tested for enrichment within the highly connected network hub genes of both developmentally differentially expressed genes and genes differentially expressed between wild and domesticated accessions. In both data sets, we observed significant enrichment for differential expression of all developmental comparisons (GSEAPreranked test, $P < 0.05$, $Q < 0.05$); however we did not find enrichment for those genes differently expressed between wild and domesticated accessions.

We also examined the enrichment of differential expression in those genes that show the highest module membership (kME). These genes are hubs for specific patterns of co-expression, and therefore likely to be biologically important (Langfelder et al. 2013). As done previously by Hu et al. (2016), module expression levels of member genes were summarized by their
eigengene value and related to the sample conditions (here, 2 cultivation conditions x 4 developmental stages = 8 conditions); 18 modules showed significant domestication-related or development-related co-expression patterns in the aggregate co-expression network, while 28 modules were found to be significant in the partitioned network. Gene set enrichment tests show that the hub genes within these modules are enriched for differential expression. Aggregate modules 1, 3, 7, 8, 11, 12, 13, 14, 17, and 25 showed significant enrichment for differential expression associated with development (GSEAPreranked, P < 0.5, Q < 0.5). Modules 3, 4, 5, 6, 7, 8, 9, 11, 12, 14, 17, 20, and 25 showed enrichment for differential expression associated with domestication. In the partitioned network, modules 1, 2, 3, 6, 7, 8, 9, 12, 13, 14, 17, 18, 23, 28, 33, 37, 41, 44, and 46, showed enrichment for differential expression during development, while module 3, 6, 7, 9, 10, 11, 12, 13, 14, 17, 18, 28, 41, and 46 showed enrichment for differential expression between wild and domesticated accessions (GSEAPreranked, P < 0.5, Q < 0.5).

**Functional annotation of modules in G. hirsutum fiber co-expression network**

Genes from modules that were significantly associated with conditions were ranked according to module membership (kME). These genes were subjected to the GSEAPreranked algorithm to ascertain Biological Process Gene Ontology (GO) enrichment within the modules. Aggregate module 1 is the largest module and shows a general trend of increasing expression during development. This module was highly enriched for terms related to cell wall biosynthesis, following our expectation given the transition from primary cell wall synthesis and elongation to secondary cell wall synthesis during this time frame (GSEA, P < 0.05, Q < 0.05; Suppl. Table 1).

Aggregate module 2 showed decreasing expression during development, and was enriched for cellular regulation and ribosomal components (GSEA, P <0.05, Q <0.05; Suppl. Table 1).

Aggregate module 3, which peaked in expression at 15 dpa, showed enrichment for terms related
to meristem development and regulation, and other organ morphogenesis terms (GSEA, $P < 0.05$, $Q < 0.05$; Suppl. Table 1). Aggregate module 14 showed enrichment for terms related to cell wall loosening and multidimensional cell growth; this same module showed higher expression in domesticated fiber, consistent with its much longer fiber cells than in the wild forms (Fig. 1).

The partitioned gene co-expression network shows enrichment for similar terms as in the aggregate gene co-expression network. Partitioned module 1, which shows increasing expression in the developmental phases studied here, shows enrichment for terms related to cell wall biosynthesis (GSEA, $P < 0.05$, $Q < 0.05$, Suppl. Table 2). Partitioned module 2, which shows decreasing expression in these developmental stages, shows enrichment for cellular components and ribosomal assembly (GSEA, $P < 0.05$, $Q < 0.05$, Suppl. Table 2). Partitioned module 3 shows enrichment for terms related to meristem maintenance and organ formation (GSEA, $P < 0.05$, $Q < 0.05$, Suppl. Table 2). Partitioned module 14 shows enrichment for plant cell wall loosening and multidimensional cell expansion (GSEA, $P < 0.05$, $Q < 0.05$, Suppl. Table 2). Partitioned module 17 and module 46 share this enrichment for cell wall loosening (GSEA, $P < 0.05$, $Q < 0.05$, Suppl. Table 2); Despite sharing enrichment for the same categories, module 14 shows higher expression in the domesticated samples, module 17 and module 46 showed higher expression in the wild samples.

**Comparison between aggregate and partitioned networks**

The modules from the aggregate gene co-expression network can be applied to the partitioned gene expression data, split into subgenomes, and tested for preservation of module structure to assess differences between the aggregate and partitioned gene expression patterns. A high $Z_{\text{summary}}$ score, a score based on density and connectivity metrics, and a low $\text{medianRank}$, a
more robust but similar score based on rank, would show high levels of aggregate module structure preservation within the partitioned data set. When applying the aggregate gene modules to the A-subgenome gene expression data, only module 26 was found to be below the $Z_{\text{summary}} = 10$ threshold, the suggested cut-off for strong support for module preservation (Fig. 4; Horvath 2011). In the application of the aggregate modules to the D-subgenome gene expression data, both module 26 and module 15 were below the threshold (Fig. 4). The fact that module 15 is found to be preserved in the comparison with only the A-subgenome shows that there may be a change in gene expression dynamics in only the D-subgenome homoeologs for that module. The general result of this analysis shows that most modules within the aggregate gene co-expression network are preserved within the two subgenomes of the polyploid.

**Homoeolog expression in genes and modules**

Differential expression was calculated between homoeologs of all gene pairs that could be diagnosed by SNPs. For both the wild and domesticated data sets overall, there is slightly higher D homoeolog usage than A homoeolog usage (5.3% and 4.4%, respectively; Table 2). When we examine the homoeolog biases in specific sample sets, we find that this bias is still observed (Table 2). Homoeolog bias was also examined for each module in the aggregate network. A similar pattern was seen as in the wild and domesticated data sets as a whole, with most modules having higher D than A homoeolog usage. Module 13, 15, and 22 were the only modules to show higher usage of A homoeologs. However, module 15 showed much stronger bias (81% difference) toward A homoeolog usage, confirming our previous result showing the preservation of module 15 only in the A-subgenome.

We also examined if homoeologs were found within the same module in the partitioned gene co-expression network. We found that 7,561 homoeolog pairs are contained within the
same module in the partitioned network, about one fifth of all gene pairs, and representing one quarter of gene pairs expressed in the data set. Thus, only about one quarter of expressed genes in fibers are identified as having the same co-expression patterns as their homoeologs in the partitioned gene co-expression network.

**Comparison between domesticated fiber co-expression network and wild fiber co-expression network**

Gene co-expression networks were constructed for the domesticated sample set and the wild sample set and modules were inferred using WGCNA. The aggregate wild co-expression network had 32 modules, while the aggregate domesticated co-expression network had 28 modules; the partitioned network had 108 and 47 modules for wild and domesticated cottons, respectively. In the aggregate network, we observed that 5 and 8 modules were significantly associated with DPA in the wild and domesticated co-expression networks; 10 and 9 modules were significantly associated with DPA in the partitioned wild and domesticated co-expression networks (ANOVA, p < 0.05).

The average correlation preservation, or D statistic, provides a measure of how similar correlation is between sets, here the wild and domesticated gene expression data. The D-statistic was 0.773 between the wild and domesticated networks in the aggregate data set, and 0.763 between the wild and domesticated networks in the partitioned data set, showing that there is divergence in co-expression patterns between wild and domesticated accessions. These numbers align with previous comparisons between wild and domesticated networks in cotton seed (Hu et al. 2016).

To better assess which modules were preserved in cotton following domestication, we calculated the $Z_{\text{summary}}$ score, as we did when comparing the aggregate network and partitioned
gene expression data set. We applied the wild module structure to the domesticated gene expression to determine which modules are poorly preserved, and most likely altered by domestication. In the aggregate network comparison, we observed that 17 of the 32 modules were strongly preserved \((Z_{\text{summary}} > 10)\), including all of the modules significantly associated with DPA in the wild gene co-expression network; the medianRank score followed the \(Z_{\text{summary}}\), showing that those modules with low \(Z_{\text{summary}}\) scores have high medianRank scores (Fig. 5). In the partitioned network, 26 of 108 modules are strongly preserved \((Z_{\text{summary}} > 10)\). Seven of the 10 modules significantly associated with DPA were not found to be strongly preserved. The wild gene co-expression network modules that are poorly preserved in the domesticated data set may represent targets of domestication.

**Overlap of differential expression and co-expression between wild and domesticated accessions**

While co-expression describes how genes are correlated in their expression patterns, differential co-expression highlights the potentially biologically meaningful changes in these relationships. We used the R package DiffCorr to look at differential co-expression between the wild and domesticated aggregate data sets (Fukushima 2013). We found that 2.9% of edges appeared to be rewired throughout the entire data set (2.8% in the partitioned network). A total of 15,364 significantly differentially co-expressed (DC) pairs \((p < 0.05, q < 0.05)\); only 1,273 genes (4.3%) were found to show significant differential co-expression \((p < 0.05, q < 0.05)\). These DC genes were enriched for cellular catabolic processes and extracellular membrane via topGO \((p < 0.05, q < 0.05;\) Alexa and Rahnenfuhrer 2016). In the partitioned data set, we found that 1,957 homoeologs were differentially correlated, 76.9% of those expected if both homoeologs of the
gene pairs that are differentially correlated in the aggregate data sets are also differentially correlated in the partitioned data set.

The overlap of differential co-expression and differential expression identifies those genes that are exhibiting the most extreme transcriptional changes. We compared the differentially expressed genes with the co-expressed genes, finding that 407 gene pairs overlapped between significantly differentially expressed between any condition and significantly differentially co-expressed. We subjected these overlapping genes to GO enrichment analysis to determine if any specific biological processes were highly represented. However, no GO terms were significantly enriched (P < 0.05, Q < 0.05). We confirmed that 364 and 368 corresponding homoeologs were also differentially correlated in the partitioned data set, in the A-subgenome and D-subgenome, respectively, with an overlap of 345. Twenty of the DC genes from the aggregate data set showed no differential correlation in the partitioned data set. A list of the overlapping genes and their D5 annotation can be found in Suppl. Table 3.

Discussion

In this study, we examined the gene co-expression network in both wild and domesticated fibers of *G. hirsutum*. Using a combination of weighted gene co-expression, differential expression, and differential co-expression analyses, we have shown that the overall co-expression patterns of genes in these two data sets are relatively preserved among homoeologs. Notwithstanding this relatively high module preservation among homoeologs, we have also shown that specific modules show preservation only for homoeologs from one subgenome, such as module 15 in the aggregate data set, and that homoeolog biases can be module-dependent (Table 2). Regarding domestication, we found that there were extreme changes in transcription
between wild and domesticated fiber co-expression networks following the strong directional selection represented by domestication. We also identified specific modules and genes that show differential expression and co-expression between the two cultivated conditions. This allows us to identify trends within the gene expression landscape with regards to domestication. Further, we were able to identify sets of genes that might prove important for the phenotypic shift between wild and domesticated fibers, and how they tie in to the transcriptomic network.

The intersection between differential expression and differential co-expression provides clues to cotton fiber elongation and secondary cell wall transition

In our previous study, we aimed to show the power of gene co-expression network analysis, and limited our analysis to the profilin gene family (Gallagher et al. 2016). Here, we have expanded our scope to look at the entire fiber transcriptome. In regards to differential expression, we found that approximately 5,000 and 7,000 genes were differential expressed during development for wild and domesticated fiber, respectively. These data can be integrated with our differential co-expression data to narrow the search for genes affected by domestication. Those genes that show significant differential gene expression that are also found to be differentially coexpressed show where patterns of expression have been altered. These patterns of alteration may represent parts of the transcriptional network that have been subject to domestication or improvement. While we did not find specific GO enrichment, the closest *Arabidopsis thaliana* homologs, as annotated in the D5 reference genome (Paterson et al. 2012; Yu et al. 2014), can reveal likely functions of these genes (Suppl. Table 3). Many of these overlapping genes are homologs of transcription factors of various types. Wan et al. (2014) previously showed that transcription factors showed differential expression between domesticated and naked seed or fuzzless mutants. These included MYBs, bHLH, WD40,
WRKY, bZIP, and others. Specifically, they uncovered that the cotton homolog of AtMYB5 was found to be differentially expressed; here we extend that to differential co-expression for the cotton homolog of AtMYB5 (Gorai.008G126700). In Arabidopsis, AtMYB5 causes outer seed coat differentiation and trichome development (Li et al. 2009b). It may also be responsible for initiating cell wall development in the fiber here. Several other transcription factors were also identified, such as zinc fingers transcription factors, including C2H2 and RING domains, leucine rich repeats responsive to disease, bZIPs, bHLH, and WD40.

Several cell wall growth and elongation genes were also identified in the data set. We identified homologs of expansin 15 (Gorai.005G142200) and expansin 20 (Gorai.008G013200), which loosens cell walls for elongation through acid growth, and xyloglucan transferase/hydrolases (XTH, Gorai.010G177300, Gorai.011G016500), which are proposed to be both capable of cell wall loosening and strengthening, and have been shown to increase G. hirsutum fiber length in overexpression transgenics (Cosgrove 2005; Lee et al. 2010). The homolog of expansin 15 shows higher expression in 10 dpa cotton over 15 in both sets, and has seemingly balanced co-expression between the two. However, the expansin 20 homolog shows increased expression in wild early in development (5 dpa), but changes to higher expression in domesticated cotton by 20 dpa; both wild and domesticated cotton show a general increase in expression over time. These XTH homologs are all upregulated in the wild 20 dpa fiber compared to 15 dpa fiber, but not in domesticated fiber. This, along with the patterns observed for the expansin 20 homolog, mirrors the delayed transition to secondary cell wall synthesis that has been observed in domesticated cotton fiber.

A category found in large numbers in the overlapping differentially co-expressed and differentially expressed gene data set are genes encoding reactive oxygen species (ROS)
proteins. ROS have been proposed to be involved in cell expansion (Cosgrove 2005). Hovav et al. (2008b) and Chaudhary et al. (2009) both showed that the ROS signaling pathway genes were altered by domestication; here we expand those studies to show that differential co-expression is also occurring. Cosgrove (2005) synthesized evidence for the loosening of the cell wall for cell elongation via hydroxyl radicals, a form of reactive oxygen species. Hovav et al. (2008b) and Chaudhary et al. (2009) extended this possibility to cotton fiber cell walls, making an argument for the increased expression of ROS producing and regulating genes assisting with cell wall expansion. Many genes can be involved in ROS scavenging and balance. In our data, both a glutathione peroxidase homolog (Gorai.004G215500) and a generic peroxidase family homolog (Gorai.012G046400) appear in our overlap between differential expression and differential co-expression, both exhibiting higher expression in domesticated accessions over wild and gain in correlation in the domesticated data in some cases. We also see a homolog of the NADPH oxidoreductase superfamily (Gorai.005G151000). These genes could be responsible for more effectively sequestering and balancing ROS in domesticated cotton during cotton fiber elongation.

Another candidate from the work by Hovav et al. (2008b) also appeared in the overlap data set. A cotton GAST1 homolog (Gorai.006G017000) shows down-regulation of expression in wild accessions from 10 to 15 dpa and in the domesticated accessions from 15 dpa to 20 dpa; it also showed increasing levels of co-expression in the domesticated accessions compared to the wild. These patterns support the hypothesis that GAST1 and GAST-like proteins, gibberellin-induced genes with ROS regulating ability, are involved in cell elongation in other plant tissues, as cotton has been shown to have an extended period of elongation in the domesticated fiber
before secondary cell wall biosynthesis (Shi and Olszewski 1998; Wigoda et al. 2006; Haigler et al. 2012).

Ethylene and responses to ethylene are also thought to be important in the development of fiber development. Previous reports comparing cultivated fiber and fiberless mutants have shown that expression of ethylene biosynthesis is ongoing through fiber development, and that the expression of certain fiber development genes is associated with high ethylene levels, including sucrose synthase, tubulin, and expansin (Shi et al. 2006). In our overlap data set, we find homologs of ERF1 (Gorai.008G232500), EBF1 (Gorai.001G042600), and ETR2 (Gorai.006G247800), genes related to ethylene response and regulation, that show higher expression in domesticated accessions over wild, and show increased co-expression in domesticated accessions. In the ethylene signaling pathway, ERF1 binds DNA to trigger expression of abiotic stress-response genes, while EBF1 regulates EIN3, an upstream ethylene signaling factor, and other genes through ubiquitination (Potuschak et al. 2003; Chang 2016). A homolog of ethylene signaling gene ETR2 also appears in the same module as the EBF1 homolog (Sakai et al. 1998). In corroboration with Shi et al. (2006), we also find two expansins in our overlap data set (see above), one of which shows a similar expression pattern as the ethylene response factor and related genes.

Studies have previously linked ethylene to ROS scavenging capabilities and cotton fiber elongation (Li et al. 2007; Qin et al. 2008). Recently, Zhang et al. (2016) showed that TERF1, a tomato ethylene response factor, can induce genes for ROS scavenging, such as glutathione peroxidase. The ethylene related genes that we see more highly expressed in the domesticated data sets may be activating the genes necessary to sequester and balance ROS during cotton fiber elongation, as was seen in TERF1; this could allow the fiber cell wall to loosen and elongate,
alongside other factors in cell elongation, such as XTH, with greater control over the balance of ROS than is seen in wild fiber. Molecular biological work supporting ERF-based promotion of the ROS sequestration genes will be necessary for confirmation of this hypothesis.

**Polyplody and gene co-expression**

Polyplody, or whole genome duplication, has been revealed in the evolutionary past of every seed plant species, and has been identified as ongoing and recurring in many modern day plant species (Jiao et al. 2011; Soltis et al. 2014; Wendel 2015). In allopolyploids, the transcriptomic environment is altered; while the *cis*-regulatory elements remain the same for individual genes, the *trans*-regulatory environment now contains factors from both subgenomes. The novel interactions between the *cis* and *trans* factors in the polyploid expression environment may lead to rewiring throughout the network (De Smet and Van de Peer 2012). Gene co-expression networks can help us to identify those modules that may be undergoing these changes.

Here, using cotton as a model for polyploidy-caused gene expression alteration, we have shown how gene co-expression networks can be used to identify changes due to whole genome duplication. By analyzing the data as both aggregate and partitioned expression, we are able to observe changes in gene co-expression due to polyploidy. We found that the aggregate co-expression network contained 26 modules, while the homoeolog partitioned co-expression network contained 52 modules. This shows that there must be differences in expression patterns between homoeologs, enough so that they can generate different sets of modules when their expression is partitioned. Indeed, only a quarter of the expressed homoeologs were found in the same module in the partitioned network. However, while expression may be somewhat different among homoeologs, it is not as different as it initially appears. Tests of module preservation
show that the modules from the aggregate co-expression network are mostly well-preserved in the partitioned data set (Fig. 4), with the exception of module 15 (but see below). Thus, we find that, overall, gene co-expression is not drastically changed due to polyploidy, but does show some nuanced change.

By focusing on individual co-expression modules, we can further assess this nuanced change. Homoeolog bias, the difference in contribution of each homoeolog to the total expression, can occur based on co-expression modules. Here we calculated it for domestication status, developmental stage, and module. Overall, we calculated a D-subgenome bias in expression; this mirrors results of some studies, but is counter to recent studies suggesting a more balanced bias (Hovav et al. 2008a; Yoo and Wendel 2014; Zhang et al. 2015). While our domestication and development-based calculations all showed bias towards higher expression of the D-subgenome, partitioning homoeolog bias into modules showed that certain modules are actually biased toward higher expression in the A-subgenome, in particular module 15, which showed 81.7% difference in homoeolog expression bias (Table 2). This was supported by our analysis of module preservation within the partitioned data sets, which showed that module 15 was more strongly preserved in the A-subgenome expression data set than in the D-subgenome data set. Based on GSEAPreranked, module 15 was enriched for Biological Process GO Terms involved in cell wall biosynthesis. ME15, the representative expression measure for the module, shows lower expression in domesticated cotton, with its highest peak at 20 dpa; but shows high levels of expression at both 5 and 20 dpa in wild cotton. Based on these expression values and GO terms, this module may contain A-subgenome genes that are important for the fiber phenotype that we see in polyploid versus diploid cotton, as well as for gene expression changes
leading to the domesticated fiber phenotype. This is one example of the nuance that gene co-expression networks can reveal in regards to polyploidy.

**Domestication and gene co-expression networks**

This is one of the first studies to use gene co-expression network analysis for studying cotton fiber (but see Gallagher et al. 2016; You et al. 2016; You et al. 2017). Gene co-expression networks were constructed using both aggregated and partitioned gene expression, leading to global networks with 26 and 52 co-expression modules, respectively. We further separated these into wild and domesticated data sets, and then generated wild and domesticated gene co-expression networks. These had 32 and 28 modules, respectively, in the aggregate data set, and 108 and 47 modules in the partitioned data set.

Enrichment of highly connected intramodular hub genes can show functions that acting together in fiber development. Gene set enrichment analysis (GSEA) was used to assess which modules were associated with various biological processes. We did detect enrichment for the module hub genes and other genes within the modules (Suppl. Table 1). The gene sets were enriched for processes such as cell wall biogenesis, fatty acid biosynthesis, flavonoid biosynthesis, and others. Previous studies have shown that modules, which are subnetworks within the larger gene co-expression network, are likely to have biological significance (Langfelder et al. 2013). These modules enriched for specific biological processes are starting points for continued investigation. Gene expression knock downs, specifically targeting intramodular hub genes, may further define the true interactions of the modules, and not just those that show similar expression patterns.

Here, we have shown that strong directional selection practiced by humans during the domestication and crop improvement process not only has affected gene expression itself, but
that modules of co-expressed genes have been altered. Through the module preservation test, we are able to see that only about half of the modules in the aggregate wild gene co-expression network are strongly preserved in the domesticated data set; in the case of the partitioned wild gene co-expression network, only about one quarter of the co-expression modules are strongly preserved in the domesticated data set. This shows that the loss of preservation observed in the aggregate data set appears even more severe when considered as a partitioned data set. Since only a quarter of the genes in the aggregate network are found in the same modules within the partitioned network, the partitioned approach may more accurately represent the transcriptional state of the fibers.

Conclusions

In this work, we generated RNA-seq libraries for several accessions of wild and domesticated cotton over four developmental stages focusing on fiber elongation, primary cell wall synthesis, and transition to secondary cell wall. From these data, we performed expression analyses, including differential expression, differential co-expression, and gene co-expression network analysis. Combining these three strategies, we were able to identify important large scale trends, such as co-expression modules that are only preserved in one of two subgenomes, and individual genes of interest, such as those that overlap in differential expression and differential co-expression. By comparing networks constructed solely from wild or domesticated fibers data sets, we were able to show that the topology of the wild network has been drastically altered by domestication. This only becomes more obvious by separating the data set into the two subgenomes. These advances demonstrate the advantages of using multiple analysis methods and partitioning methods when working with large scale expression data in polyploid species.
Acknowledgements

We thank Anna Tuchin and Simon Renny-Byfield for assistance with RNA extractions. This work was funded by Cotton Incorporated and the National Science Foundation Plant Genome Program. Joseph P. Gallagher is supported by NSF Graduate Research Fellowship DGE1247194.

References


Anders, S., P. T. Pyl and W. Huber. 2015. HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics. 31: 166-169


Mootha, V. K., C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P.
Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson,
J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D.
Altshuler and L. C. Groop. 2003. PGC-1 alpha-responsive genes involved in oxidative

Page, J. T., Z. S. Liechty, M. D. Huynh and J. A. Udall. 2014. BamBam: genome sequence
analysis tools for biologists. *BMC Res Notes.* 7: 829

Showmaker, S. Q. Shu, J. Udall, M. J. Yoo, R. Byers, W. Chen, A. Doron-Faigenboim,
Saranga, B. E. Scheffler, J. A. Scheffler, D. M. Stelly, B. A. Triplett, A. Van Deunze, M.
F. S. Vaslin, V. N. Waghmare, S. A. Walford, R. J. Wright, E. A. Zaki, T. Z. Zhang, E. S.
2012. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable
Wheat Genome Sequencing, K. F. Mayer and O. A. Olsen. 2014. Genome interplay in the
grain transcriptome of hexaploid bread wheat. Science. 345: 1250091

2003. EIN3-dependent regulation of plant ethylene hormone signaling by two arabidopsis
F box proteins: EBF1 and EBF2. Cell. 115: 679-689

Qin, Y. M., C. Y. Hu and Y. X. Zhu. 2008. The ascorbate peroxidase regulated by H(2)O(2) and
ethylene is involved in cotton fiber cell elongation by modulating ROS homeostasis.
Plant Signal Behav. 3: 194-196

R Core Team. 2015. R: A language and environment for statistical computing. Vienna, Austria:

expression in developing fibres of Upland cotton (Gossypium hirsutum L.) was massively
altered by domestication. Bmc Biology. 8: 139

1998. ETR2 is an ETR1-like gene involved in ethylene signaling in Arabidopsis. Proc
Natl Acad Sci U S A. 95: 5812-5817

Shi, L. and N. E. Olszewski. 1998. Gibberellin and abscisic acid regulate GAST1 expression at
the level of transcription. Plant Mol Biol. 38: 1053-1060

Wang and Y. X. Zhu. 2006. Transcriptome profiling, molecular biological, and
physiological studies reveal a major role for ethylene in cotton fiber cell elongation. Plant
Cell. 18: 651-664


Table 1. RNA sequencing libraries, read counts, and read mapping success. The bolded library was replaced the other *G. hirsutum* var. *yucatanense* 10 dpa sample.

<table>
<thead>
<tr>
<th>Accession</th>
<th>Status</th>
<th>Stage</th>
<th>Total Reads</th>
<th>Total Mapped</th>
<th>A&lt;sub&gt;T&lt;/sub&gt; Reads</th>
<th>D&lt;sub&gt;T&lt;/sub&gt; Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRB252</td>
<td>Dom</td>
<td>5 dpa</td>
<td>55,901,222</td>
<td>19,656,544</td>
<td>9,623,973</td>
<td>10,032,571</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 dpa</td>
<td>29,258,894</td>
<td>10,440,615</td>
<td>5,120,169</td>
<td>5,320,446</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15 dpa</td>
<td>33,262,075</td>
<td>11,777,189</td>
<td>5,768,112</td>
<td>6,009,077</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20 dpa</td>
<td>34,011,731</td>
<td>11,478,997</td>
<td>5,732,263</td>
<td>5,746,734</td>
</tr>
<tr>
<td>Maxxa</td>
<td>Dom</td>
<td>5 dpa</td>
<td>44,718,091</td>
<td>15,892,094</td>
<td>7,723,781</td>
<td>8,168,313</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 dpa</td>
<td>29,367,962</td>
<td>10,040,583</td>
<td>4,933,497</td>
<td>5,107,086</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15 dpa</td>
<td>31,686,165</td>
<td>7,356,773</td>
<td>3,635,382</td>
<td>3,721,391</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20 dpa</td>
<td>5,539,104</td>
<td>864,837</td>
<td>440,826</td>
<td>424,011</td>
</tr>
<tr>
<td>TM-1</td>
<td>Dom</td>
<td>5 dpa</td>
<td>38,014,167</td>
<td>13,443,793</td>
<td>6,593,740</td>
<td>6,850,053</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 dpa</td>
<td>20,985,382</td>
<td>7,464,153</td>
<td>3,681,456</td>
<td>3,782,697</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15 dpa</td>
<td>29,554,346</td>
<td>10,116,872</td>
<td>4,965,178</td>
<td>5,151,694</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20 dpa</td>
<td>29,510,081</td>
<td>9,824,471</td>
<td>4,832,783</td>
<td>4,991,688</td>
</tr>
<tr>
<td>var. yucatanense</td>
<td>Wild</td>
<td>5 dpa</td>
<td>27,776,650</td>
<td>7,272,897</td>
<td>3,590,323</td>
<td>3,682,574</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 dpa</td>
<td>29,318,675</td>
<td>10,033,985</td>
<td>5,013,185</td>
<td>5,020,800</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 dpa</td>
<td>16,282,320</td>
<td>10,906,980</td>
<td>5,406,412</td>
<td>5,500,568</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15 dpa</td>
<td>26,232,665</td>
<td>8,815,079</td>
<td>4,403,285</td>
<td>4,411,794</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20 dpa</td>
<td>38,162,310</td>
<td>2,293,628</td>
<td>1,137,897</td>
<td>1,155,731</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TX 665</td>
<td>Wild</td>
<td>5 dpa</td>
<td>10 dpa</td>
<td>15 dpa</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>29,684,989</td>
<td>29,511,233</td>
<td>37,612,622</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>8,311,319</td>
<td>7,293,986</td>
<td>13,101,674</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4,087,015</td>
<td>3,588,812</td>
<td>6,396,135</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4,224,304</td>
<td>3,705,174</td>
<td>6,705,539</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 dpa</td>
<td>15 dpa</td>
<td>20 dpa</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>27,953,610</td>
<td>33,681,294</td>
<td>35,291,246</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>9,067,078</td>
<td>11,124,041</td>
<td>10,564,456</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4,484,549</td>
<td>5,550,605</td>
<td>5,285,810</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4,582,529</td>
<td>5,573,436</td>
<td>5,278,646</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TX 2095</td>
<td>Wild</td>
<td>5 dpa</td>
<td>10 dpa</td>
<td>15 dpa</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>31,797,359</td>
<td>26,439,223</td>
<td>33,681,294</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>9,357,464</td>
<td>8,966,075</td>
<td>11,124,041</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4,591,086</td>
<td>4,361,536</td>
<td>5,550,605</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4,766,378</td>
<td>4,604,539</td>
<td>5,573,436</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 dpa</td>
<td>15 dpa</td>
<td>20 dpa</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>35,291,246</td>
<td>33,681,294</td>
<td>35,291,246</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,564,456</td>
<td>11,124,041</td>
<td>10,564,456</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5,285,810</td>
<td>5,550,605</td>
<td>5,285,810</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5,278,646</td>
<td>5,573,436</td>
<td>5,278,646</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Homoeolog expression bias by sample and module

<table>
<thead>
<tr>
<th>Set</th>
<th>A</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LFC &gt; 0</td>
<td>LFC &gt; 0.5</td>
</tr>
<tr>
<td>Wild</td>
<td>6307</td>
<td>5598</td>
</tr>
<tr>
<td>Domesticated</td>
<td>7375</td>
<td>6507</td>
</tr>
<tr>
<td>Wild 5 DPA</td>
<td>2785</td>
<td>2784</td>
</tr>
<tr>
<td>Wild 10 DPA</td>
<td>2141</td>
<td>2140</td>
</tr>
<tr>
<td>Wild 15 DPA</td>
<td>1990</td>
<td>1990</td>
</tr>
<tr>
<td>Wild 20 DPA</td>
<td>2209</td>
<td>2207</td>
</tr>
<tr>
<td>Dom 5 DPA</td>
<td>3812</td>
<td>3810</td>
</tr>
<tr>
<td>Dom 10 DPA</td>
<td>3425</td>
<td>3422</td>
</tr>
<tr>
<td>Dom 15 DPA</td>
<td>3590</td>
<td>3588</td>
</tr>
<tr>
<td>Dom 20 DPA</td>
<td>2580</td>
<td>2580</td>
</tr>
<tr>
<td>Module 0</td>
<td>280</td>
<td>250</td>
</tr>
<tr>
<td>Module 1</td>
<td>1800</td>
<td>1561</td>
</tr>
<tr>
<td>Module 2</td>
<td>2180</td>
<td>1869</td>
</tr>
<tr>
<td>Module 3</td>
<td>498</td>
<td>478</td>
</tr>
<tr>
<td>Module 4</td>
<td>360</td>
<td>262</td>
</tr>
<tr>
<td>Module 5</td>
<td>339</td>
<td>298</td>
</tr>
<tr>
<td>Module 6</td>
<td>251</td>
<td>237</td>
</tr>
<tr>
<td>Module 7</td>
<td>220</td>
<td>157</td>
</tr>
<tr>
<td>Module 8</td>
<td>242</td>
<td>201</td>
</tr>
<tr>
<td>Module</td>
<td>Start</td>
<td>End</td>
</tr>
<tr>
<td>--------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>9</td>
<td>229</td>
<td>179</td>
</tr>
<tr>
<td>10</td>
<td>118</td>
<td>77</td>
</tr>
<tr>
<td>11</td>
<td>117</td>
<td>101</td>
</tr>
<tr>
<td>12</td>
<td>128</td>
<td>125</td>
</tr>
<tr>
<td>13</td>
<td>124</td>
<td>88</td>
</tr>
<tr>
<td>14</td>
<td>118</td>
<td>110</td>
</tr>
<tr>
<td>15</td>
<td>216</td>
<td>209</td>
</tr>
<tr>
<td>16</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>17</td>
<td>61</td>
<td>59</td>
</tr>
<tr>
<td>18</td>
<td>101</td>
<td>69</td>
</tr>
<tr>
<td>19</td>
<td>61</td>
<td>54</td>
</tr>
<tr>
<td>20</td>
<td>87</td>
<td>79</td>
</tr>
<tr>
<td>21</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>22</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>23</td>
<td>39</td>
<td>21</td>
</tr>
<tr>
<td>24</td>
<td>27</td>
<td>14</td>
</tr>
<tr>
<td>25</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>26</td>
<td>20</td>
<td>19</td>
</tr>
</tbody>
</table>
Figure 1. Cotton fibers from the samples examined here. Each sample shows a single seed.

TX2095, TX665, var. *yucatanense* = wild cotton. CRB252, Maxxa, TM-1 = domesticated cotton.
Figure 2. PCA of RNA-seq libraries. PC1 generally follows developmental stage. The red arrow denotes the *G. hirsutum* var. *yucatanense* sample that grouped with the 20 dpa samples, and was removed from the analysis. The grey circle highlights the *G. hirsutum* var. *yucatanense* sample from a previous data set that replaced the bad sample.
Figure 3. Dendrogram, gene co-expression module assignment, and differential expression in *G. hirsutum* fiber expression data set. The dendrogram shows the clustering of genes based on co-expression patterns, which was then used to group genes into modules. Aggregate network modules are shown by colors in the first bar, while partitioned network modules are shown in the next two bars; these two bars share colors, as each bar represent homoeolog from a specific subgenome within the same network. The next three columns show the gene significance value for each node (gene pair or homoeolog) for developmental stage. Blue represents negative
correlation, i.e. expression is higher at earlier developmental stages, while red represents increasing correlation, i.e. higher expression in later stages. Green and purple bars show genes that are differential expressed in a given comparison for the aggregate (green) or partitioned (purple) gene co-expression network.
Figure 4. Module preservation of aggregate gene co-expression network topology in A- and D-subgenomes. The top two graphs show aggregate module preservation in the A-subgenome fiber expression data, while the bottom two graphs show the aggregate module preservation in the D-subgenome fiber expression data. Module numbers are shown on graphs; the same numbers correspond to the same aggregate modules. Red dashed line, $Z_{\text{summary}} = 10$. 
Figure 5. Module preservation of aggregate wild cotton fiber gene co-expression network in the domesticated cotton fiber gene expression data. Red dashed line, $Z_{\text{summary}} = 10$. Blue dashed line, $Z_{\text{summary}} = 2$.

Supplementary Files

Table S1. Biological Process Gene Ontology terms enriched in significant modules in the aggregate gene co-expression network.

Table S2. Biological Process Gene Ontology terms enriched in significant modules in the partitioned gene co-expression network.

Table S3. Overlap of genes found to be differentially expressed and differentially co-expressed.
Abstract

While previous studies have identified evolutionary trends of genes in pathways in diploids, the effects of polyploidy on the evolution of gene pathways and their expression dynamics have been less thoroughly explored. Here, we use targeted sequence capture to calculate evolutionary rates of genes in the anthocyanin biosynthesis pathway (ABP) across a broad sampling of *Gossypium hirsutum* and *G. barbadense* accessions, and to evaluate these rates in the context of pathway features. Although pathway features and evolutionary rates appeared to be uncorrelated, we do find a significant difference in evolutionary rates between the two subgenomes of polyploid cotton. We generated a gene co-expression network to assess global transcriptomic patterns and expression of duplicated genes in the ABP. Diploid networks are generally preserved within the polyploid data sets, although some modules show divergence. Homoeolog bias and expression level dominance and transgression differ depending on module and based on tissue and accession. Evolutionary rates are significantly correlated with gene co-expression features, specifically with expression level, module assignment, and gene co-expression connectivity. We discuss broad implications for polyploid pathway evolution and expression, as well as for polyploid gene co-expression networks in general.
Introduction

Polyploidy, or whole genome duplication (WGD), is an evolutionarily significant phenomenon, due to the increase in potential outcomes afforded by duplication of the entire complement of chromosomes and genes. Many have explored and debated the evolutionary trajectories of polyploid species and their duplicated genes, particularly when WGD is combined with hybridization in the phenomenon of allopolyploidy (Stephens 1951; Lynch and Force 2000; Des Marais and Rausher 2008; Soltis et al. 2014). This interest, as well as advances in molecular and bioinformatic tools, has generated a great deal of insight regarding both the frequency of polyploidy and the fate of duplicated genes and genomes. Polyploidy occurs ubiquitously throughout the evolutionary history of angiosperms, being both recurrent and ongoing in many plant lineages (Soltis and Soltis 1999; Wood et al. 2009; Jiao et al. 2011; Wendel 2015; Soltis et al. 2016). Even plants with small genomes, such as Arabidopsis thaliana (125 Mb) and Utricularia gibba (82 Mb), have been found to harbor multiple episodes of WGD within their evolutionary past (The Arabidopsis Genome Initiative 2000; Ibarra-Laclette et al. 2013).

Following polyploidy, the resulting duplicated genes may experience several fates, the most frequent being deletion via genome fractionation. In this process, mutational and deletional mechanisms break or eliminate functional genes or their regulatory elements. This process may be biased with respect to progenitor genome (Woodhouse et al. 2010; Freeling et al. 2012). Fractionation helps to explain how polyploid genomes slowly return to a more diploidized state, and can account for the observation of plants with small genomes despite an evolutionary history of genome expansion through polyploidy (Wendel 2015).

Several hypotheses have been posited to explain the types of genes that are retained or lost following polyploidy, the most prominent of which is the gene balance hypothesis. This
hypothesis proposes that genes involved in generating multi-molecular complexes are more likely to be retained following WGD than genes whose products function as single proteins due to stoichiometric balance requirements among components of multimeric complexes (Birchler et al. 2005; Birchler and Veitia 2012). Differential loss of a gene encoding one subunit of the complex may lead to wasteful excess of other subunits retained in duplicate; large-scale genomic surveys suggest that this is indeed a factor in gene retention following polyploidy (Bekaert et al. 2011).

Additional hypotheses for duplicate gene retention relate to divergence in function for duplicated genes, i.e. neofunctionalization and subfunctionalization (reviewed in Conant and Wolfe 2008). Neofunctionalization involves new function(s) arising for one of both gene duplicates. In subfunctionalization, the function of the duplicated gene copies is maintained, while expression occurs in new contexts. Routes to subfunctionalization include (1) duplication-degeneration-complementation (DDC), whereby duplicated genes undergo complementary degenerative regulatory mutations, and (2) escape from adaptive conflict (EAC), in which duplicated genes with multiple functions become better adapted for a subset of their original functions (Force et al. 1999; Des Marais and Rausher 2008). While these paths for duplicate gene retention and evolution have been considered in the context of genes or whole genomes, they have been infrequently evaluated in the context of interconnected biosynthetic pathways and gene regulatory networks (although see Bekaert et al. 2011; Pfeifer et al. 2014; Hu et al. 2016).

Likewise, gene pathway evolution has recently become better understood for diploid species, although the concepts derived from this research have not yet been applied to polyploid systems. Research by Rausher et al. (1999) on the anthocyanin biosynthetic pathway demonstrated that the relative evolutionary rates of genes in a pathway may be influenced by
pathway features, with the foremost being position in the pathway (Lu and Rausher 2003; Rausher et al. 2008). Subsequent research has also highlighted that pathway branching, pleiotropy, and pathway connectivity can influence evolutionary rates (Ramsay et al. 2009; Yang et al. 2009; Olson-Manning et al. 2012).

Combining concepts from polyploid and pathway evolution research leads to several possibilities for the suite of duplicates by polyploidy, or homoeologs. From the perspective of gene retention, homoeologous genes may experience loss (or nonfunctionalization) or they may be retained in duplicate due to subfunctionalization or neofunctionalization. These latter two possibilities are particularly interesting because duplication of the whole pathway via polyploidy means that the pathway itself, or sections of it, has the opportunity to undergo sub- and neofunctionalization. Changes in polyploid expression dynamics, including homoeolog expression bias (reviewed in Grover et al. 2012b) and altered co-expression patterns among pathway genes (Hu et al. 2016) may also be influenced by pathway dynamics. New interactions may arise or old interactions may be lost in the new cellular environment with a doubled gene complement, relative to that in the diploid progenitors. Furthermore, changes in co-expression might lead to changes in expression ratios of pathway genes and ultimately the proportions of final metabolites.

The anthocyanin biosynthetic pathway (ABP) is one of the best described pathways in plants. First studied by Mendel as a segregating characteristic in peas, the ABP and resulting pigments comprise a model system for explorations in evolution, genetics, and molecular biology (Grotewold 2006; Smith et al. 2013). Anthocyanins are a subset of the flavonoids that make up the blue, purple, and red pigments we see in the petals of the flowering plants (Grotewold 2006). However, these pigments occur throughout the plant, performing functions such as UV and pest
protection (Agati et al. 2013). While only eight proteins make up the core ABP, there are several branches from the pathway that lead to other flavonoid classes, such as flavonols, proanthocyanidins, and flavones (Fig. 1; Grotewold 2006). The ABP has been studied in model plants such as *Arabidopsis*, maize, snapdragon, petunia, and morning glory, which has yielded a great understanding of the pathway easily extended to understanding the evolution of this pathway in a neoallopolyploid context.

A promising model, in this sense, is the genus *Gossypium*, which contains the various cotton species. Existing research into the anthocyanin pathways includes evaluating the segregation of floral colors and petal spots in the genus (McLendon 1912); evaluating mutants and variation in floral and vegetative color (Harland 1929; Silow 1942; Stephens 1948a); and generating the flavonoid profile for 29 cotton species (Parks et al. 1975). More recently, transcriptomic changes have been described for ABP genes as the flowers of *Gossypium hirsutum* senesce (Tan et al. 2013). As a polyploid model system, *Gossypium* has been extensively studied for genomic, transcriptomic, and proteomic changes (e.g. Flagel et al. 2012; Yoo et al. 2013; Hu et al. 2014; Zhang et al. 2015).

The genus *Gossypium* arose between 5 and 10 million years ago (Mya) and contains more than 50 species distributed pantropically throughout the world. Between 1 and 2 Mya, a member of the African A-genome clade traveled across the ocean to the New World, where it hybridized with a member of the native D-genome clade to form an allotetraploid, the ancestor of all modern allotetraploid cottons (Wendel and Grover 2015). This resulted in a single allotetraploid lineage that has diversified into seven species and spread throughout the New World tropics and the Pacific Islands (Krapovickas and Seijo 2008; Grover et al. 2012a; Grover et al. 2015b; Gallagher et al. 2017). Four members of the genus have independently undergone domestication,
two diploids from the A-genome clade, and two members of the AD-genome polyploid clade, *G. hirsutum* and *G. barbadense* (i.e. Upland cotton and Pima cotton, respectively); these last two make up the vast majority of global cotton cultivation.

Here, we address the evolution of the anthocyanin biosynthetic pathway in the genus *Gossypium* L. following hybridization and polyploidization. Our null hypothesis is that genic evolution has been equivalent between homoeologs, that gene expression levels have been maintained or averaged relative to that of the diploid progenitors, and that gene co-expression relationships have also been maintained or averaged relative to the diploid progenitors. Alternatively, we may find faster rates of evolution in one subgenome over the other and drastic alteration of gene expression and co-expression patterns in the polyploids. Toward this end, we performed targeted sequence capture of anthocyanin biosynthesis genes on a set of *Gossypium hirsutum* L. and *Gossypium barbadense* L. polyploids, as well as modern representatives of their progenitor genomes, *G. raimondii* Ulbr. and *G. arboreum* L. We examine evolutionary rates following polyploidy among genes and subgenomes, and by position (upstream, downstream) in the pathway. We also evaluated RNA-seq data for the same species, to examine the effects of polyploidy on gene expression and gene co-expression. We find that while genic evolution has favored a faster rate of substitutions in the A-subgenome over the D-subgenome, evolutionary rates are correlated between the two. Many of the genes in the ABP pathway show co-expression and are grouped together in the same co-expression module, both when considered in aggregate and as homoeologs. Evolutionary rates of ABP genes are correlated with expression, module assignment, and co-expression connectivity. In the co-expression network overall, diploid expression patterns are generally well preserved, but homoeolog expression bias and non-additive expression are found throughout the modules.
Materials and Methods

**Bioinformatic identification of *Gossypium* ABP genes**

As described in Grover et al. (2015a), genes from the ABP were identified using OrthoMCL (Li et al. 2003). Proteomes of *Arabidopsis thaliana, Brassica rapa*, *Glycine max* (L.) Merr., *Gossypium raimondii* Ulbr., *Malus domestica, Medicago truncatula, Populus trichocarpa* Torr. & A. Gray, *Solanum lycopersicum* L., *Theobroma cacao* L., and *Vitis vinifera* L. were downloaded from Phytozome and were used to generate clusters of orthologous groups (Goodstein et al. 2012). *G. raimondii* genes in clusters that contained the *A. thaliana* ABP genes, as identified from NCBI GenBank, were called as being putative ABP genes (NCBI Resource Coordinators 2017).

**Plant materials, DNA extraction, and targeted sequence capture**

*Gossypium arboreum* L., *G. raimondii* Ulbr., *G. hirsutum* L., and *G. barbadense* L. plants were grown under a 10-hr photoperiod in the Pohl Conservatory at Iowa State University (Table 1). DNA was extracted from young leaves using the Qiagen DNEasy Plant Mini Kit (https://www.qiagen.com/us/). Library construction, target gene hybridization, and DNA sequencing followed the MYbaits version 2 protocol from Mycroarray (Ann Arbor, MI, USA) and were conducted as described in Grover et al. (2017). Briefly, Illumina libraries were generated by the Genomics Core Facility of West Virginia University (http://genomics.as.wvu.edu/) and the Iowa State University DNA Facility (http://dna.biotech.iastate.edu/). Illumina libraries were captured as per the MYbaits version 2 protocol and sequenced by the Beijing Genomics Institute (BGI, Hong Kong, China) as PE 150 on the HiSeq 2500 (NCBI PRJNA351864); Further details can be found in Grover et al. (2017), and a list of targeted ABP genes can be found in Table 2.
DNA sequence data processing

Raw Illumina reads were filtered and DNA sequences were trimmed via sickle and mapped to the reference *G. raimondii* genome using the SNP-tolerant GSNAP and with a cotton-specific SNP database (Wu and Nacu 2010; Joshi and Fass 2011; Paterson et al. 2012; Page et al. 2013; Page et al. 2014). For polyploid accessions, reads were partitioned into subgenomes using polyCat and the SNP database (Page et al. 2013). Assembled gene sequences were extracted from mapping files using bam2consensus (Page et al. 2014).

Rates of evolution

Tajima’s 1D relative rate test was performed in R by utilizing the seqinr package (Tajima 1993; Charif and Lobry 2007). Trifurcating trees were formed with *G. arboreum*, *G. raimondii*, and one gene candidate from either the A- or D-subgenome; the alternate diploid was used as the outgroup for the test. Muse and Weir’s relative rate test was performed in HyPhy under the Kimura 2-parameter model, with model parameters being calculated globally (Muse and Weir 1992; Pond et al. 2005). Trifurcating trees were formed in the same way as described above for Tajima’s 1D relative rate test. Multiple comparisons correction were made with the Bonferroni correction.

Synonymous (*K_s*) and nonsynonymous (*K_a*) rates of evolution were calculated in R using the kaks function of the seqinr package (Charif and Lobry 2007; R Core Team 2015). Since it previously has been shown that there are different average rates of substitution between the diploid and their respective subgenomes, the values for the evolutionary rates comparing *G. raimondii* to the D-subgenomes of polyploid species were normalized per previously recorded values for *K_a* and *K_s* from Flagel et al. (2012). These rates were then compared using the Wilcoxon rank sum test in R (R Core Team 2015).
Pathway features were examined in the form of pathway position, by categorizing enzymes from most upstream to most downstream, and those at branch points (Figure 1). Trends in position were examined by looking at tau correlation of the evolutionary rates of genes with indices of their position. The effect of connectivity was examined by grouping enzymes into branching and non-branching enzymes and comparing means of the two groups.

**Gene expression analysis**

RNAs were extracted from 7 am and 5 pm leaf (SD7, SD5), meristems (SDM), pre-anthesis petals (SDP9), and fully opened petals (SDPF) from the A-genome (*G. arboreum*) and D-genome (*G. raimondii*) model diploid progenitors; a synthetic diploid hybrid of these two species (A2xD5); and four accessions of each *G. hirsutum* and *G. barbadense* (Table 1). All samples were collected in triplicate except for the A2 meristem and GB-0303 meristem, which were only collected in duplicate. Plants were kept under the 10-hr day conditions in the Pohl Conservatory at Iowa State University. RNAs were extracted using the Sigma-Aldrich Spectrum Plant Total RNA Kit (Sigma-Aldrich); those with RIN scores greater than 6.0 were sent to BGI for library construction and RNA-seq (BGI Americas, https://www.bgi.com/us/). Resulting reads were mapped to the *G. raimondii* reference genome using the SNP-tolerant mode of GSNAP and the same cotton-specific SNP-indices used for the genomic read mapping (Wu and Nacu 2010; Page et al. 2013). Again, polyploids were parsed into subgenomes with PolyCat (Page et al. 2013; Page et al. 2014). Read counts were calculated using HTseq-count (Anders et al. 2015). RNA-seq data processing, differential gene expression, weighted co-expression network analysis, and gene set enrichment analysis were performed as described in Chapter 3.
Homoeolog expression bias and expression level dominance

Homoeolog expression bias, the difference in expression levels between the two subgenomes of the polyploid, was calculated for each gene using DESeq2 (Love et al. 2014). The number of genes showing differential expression between subgenomes was called and the percent difference calculated for each tissue and accession combination. Additionally, homoeolog bias was separated into each module for each tissue and accession combination.

The combined expression can be considered in relation to both the progenitor species and a calculated midparent value (Grover et al. 2012b) and categorized as additive (i.e. matching the midparent value or progenitor expression levels when these are equal), or non-additive (i.e. different from the midparent value and exceeding expression of one or both of the diploids). Non-additive expression patterns were further classified into expression level dominance (expression level matches one, but not both, diploids, and is distinct from the midparent value), and transgressive expression (expression is significantly greater or less than both diploids). DESeq2 was used to calculate the differential expression for each gene and categorize it as described above (Love et al. 2014).

Correlation of evolution of anthocyanin biosynthesis pathway genes with gene expression and co-expression networks

Evolutionary rates ($K_a$, $K_s$, and $K_a/K_s$) for each gene from the anthocyanin biosynthetic pathway were tested for correlation with gene expression, gene co-expression modules, and gene co-expression connectivity. Kendall’s tau rank correlation was used to correlate expression and connectivity with the evolutionary rates; an ANOVA was used to test the effect of the module assignment of each gene on evolutionary rates.
Results

Identification of genes and targeted sequence capture

A total of 28 candidate genes in the *Gossypium* anthocyanin biosynthesis pathway were bioinformatically identified (Table 2). For most genes, only one or two candidates were identified in the D5 reference genome, but several genes had a higher number of putative copies. Chalcone synthase (CHS), flavanol synthase (FLS), and basic helix-loop-helix (bHLH) transcription factors were identified as having six, three, and three copies respectively (Table 2).

Targeted sequence capture was conducted on multiple *Gossypium* species and accessions, including 37 *G. hirsutum* and 8 *G. barbadense* representatives spanning the wild to domesticated continuum (Table 1). Examining the sequence recovery of targeted genes, 17 of the 28 ABP genes were recovered with at least 90% coverage of exonic regions in all accessions. One accession had more than 4 genes with 10% or greater ambiguous sites in the exonic sequence (TX786), indicating a lower capture success rate. Two genes also exhibited lower success capture, as indicated by 10% or more ambiguous nucleotides in most accessions (Gorai.008G121600, a bHLH candidate; Gorai.012G001200, a WD40 repeat candidate); these two genes were excluded from further analyses. An anthocyanidin synthase (ANS) gene candidate, Gorai.008G242100, had 29 sequences with higher than 10% ambiguous sites due to poor assembly at the 3’ end of the coding sequence; this region of the gene was trimmed for further analyses. Within gene alignments, sequences with more than 20% ambiguous sites were eliminated. For most alignments, this was required for only one or two sequences; however, for FLS candidate Gorai.012G026100, all the A-subgenome copies from *G. barbadense* were eliminated. To verify that these missing sequences were not contained in the reads that either (1) were unassignable to the A- or D-subgenome or (2) contained both A and D SNPs, we de novo
assembled these read partitions for *G. barbadense*. As no further sequences were recovered, it is likely that either this homoeolog was lost in the A-subgenome or gene conversion has overwritten the original homoeolog. This gene was also removed from further analyses for *G. barbadense* only. Thus, 52 genes, representing 26 homoeologous gene pairs, were used for further analysis in *G. hirsutum*; 51 genes, representing 25 homoeologous gene pairs and one unpaired gene, were used for *G. barbadense*.

**Comparison of evolutionary rates of genes between diploids and polyploids**

The number of unique substitutions in genes from the diploids and polyploids was calculated using seqinr (Fig. 2; Charif and Lobry 2007). Almost all genes showed differential accumulation of substitutions between the diploids and the corresponding polyploid subgenome. Eleven genes exhibited more substitutions in the A-subgenome of the polyploid than in *G. arboreum* (using *G. raimondii* as an outgroup), nearly identical in number to the twelve that exhibited more substitutions in *G. arboreum*; one gene did not exhibit any changes in either lineage (Fig. 2, red bars). In the *G. raimondii*/D-subgenome comparison (rooted by *G. arboreum*), fourteen genes showed more substitutions in the D-subgenome of the polyploid, and eleven had greater substitutions in *G. raimondii*; again, one gene did not show variation between lineages (Fig. 2, blue bars).

In order to determine if the genes from the polyploids were evolving at a different rate than the diploids, Tajima’s 1D relative rate test, which statistically tests the differential accumulation of substitutions between the diploid and their corresponding polyploid genome, was performed (Tajima, 1993). The number of unique substitutions between the two diploid progenitors and an A- or D-subgenome homoeolog (using the alternate diploid progenitor as an outgroup) was calculated for each accession and used to perform a $\chi^2$ test. We found that 2 of 26
genes showed at least one instance of deviation from a constant molecular clock for the *G. arboreum*/A-subgenome comparisons whereas none were detected for the *G. raimondii*/D-subgenome comparisons (*p* < 0.001). No genes showed uniform deviation for all accessions, and the two genes with the highest rate of differential substitution accumulation across subgenomes were Gorai.001G087200.A (MYB) and Gorai.004G212200.A (bHLH), which had 6% and 10% of accessions displaying a higher substitution rate in the polyploid gene than in the diploid gene. To corroborate the results from Tajima’s 1D relative rate test, we also performed Muse and Weir’s likelihood ratio test of relative rates using the Kimura 2-parameter model of evolution (Kimura 1980; Muse and Weir 1992). For this test, the same two genes as in the Tajima’s relative rate test each showed higher rate of substitution in the polyploid (*p* < 0.001) for 12% of accessions. Collectively, these data suggest that there is little difference in accumulation of substitutions when comparing the diploid and polyploid species.

**Comparison of evolutionary rates of genes within polyploids**

To test the hypothesis of equal evolutionary rates for the two cotton subgenomes, we calculated *K*<sub>a</sub> and *K*<sub>s</sub> between the A- and D-subgenomes relative to their respective progenitors. Flagel et al. (2012) previously calculated mean *K*<sub>a</sub> and *K*<sub>s</sub> for each subgenome and its respective model diploid progenitor using a large set of expressed sequence tags (EST). They calculated that the *K*<sub>a</sub> of the A-subgenome was 0.002 for both *G. hirsutum* and *G. barbadense*, and that the *K*<sub>s</sub> was 0.006 and 0.005, respectively; likewise, for the D-subgenome, they calculated a *K*<sub>a</sub> of 0.003 for both species and a *K*<sub>s</sub> of 0.010 and 0.009 for *G. hirsutum* and *G. barbadense*. While these numbers are small, for our analysis, we sought to analyze differences in evolutionary rates beyond what is caused by the closeness of the model progenitor species. Thus, we normalized the *K*<sub>a</sub> and *K*<sub>s</sub> values from the D-subgenome based on these genome-wide values from Flagel et
al. (2012) to make them comparable to those values from the A-subgenome. For *G. hirsutum*, 23 of 26 gene candidates showed statistically significant different normalized rates of synonymous substitution (Wilcoxon rank sum test, p-value < 0.001; Figure 3). Of these, 14 had a higher rate of substitution in the A-subgenome and 9 had a higher rate of substitution in the adjusted D-subgenome. For *G. barbadense*, 25 of 26 genes showed statistically significant differences in normalized rates of synonymous substitution (Wilcoxon rank sum test, p-value < 0.001; Figure 4). Sixteen of these showed higher $K_s$ values in the A-subgenome than the adjusted D-subgenome, while 9 showed the opposite trend. Again, these data collectively suggest that rates of evolution are slightly faster overall in the A- than in the D-subgenome.

For nonsynonymous substitution rates, there were fewer gene candidates for which the differences in rate were significantly different. In *G. hirsutum*, of the 21 significantly different comparisons, 15 were biased towards higher $K_a$ in the A-subgenome than in the adjusted D-subgenome, and 6 were biased in the other direction (Wilcoxon rank sum test, p-value < 0.001; Figure 3). In *G. barbadense*, there were 15 statistically significant comparisons between the A- and D-subgenomes, with 13 showing higher nonsynonymous rates in the A-subgenome than in the adjusted D-subgenome (Wilcoxon rank sum test, p-value < 0.001; Figure 4). These tests thus mirror the other tests in showing faster rates, overall, of molecular evolution in the A than in the D subgenome.

**Correlation of pathway features and subgenomes with rates of evolution**

To understand the correlation of evolutionary rate with pathway features in the *Gossypium* anthocyanin biosynthesis pathway, we assigned pathway position values to the genes of the pathway, not including transcription factors since the original hypothesis refers to enzymes in metabolic pathways only, by numbering enzymes relative to position in the pathway (Figure
Kendall’s tau correlation tests reveal that position values correlate poorly with Ks and Ka for the genes in the *Gossypium* anthocyanin biosynthesis pathway. Neither value produced a significant correlation with position for the A- or the D-subgenomes of *G. hirsutum* or *G. barbadense*. Correlations were also computed for whether genes were branch points in the pathway and Ks and Ka; these correlations were not significant for either genome from either species.

Finally, correlations were calculated for Ks and Ka values between subgenomes. Both measures were significantly correlated between the two subgenomes (p < 0.05). This shows that general patterns of evolutionary rates have not changed between the two subgenomes, suggesting that the gene pairs are evolving at rates relatively constant to each other.

**Association of pathway genes with phenotype**

The anthocyanin biosynthetic pathway controls production of anthocyanins throughout the plant, but in cotton, the most obvious resulting phenotype is concentration in the petal spot found at the base of each petal in most members of the genus. In *G. hirsutum*, however, the petal spot has been lost in some accessions. We examined gene sequence alignments for stop codons, indels, or amino acid changes that were shared by all accessions that had lost the petal spot, and were not shared by accessions that still bore the petal spot. While several stop codons and amino acid changes were identified in accessions that lack the petal spot, none of these were exclusive to those accessions; each was shared by an accession that retains the petal spot (Table 3). We also examined the association of Ka/Ks, representing change in the protein sequences of the anthocyanin biosynthetic genes relative to the rate of silent substitution, and the petal spot phenotype of the *G. hirsutum* accessions examined. However, no genes showed a significant correlation between petal spot and Kd/Ks (ANOVA, p < 0.001)
Gene expression across various tissues

To better understand the expression evolution of duplicated pathways in polyploids, we extracted RNA, built RNA-seq libraries, and mapped sequenced reads to the *G. raimondii* reference genome (Paterson et al. 2012). Following mapping, read counts were calculate for each gene model using HTseq-count (Anders et al. 2015). In addition, in polyploids, the mapped reads were also parsed into subgenomes. Reads belonging to each homoeolog were then counted, resulting in read counts for both copies, as well as an aggregated read count. In total, 165 RNA-seq libraries were constructed from 2 diploids, a diploid hybrid, and 8 allopolyploids (Table 1).

Read counts were normalized with the DEseq2 variance stabilizing transformation, and a principal components analysis (PCA) was performed to delineate the structure of the data (Fig. 3; Love et al. 2014). The samples primarily grouped according to collection tissue: fully opened (0 DPA) petals, preanthesis (-4 DPA) petals, meristems from flowering branches, and fully expanded leaves collected at 7 am and 5 pm. Leaves grouped together but showed some slight divergence based on collection time. One sample from a preanthesis petal of *G. arboreum* grouped with the fully expanded petals, despite the other three preanthesis petals *G. arboreum* samples grouping together. This may have been due to an error in labelling or in collecting; regardless, this sample was removed from further analysis.

Gene co-expression network analysis

Weighted gene co-expression networks were constructed using the WGCNA package in R (Langfelder and Horvath 2008). An all sample network was constructed with 36,815 aggregated homoeologous gene pairs as nodes. We first examined these networks for overall patterns in gene co-expression. Both networks were shown to have scale free topology, in that a small number of genes showed high connectivity. Eighty-two gene pairs were identified as hub
genes, genes with high connectivity, in the aggregated network \((K > 0.9)\). In the aggregated network, 36 modules, or clusters of gene pairs with similar expression patterns, were identified. An ANOVA was used to test the significance of each module eigengene (ME), the first principal component of the expression profiles of all module gene members, with each tissue and accession type. Of these modules, 34 were significantly associated with tissue and accession type \((p < 0.05)\). Among these, ABP genes were detected in module 1, 2, 3, 6, 10 and 21 (Table 2).

Previous studies have suggested that modules, clustered subnetworks within the whole co-expression network, are likely to be biologically meaningful (Langfelder et al. 2013). GSEAPreranked analysis was performed on the module membership \((kME)\) values for each module to test enrichment for GO Biological Processes (Subramanian et al. 2005; Paterson et al. 2012). Every aggregate module showed enrichment for at least one GO Biological Process \((P < 0.05, Q < 0.05)\). Many general cellular biological processes were the highest enriched GO terms within the modules (e.g. translational termination, translational elongation, rRNA processing, ribosome biosynthesis). Plant-related GO terms were also highly enriched, including many terms connected to photosynthesis, sucrose biosynthesis, and plant cell walls. Of note for this study, module 31 and module 35 were both enriched for flavonoid biosynthesis, although our specific anthocyanin genes were not found in these modules; this may be due to the small sizes of module 31 and 35 compared to those modules in which the anthocyanin genes are found. The modules that contained the anthocyanin genes showed enrichment for a variety of GO categories (Supplementary Table 1). Module 1, which contained half the ABP genes, including representatives of almost every gene product, was enriched for a large number of biological processes related to translation, ribosome assembly, DNA replication, and other general cellular processes. Module 2 contained several ABP genes and was enriched primarily for terms related
to photosynthesis. Three ABP genes were found in module 3; it was enriched for vesicle transport, cell wall organization, and extracellular polysaccharide biosynthesis. Module 6 contains an F3H, ANS, and FLS, and is enriched for respiratory functions, vesicle transport, glycolysis, and cytoskeleton organization. Module 10 contains translation and transcription-related terms. Module 21 is only significant for retrograde transport. These, and the other functions enriched in these modules, are co-expressed with the ABP genes; they represent those cellular and biological processes most closely associated with the ABP.

**Diploid co-expression networks and preservation in polyploid data sets**

Individual gene co-expression networks were generated for both diploid species, which were compared to each polyploid network constructed with aggregated expression of homoeolog gene pairs. To evaluate how the topology of each parental diploid network is preserved and modified since polyploidization, we computed the preservation statistics $Z_{\text{summary}}$ and $\text{medianRank}$ scores, for diploid modules as compared to the polyploid network. The $Z_{\text{summary}}$ score is a composite metric of various connectivity- and density-based preservation metrics, and a $Z_{\text{summary}}$ score of greater than 10 is the suggested cut-off for strong support of module preservation, while a $Z_{\text{summary}}$ score of less than 2 shows lack of support for module preservation (Horvath 2011). The $\text{medianRank}$ score is also a composite score of connectivity- and density-based preservation metrics and is robust to module size variation. Among the 13 A-genome network modules, modules 11, 12, and 13 were the least preserved, showing $Z_{\text{summary}}$ scores below 10 (and below 2 in the case of module 13) and the highest $\text{medianRank}$ values (Fig. 5). The preservation test of D-genome network modules showed that, out of 16 D-genome modules, modules 6, 13, and 16 showed $Z_{\text{summary}}$ scores below 10; the $\text{medianRank}$ values, however, showed module 16 as being preserved (Fig. 6). The small size of module 16 is most likely the
reason for its low $Z_{\text{summary}}$ preservation score. The only ABP gene to appear in one of the poorly preserved modules was FSII (Gorai.013G141500) in module 6 of the D-genome network.

In addition, we also compared the modular topology of the A and D genome networks against their corresponding subgenome homoeolog gene network in allopolyploids. For both *G. hirsutum* and *G. barbadense*, we found that diploid network modules showed similar patterns of preservation when tested against the subgenome networks as when these modules were tested against the aggregated network, i.e. the same modules showed strong, moderate, and low support. This suggests that the diploid module structure is preserved when considering both the aggregate gene expression and the homoeolog-specific expression.

**Preservation and consensus network construction and analysis**

By constructing a consensus co-expression network for all diploid and polyploid aggregated networks, a similarity measure of network topology, the average preservation correlation, or D-statistic, was calculated for each pair of networks (Table 4). As expected, D-statistics were high between accessions from the same species (0.94—0.96 in *G. hirsutum*; 0.95—0.97 in *G. barbadense*). Between the two polyploid species, a slightly lower level of preservation was observed, with the lowest D-statistic of 0.92 between TX2095 and Pima S-6 and the highest D-statistic of 0.96 between GB-303 and TX2094. The two diploids shared an average correlation preservation of 0.89; Comparing diploids to polyploids, there is a reduction in preservation of correlation, decreasing to between 0.88 and 0.91 for the A-genome, and 0.91 and 0.95 in the D-genome. This shows that there is greater co-expression similarity between the D-genome and the polyploid cottons. The diploid hybrid also showed a higher expression similarity to the D-genome than to the A-genome expression patterns. It showed slightly higher levels of preservation in the comparison with *G. hirsutum*, and showed preservation only slightly
higher than it did for the D-genome for *G. barbadense*. This result suggests that the preservation in expression between the diploid hybrid and each polyploid accession is similar to the level of preservation observed between each polyploid accession within a given species.

To further assess similarity between the accessions, a consensus gene co-expression network was generated from scaled TOMs for each accession. The aggregate consensus network contained 19 modules. These consensus modules were then used for comparisons of homoeolog expression bias and expression level dominance. A consensus co-expression network was also generated for the partitioned data set, in which homoeolog expression was counted separately. Twenty-six modules were generated. Module assignment was compared for homoeologs, revealing that 14,661 genes were found within the same module, roughly half of the homoeolog pairs expressed in this study.

**Homoeolog bias**

Homoeolog expression bias, or bias in total contribution to expression of a homoeologous pair, was calculated for each polyploid genome and the A2 X D5 hybrid as a whole (Suppl. Table 2; Grover et al. 2012b). In all comparisons, there was slight bias toward higher D-subgenome expression. The hybrid showed between 9.2% and 10.1% percent difference favoring expression of the D-subgenome. This bias was reduced in both *G. hirsutum* (4.5% to 6.6%) and *G. barbadense* (6.2% to 7.5%).

Homoeologous biases were further broken down to consider tissue specific biases within each accession (Suppl. Table 2). All comparisons, regardless of tissue or accession, showed homoeolog expression bias in favor of the D-subgenome, except for comparison between the subgenomes in the 7 am leaf of Cascot L-7, a domesticated accession; this tissue showed extreme
bias (27.82% A-biased) in the opposite direction of all the other leaves, even other *G. hirsutum* accessions (4.9% to 8.2% D biased; Suppl. Table 2).

Aside from the Cascot 7 am leaf, all other tissues showed biases in the same direction, with the approximately the same degree of bias. In 5 pm leaf, the biases favored the D-subgenome and ranged from 6.6% in Pima to 10.7% in Maxxa. In meristems, there was more extreme bias, with A2 x D5 showing the highest bias of 20.7% in the direction of the D subgenome. The other accessions also showed higher bias in this tissue, ranging from 10.6% in TX2095, closer to the highest percent in other comparisons, to 16.9% in PhytoGen 76. In preanthesis petals, the A2 x D5 hybrid again shows the highest percentage of bias (13.2%), with GB 0303 showing only 8.1% bias. Finally, in fully expanded petals, there is a relatively low bias percentage, only ranging from 0.9% in the A2 x D5 hybrid to 5.5% in K101. Interestingly, while the A2 x D5 hybrid showed close to the highest level of homoeolog bias in the other comparisons (excluding the outlier of Cascot L-7 7 am leaf), it showed the lowest level of homoeolog bias in the fully expanded petals. These data show that, in general, the allopolyploids and the A2 x D5 hybrid show bias towards higher D-subgenome expression. While the hybrid and the polyploids both have high levels of homoeolog bias, there are more genes showing bias in the polyploids despite the higher percentage of homoeolog bias in the diploid hybrid; these differences must be accounted for by concomitant polyploidy and the 1-2 million years of evolution since allopolyploidization.

Homoeolog bias can also be considered within the network context. For each of the tissues described above, the number of biased genes within each aggregate consensus module was calculated (Figure 7). The heatmap largely mirrors what we observed in the homoeolog bias comparisons above. It reveals, however, that certain modules show bias toward the A-
subgenome. Module 14 displays A-subgenome bias across all the accessions and tissues, while modules 11 and 12 display tissue-specific A-subgenome bias in both leaves, with the addition of meristems for module 12. Modules 17, 18, and 19 shows some more variability in their percent differences in each accession and tissue; this may be expected as these modules are smaller and stochastic differences in homoeolog expression will have a greater effect on the overall percent difference between the two subgenomes.

**Total expression level**

While homoeolog expression bias can inform us as to which subgenome is contributing most to expression, we can also consider the combined expression level of each gene. Here, we calculated the combined expression levels for both homoeologs in all polyploids and our diploid hybrid to categorize gene pairs with respect to their cumulative expression pattern (Fig. 8, Suppl. Table 3). Within each tissue, the highest number of genes was assigned to the conserved category, ranging from 27% to 69%, with the other additive expression categories appearing as the next highest. Transgressive expression was found in between 0% and 12% of genes. In all tissues except 5 pm leaf, the polyploids show higher levels of transgressive expression than the diploid hybrid. Genes categorized as showing expression level dominance (ELD) also show relatively low levels, ranging from 0% to 7% in all ELD categories. When splitting these into subgenome and direction, A-subgenome high ELD shows between 0% and 4%, A-subgenome low ELD shows between 0% and 6%, D-subgenome high shows between 0% and 5%, and D-subgenome low shows between 0% and 7%. Similar to what was observed for the transgressive expression categories, the polyploids generally showed higher proportions of expression level dominance than the diploid hybrid.
To evaluate whether expression level dominance was associated with module structure, we tallied the number of times a gene from a given aggregate consensus module was assigned to each expression category, for all tissues and accessions (Figure 8). We found that, similar to the whole data set, the “conserved” category appeared most frequently for every module except modules 14 and 17, which showed higher numbers of “unassigned” genes. Following these conserved and unassigned categories, the additive expression patterns, with either A or D as the higher value, are the next most frequent categories for most modules; unsurprisingly, given the A-homoeolog bias we observed in module 14, the number of cases where there is D-high additive expression is proportionately lower than observed in other modules. Transgressive expression and expression level dominance were found in all modules.

**Connecting gene evolution and expression in the anthocyanin biosynthetic pathway**

To connect our study of genic evolution with gene expression, we examined the anthocyanin biosynthetic pathway within our expression data set. Within the data set, 51 of the 52 genes were expressed. Only one of the candidate MYB genes, Gorai.001G087200 from the D-subgenome, was found to show no expression within our data set.

Previous studies have connected gene evolution with gene expression, suggesting that genes with higher levels of expression will exhibit lower levels of nonsynonymous substitution, K_a (Yang and Gaut 2011). Here we examined the correlation of these two metrics for the anthocyanin biosynthetic pathway genes. For both subgenomes, we found that the expression of the anthocyanin biosynthetic genes was significantly negatively correlated with nonsynonymous rate (Kendall’s tau, p < 0.05).

Since modules represent clusters of genes with similar expression patterns, and therefore genes that might be acting in conjunction with each other, we hypothesized that genes within the
same modules would show similar evolutionary rates. We tested the membership of the anthocyanin biosynthetic genes within each of the module against the three evolutionary rate categories. We found that there was an effect for each evolutionary rate associated with module for both polyploid species (p < 0.05, Fig. 9). We observed outliers primarily in module 0, which contains genes that are not assigned to modules. Without module 0 in the test data set, we still observed a significant effect of module on the rate of evolution for the genes in the ABP.

We also hypothesized that there would be an effect of connectivity on gene evolution. Genes that are highly connected within the module may exhibit more evolutionary constraint, as change in their protein sequence may be more likely to have widespread effects. We calculated the correlation between network-wide connectivity and evolutionary rates, for each subgenome. Since module 0 represents genes that do not follow specific expression patterns, those genes that are part of module 0 were excluded from this analysis. We found a significant correlation between $K_a$ and connectivity in both the A- and D-subgenome (Kendall’s tau, p < 0.05). $K_s$ was not significant in either comparison. $K_a/K_s$, however, was also significantly correlated with connectivity in both subgenomes (Kendall’s tau, p < 0.05). Contrary to what we hypothesized, we found a positive correlation of connectivity and protein evolution, by means of both $K_a$, and $K_a/K_s$.

**Connecting petal spot phenotype with gene expression**

To better understand the cause of the petal spot loss observed in *G. hirsutum*, differential gene expression analysis of aggregate gene expression was performed. Comparing the petal tissues from the *G. hirsutum* accessions with and without a petal spot, we observed only a single differentially expressed gene in our set of ABP candidate genes: Gorai.007G099300, a bHLH gene. It decreased in expression in *G. hirsutum* accessions without the petal spot, but it was not
one of the genes found with a stop codon in its sequence alignment. This gene is another candidate for the change in petal spot that we observe in *G. hirsutum*.

**Discussion**

Here, we used targeted sequence capture to determine gene sequences across a wide panel of polyploid cottons, from which we were able to calculate evolutionary rates. At the same time, we were able to generate a large gene expression data set from which we constructed a gene co-expression network. With these two data sets combined, we were able to better understand the evolution and expression dynamics of the anthocyanin biosynthesis pathway in cotton, as well as the expression dynamics of polyploid cotton more generally.

**The evolution of biosynthetic pathways in a polyploid**

Previous work has shown that pathway features -- including pathway position, pathway branching, and pleiotropy -- may affect the evolutionary rate for genes in a pathway (Rausher et al. 1999; Lu and Rausher 2003; Rausher et al. 2008; Ramsay et al. 2009; Yang et al. 2009; Olson-Manning et al. 2012). Here, for the first time, the evolutionary rates of genes in a duplicated pathway in a neoallopolyploid were tested for association with those pathway features. We did not find that pathway position or branching were significantly associated with the rate of protein evolution in the polyploid cotton studied here. However, the lowest rates of protein evolution were found in five genes that are found at branch points within the pathway (three CHS, one F3H, and one DFR), all of which make up 3 of the 4 most upstream genes in the pathway. All genes are being expressed, except a bHLH gene from the A-genome (Gorai.001G087200).
We hypothesized that the evolutionary rates of the ABP would equivalent between homoeologous genes. However, when we compared subgenomes, we found that, even after adjusting for previously observed differences in levels of $K_a$ and $K_s$, we still observed higher rates of evolution in the A-subgenome for the majority of genes. This suggests that the A-subgenome ABP is generally evolving faster than the D-subgenome ABP, leading us to predict that the A-subgenome would be more likely to degenerate, subfunctionalize, or neofunctionalize than the D-subgenome. Indeed, the majority of stop codons appear in A-subgenome genes (Table 3). Despite this difference in evolutionary rates, we also observed rate correlation between the subgenomes. This suggests that, even though one subgenome may be evolving at a faster rate, they are both evolving at a consistent rate relative to each other. This may be partly because of pathway features maintaining constraint on certain genes, but could also be due to inherent features of the gene and its genomic landscape, such as local mutation rate variation or recombination rate (Gaut et al. 2011; Yang and Gaut 2011).

We found that the genes of the pathway were distributed across several different modules in the aggregate gene co-expression network, but about half of them were found within aggregate module 1 (Table 2). This is somewhat unsurprising since this is the largest module in the gene co-expression network. Yet, it is still interesting to see that the majority of them grouped together in one module, including at least one member from the majority of the different gene types (Table 2). Previous reports have suggested that co-expression modules can be biological meaningful, and the fact that 13 of the ABP genes were found in this module supports this (Langfelder et al. 2013).

The majority of ABP genes were also found within the same module in both of the diploids. We had hypothesized that the genes within the ABP would show equivalent expression
and co-expression in relation to their diploid progenitors. We tested this hypothesis by comparing the network topology of each diploid with their respective polyploid subgenome expression levels using module preservation statistics. We found that those modules that contain the ABP genes generally show strong preservation, a sign that co-expression relationship is preserved between the diploid and the polyploid genes. The exception to this was the flavone synthase II (FSII) candidate Gorai.013G141500 in the D-genome; this gene was found in module 6 in the D-genome network, which showed weak support of preservation.

Another of our hypotheses was that the evolutionary rates of the ABP would be correlated with their expression and co-expression patterns. We calculated the correlation of the ABP gene evolutionary rates with gene expression itself and with the whole gene co-expression network connectivity of each gene, and we looked at the association of the evolutionary rates with the co-expression modules. Each of these comparisons proved significant, although the direction varied. Expression and evolutionary constraint appear to be correlated, i.e. highly expressed genes have low K_a values, as previously reported (Gaut et al. 2011; Yang and Gaut 2011). Evolutionary rate also appeared to be associated with specific modules, suggesting that some co-expression modules share a similar evolutionary rate (Fig. 9); this effect, however, appears to be relatively small.

Interestingly, the correlation of connectivity and evolutionary rate led to a result we were not expecting: the connectivity of a given gene was positively and significantly correlated with K_a, or nonsynonymous substitution rate, as well as K_a/K_s. Previous reports have suggested that we would expect the opposite, that highly connected genes would be more evolutionarily constrained (Mahler et al. 2017); they, however, examined many more genes throughout the whole genome, while we examined solely those genes that are putatively part of the ABP. In the
context of the pathway, this may mean that those genes that are evolving most quickly are also
the one that are most connected to the other expressed genes in the data set, although not
necessarily the other pathway genes.

This unexpected correlation may also be an effect of polyploidy. In the DDC hypothesis,
we would expect gene copies to degenerate following duplication. Some of these may be lost,
while some may eventually undergo subfunctionalization (Force et al. 1999). Here, we may be
observing that the genes that are highly connected are not subject to the same constraint as
previously. However, while a number of genes showed a higher accumulation of substitutions
than in the diploids, we did not observe a significant increase in accumulation of substitutions
between the diploid and polyploid accessions, as we would expect if the polyploids were actually
accumulating substitutions at a higher rate (Fig. 2). This phenomenon will bear further
investigation as to whether this is pathway-specific, i.e. only found in the ABP, or if it is a
general feature of pathways in polyploidy.

Connecting petal spot phenotype with evolution and expression of the ABP

In our study of the ABP, we attempted to connect the petal spot phenotype with the
evolution and expression of the genes the pathway. Although we do not have a definitive answer,
several possibilities have been recognized. First, we looked at the presence of stop codons within
the ABP genes. While none of the stop codons were unique to those accessions without petal
spots, there are two genes that stick out as possible culprits. The chalcone synthase (CHS)
Gorai.006G000200 has stop codons in both its A and D copies, and thus would appear to be a
likely culprit of petal spot loss. However, this stop codon is present in almost all accessions
studied here, and thus is unlikely to be the source of this loss unless there have been other
changes in the pathway.
Gene expression data leads us to another possible gene responsible for this petal spot loss: Gorai.007G099300. This bHLH gene was the only gene from our ABP gene list that showed differential expression between the G. hirsutum accessions with and without petal spots, and it bears further investigation via more targeted means. As a bHLH gene, it may be one of the canonical transcription factors that controls the expression of the biosynthetic genes within the pathway (Gonzalez et al. 2008).

A final possibility is that we missed the genes responsible for production or lack of pigment in the petal spot. Previous studies in A-genome species have suggested that the petal spot forms early in floral development, prior to the petal stages examined here (Stephens 1948b). If the petal spot loss is due to expression changes, studies of earlier tissues may be required in order to truly assess the loss of the petal spot.

**Polyploidy and gene co-expression networks**

Here, we have examined the gene co-expression network of polyploid cotton over a wide variety of tissues and accessions, including two different polyploid species, their diploid progenitors, and hybrid generated from the two polyploids. We hypothesized that gene co-expression would be maintained between the diploids and polyploids. We found relatively strong preservation of correlation in gene expression between the diploid and polyploid species, supporting our hypothesis, as well as between the diploid hybrid and the polyploids. In addition, we observed a large number of co-expression modules that were preserved between the diploids and polyploids, both when expression is considered as a whole, and when it is parsed into homoeologous genes (Fig. 5, Fig. 6). However, we did observed several modules (11, 12, and 13 in the A-genome, 6 and 13 in the D-genome), that showed weak or no support for preservation.
These show that there is some divergence between the diploid co-expression patterns and those of the polyploids.

Homoeolog bias, the contribution of each subgenome to the total, is an important feature of polyploidy. Cis-regulatory elements that have diverged along with the diploid progenitor species are recombined in the polyploid species, coming to exist in a shared trans-regulatory environment. This can lead to differential contributions to gene expression. In addition, this can change over time. This change can be observed in comparisons with the diploid hybrid, which has a parentage similar to that of polyploid cotton. Here, we observed homoeolog bias favoring the D-subgenome in both polyploid species. The diploid hybrid shared this direction, but was more extreme in the proportion of this bias. Interestingly, the actual number of genes that showed this bias was higher in the polyploids than in the diploid hybrid, and is similar to what is seen in rice (Xu et al. 2014). Previous reports in cotton, however, have suggested a relatively balanced level of homoeolog expression bias, and with fewer significantly biased genes (Yoo et al. 2013; Yoo and Wendel 2014; Zhang et al. 2015); however, these reports have not included all the tissues found here.

We also observed that homoeolog expression bias was found in most modules. While the majority of the aggregate modules showed D-subgenome bias, aggregate consensus module 14 showed a consistent A-homoeolog bias through all genomes and tissues, showing the strongest percent differences in the diploid hybrid, and in the leaf and meristem tissues; aggregate consensus modules 11 and 12 also showed A-subgenome bias, although this was limited to specific tissues (Fig. 7). Previous studies in cotton have shown that genes that show differential homoeolog bias, are enriched for different biological processes. Hovav et al. (2008) showed a similar pattern in cotton, finding that genes enriched for several biological categories had
homoeolog biases that changed throughout cotton fiber development. These modules that show strong, shared expression bias may provide clues to biological processes that are controlled primarily by the genes of one subgenome or the other.

We tested our hypothesis that gene expression levels are equivalent in diploids and polyploids by assigning combined expression level categories. We found that additive categories, where the expression of a gene pair matched the midparent value, were the most common, showing initial support for our hypothesis (Suppl. Table 3). In most tissues, however, we also observed that the number of transgressively expressed genes and the number of genes showing expression level dominance increased in the polyploids relative to the diploid hybrid, similar to what Yoo et al. (2013) reported for leaf.

We further sought to connect our hypotheses regarding equivalent gene co-expression and gene expression by counting these expression level categories for each gene co-expression module. While most modules appeared to maintain relatively similar levels of each category, consensus module 14, which also showed tissue-wide homoeolog expression bias, showed higher levels of A-subgenome expression level dominance when the A-subgenome was more highly expressed, but also higher levels of D-subgenome expression level dominance when the D-subgenome is more lowly expressed. While previously reported that expression level dominance was caused by change in the expression of the homoeolog from the non-dominant parent (Yoo et al. 2013), that would lead to decreased A-subgenome expression and increased D-subgenome expression, counter to what has been observed. However, a high number of additive A-subgenome high expression genes (category 2, Fig. 8), may help to explain this phenomenon. Connecting these combined categories, which show changes on the genic level, with co-
expression modules, which show more global trends, will increase our understanding of how the transcriptional network is altered by polyploidy overall.

Conclusions

Although much remains to be learned about polyploidy, this study broadens our understanding of duplication and its effects on pathway evolution. For the first time, we have examined the evolution of a pathway in a neopolyploid, finding that one subgenome is exhibiting a faster rate of evolution than the other. In addition, we have added to the growing body of research regarding co-expression networks and polyploidy (Pfeifer et al. 2014; Hu et al. 2016), showing here that polyploids generally preserve the co-expression structures of their diploid progenitors, despite active expression changes in the form of homoeolog expression bias and expression level transgression and dominance. Further investigation into the expression dynamics of polyploids will illuminate the cis- and trans-regulatory changes that maintain the preservation of overall gene co-expression and yet allow for the change in homoeolog gene expression that we see here.

Acknowledgements

We thank Josef Jareczek and Anna Tuchin for assistance in the greenhouse. This work was funded by Cotton Incorporated, the National Science Foundation Plant Genome Program, and the Molecular, Cellular, and Developmental Biology Program through grant MCB 1118646. Joseph P. Gallagher is supported by NSF Graduate Research Fellowship DGE1247194.
References


Anders, S., P. T. Pyl and W. Huber. 2015. HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 31: 166-169


*Planta.* 240: 1237-1251


McLendon, C. A. 1912. *Mendelian inheritance in cotton hybrids*. Experiment, GA:


Tables and Figures

Table 1. Species and accessions subjected to targeted sequence capture and RNA-seq. All accessions except the diploid hybrid were subjected to targeted sequence capture. Bold accessions were sequenced via RNA-seq for five different tissues (see Methods). *Gossypium hirsutum* are marked with petal phenotypes taken from CottonGen (www.cottongen.org).

<table>
<thead>
<tr>
<th>Species</th>
<th>Accessions</th>
<th>Petal Spot Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Gossypium arboreum</em></td>
<td>A2-101</td>
<td></td>
</tr>
<tr>
<td><em>Gossypium raimondii</em></td>
<td>D5</td>
<td></td>
</tr>
<tr>
<td><em>G. arboreum</em> x <em>G. raimondii</em></td>
<td>A2xD5</td>
<td></td>
</tr>
<tr>
<td><em>Gossypium hirsutum</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cascot L-7</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Maxxa</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>TX2094</td>
<td>Medium</td>
<td></td>
</tr>
<tr>
<td>TX2095</td>
<td>Medium</td>
<td></td>
</tr>
<tr>
<td>ARK2402</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Coker315</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>CRB252</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>FM958</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>LKT511</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>PM145</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>TAMCOT</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>TM-1</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>TX1009</td>
<td>Medium</td>
<td></td>
</tr>
<tr>
<td>TX1037</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>TX1055</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>TX1107</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>TX1110</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>TX1120</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>TX1182</td>
<td>Deep</td>
<td></td>
</tr>
<tr>
<td>TX1226</td>
<td>Medium</td>
<td></td>
</tr>
<tr>
<td>TX1228</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>TX1236</td>
<td>Deep</td>
<td></td>
</tr>
<tr>
<td>TX1748</td>
<td>Deep</td>
<td></td>
</tr>
<tr>
<td>TX1982</td>
<td>Deep</td>
<td></td>
</tr>
<tr>
<td>TX1988</td>
<td>Deep</td>
<td></td>
</tr>
<tr>
<td>TX1996</td>
<td>Deep</td>
<td></td>
</tr>
<tr>
<td>TX2002</td>
<td>Medium</td>
<td></td>
</tr>
<tr>
<td>TX2089</td>
<td>Medium</td>
<td></td>
</tr>
<tr>
<td>TX2090</td>
<td>Faint</td>
<td></td>
</tr>
<tr>
<td>TX2091</td>
<td>Deep</td>
<td></td>
</tr>
<tr>
<td>Variety</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>TX2092</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>TX44</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>TX480</td>
<td>Medium</td>
<td></td>
</tr>
<tr>
<td>TX665</td>
<td>Deep</td>
<td></td>
</tr>
<tr>
<td>TX672</td>
<td>Medium</td>
<td></td>
</tr>
<tr>
<td>TX786</td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>

*G. barbadense*

- **Phy76**
- **Pima S-6**
- **GB0303**
- K101
- GB0262
- GB0287
- GB0369
- GPS52
Table 2. Anthocyanin genes and homologs in the D5 reference genome. The module to which each gene was assigned is shown for each network. Candidate gene IDs for *G. raimondii* were determined as described in the methods. Mods = modules.

<table>
<thead>
<tr>
<th>Gene Product</th>
<th>Aggr Mods</th>
<th>A2 Mods</th>
<th>D5 Mods</th>
<th>Aggr cMods</th>
<th>Part cMods A</th>
<th>Part cMods D</th>
<th>G. raimondii Gene ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHS</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>Gorai.005G035100</td>
</tr>
<tr>
<td>CHS</td>
<td>21</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Gorai.006G000200</td>
</tr>
<tr>
<td>CHS</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>Gorai.009G250900</td>
</tr>
<tr>
<td>CHS</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>Gorai.009G339300</td>
</tr>
<tr>
<td>CHS</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>8</td>
<td>1</td>
<td>Gorai.011G161200</td>
</tr>
<tr>
<td>CHS</td>
<td>3</td>
<td>1</td>
<td>11</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>Gorai.011G161300</td>
</tr>
<tr>
<td>CHI</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>Gorai.013G023400</td>
</tr>
<tr>
<td>F3H</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>Gorai.007G194700</td>
</tr>
<tr>
<td>F3H</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>Gorai.008G062900</td>
</tr>
<tr>
<td>F3'4'H</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>16</td>
<td>1</td>
<td>Gorai.008G198200</td>
</tr>
<tr>
<td>F3'5'H</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>Gorai.011G134900</td>
</tr>
<tr>
<td>DFR</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>Gorai.009G200600</td>
</tr>
<tr>
<td>DFR</td>
<td>21</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Gorai.010G008900</td>
</tr>
<tr>
<td>ANS</td>
<td>6</td>
<td>1</td>
<td>11</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>Gorai.004G205900</td>
</tr>
<tr>
<td>ANS</td>
<td>2</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>Gorai.008G242100</td>
</tr>
<tr>
<td>UFGT</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>13</td>
<td>1</td>
<td>Gorai.002G066600</td>
</tr>
<tr>
<td>FLS</td>
<td>6</td>
<td>7</td>
<td>11</td>
<td>6</td>
<td>11</td>
<td>11</td>
<td>Gorai.009G243600</td>
</tr>
<tr>
<td>FLS</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>Gorai.012G026100</td>
</tr>
<tr>
<td>FLS</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>8</td>
<td>6</td>
<td>Gorai.012G026200</td>
</tr>
<tr>
<td>FSII</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Gorai.013G141500</td>
</tr>
<tr>
<td>MYB</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>Gorai.001G087200</td>
</tr>
<tr>
<td>MYB</td>
<td>1</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>Gorai.001G087400</td>
</tr>
<tr>
<td>bHLH</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>Gorai.004G212200</td>
</tr>
<tr>
<td>bHLH</td>
<td>1</td>
<td>8</td>
<td>14</td>
<td>13</td>
<td>2</td>
<td>2</td>
<td>Gorai.007G023500</td>
</tr>
<tr>
<td>bHLH</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>Gorai.007G099300</td>
</tr>
<tr>
<td>WD40</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>Gorai.004G125200</td>
</tr>
</tbody>
</table>

Table 3. Stop codons in anthocyanin biosynthesis pathway genes in *Gossypium* polyploids.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Subgenome</th>
<th>Gene Product</th>
<th>Position</th>
<th>Number of Accessions Affected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gorai.001G3134900</td>
<td>A</td>
<td>F3'S'H</td>
<td>445</td>
<td>28</td>
</tr>
<tr>
<td>Gorai.002G0666000</td>
<td>D</td>
<td>UFGT</td>
<td>448</td>
<td>11</td>
</tr>
<tr>
<td>Gorai.004G2122000</td>
<td>A</td>
<td>bHLH</td>
<td>991</td>
<td>4</td>
</tr>
<tr>
<td>Gorai.004G2122000</td>
<td>A</td>
<td>bHLH</td>
<td>1099</td>
<td>12</td>
</tr>
<tr>
<td>Gorai.004G2122000</td>
<td>A</td>
<td>bHLH</td>
<td>1432</td>
<td>8</td>
</tr>
<tr>
<td>Gorai.004G2122000</td>
<td>A</td>
<td>bHLH</td>
<td>1507</td>
<td>6</td>
</tr>
<tr>
<td>Gorai.006G0002000</td>
<td>D</td>
<td>CHS</td>
<td>37</td>
<td>36</td>
</tr>
<tr>
<td>Gorai.006G0002000</td>
<td>A</td>
<td>CHS</td>
<td>163</td>
<td>55</td>
</tr>
</tbody>
</table>

**Table 4.** Average preservation correlation, or D-statistic, based on consensus module eigengenes for each accession

| D5          | 0.892 |
| A2xD5       | 0.907 | 0.937 |
| Maxxa       | 0.889 | 0.918 | 0.951 |
| Cascot      | 0.891 | 0.930 | 0.946 | 0.965 |
| TX2094      | 0.913 | 0.945 | 0.967 | 0.959 | 0.966 |
| TX2095      | 0.890 | 0.936 | 0.955 | 0.943 | 0.951 | 0.957 |
| Phy         | 0.880 | 0.905 | 0.936 | 0.956 | 0.953 | 0.941 | 0.945 |
| Pima        | 0.899 | 0.913 | 0.940 | 0.944 | 0.939 | 0.946 | 0.918 | 0.949 |
| GB          | 0.898 | 0.934 | 0.957 | 0.957 | 0.955 | 0.961 | 0.952 | 0.960 | 0.955 |
| K101        | 0.890 | 0.912 | 0.940 | 0.946 | 0.940 | 0.941 | 0.945 | 0.968 | 0.963 | 0.965 |
Figure 1. The anthocyanin biosynthesis pathway. Blue boxes represent those genes that are considered branching enzymes. Numbers correspond to pathway position. Flowers on the right show some of the variation in petal color and petal spot seen in *Gossypium*. 
Figure 2. The average number of substitutions between a subgenome and its respective diploid progenitor. The more negative, the greater the number of substitutions in the polyploid subgenome. The more positive, the greater the number of substitutions in the diploid progenitor. Red = A-subgenome vs. *G. arboreum*, blue = D-subgenome vs *G. raimondii*. 
**Figure 3.** Evolutionary rates for subgenomes of *G. hirsutum*. The top graph shows $K_a$ comparison between A-subgenome and D-subgenome, while the bottom graph shows $K_s$ comparison; both after adjustment based on previously reported evolutionary rates. Stars denote where the two subgenome were significantly different in their rates of substitution.
**Figure 4.** Evolutionary rates for subgenomes of *G. barbadense*. The top graph shows $K_a$ comparison between A-subgenome and D-subgenome, while the bottom graph shows $K_s$ comparison; both after adjustment based on previously reported evolutionary rates. Stars denote where the two subgenome were significantly different in their rates of substitution.
Figure 5. Module preservation of A-genome *G. arboreum* gene co-expression network modules in the aggregated data set for *G. hirsutum* and *G. barbadense*. Red dashed line, $Z_{\text{summary}} = 10$, blue dashed line, $Z_{\text{summary}} = 2$. Module preservation of A-genome *G. arboreum* modules in the partitioned data set was similar and is not shown here.
Figure 6. Module preservation of D-genome *G. raimondii* gene co-expression network modules in the aggregated data set for *G. hirsutum* and *G. barbadense*. Red dashed line, $Z_{\text{summary}} = 10$, blue dashed line, $Z_{\text{summary}} = 2$. Module preservation of D-genome *G. raimondii* modules in the partitioned data set was similar and is not shown here.
Figure 7. Homoeolog bias by consensus module and tissue and accession. Blue denotes higher expression of the D-subgenome, while red represent higher expression of the A-subgenome. SD5 = 5pm leaf, SD7 = 7 am leaf, SDM = meristem, SDP9 = preanthesis petals, SDPF = fully expanded petals.
Figure 8. Expression level categories for each module. The bar graphs show the relative expression of the A (red), D (blue), and polyploid species for each category. For each consensus module, the total number of times the category appeared was calculated. The numbers of each category on the table correspond with the numbers on the bar graphs.
Figure 9. Evolutionary rates of ABP genes in each partitioned consensus module. ANOVA were calculated for each evolutionary rate and module assignment. Test statistics and p-values are reported. *G. hirsutum* is represented in the red, *G. barbadense* is represented in the blue.
Supplementary Files

Table S1. Biological Process Gene Ontology terms enriched in modules containing ABP genes
Table S2. Homoeolog bias across tissues, accessions, and genomes.
Table S3. Combined expression level categorization for each accession and tissue.
CHAPTER 6. CONCLUSIONS

Here, I review the research objectives I put forth in Chapter 1, and discuss how each chapter has contributed to achieving these objectives.

1. To understand how genes in a pathway evolve following polyploidy, especially compared to the current understanding of pathway features and their effect on evolution

   In Chapter 5, I profiled the evolution of genes in the ABP pathway as a model biosynthetic pathway to study the effects of polyploidy on pathway evolutionary rates and dynamics. I found that, despite previous reports regarding the evolutionary patterns of diploids being correlated with pathway position and branching, these correlations did not hold true for polyploid cotton. Additionally, I found that the genes from one subgenome were generally evolving more quickly than those from the other, even after adjusting for expected differences. While this may seem like a feature of polyploidy, it may also just be a feature of polyploid cotton, or of this specific pathway. Neither subgenome showed significantly different rates of substitution from its diploid progenitor, so it does not appear that this is an effect of polyploidy. Expansion into further pathways or more polyploid species may further elucidate this possibility.

2. To understand how polyploidy affects the evolution of gene expression, both for genes in a pathway, and for the whole gene co-expression network

   In Chapter 2, I review the possible changes to expression and evolution caused by polyploidy, providing a short example of how polyploidy and gene co-expression networks might be combined by focusing on the profilin family of genes. In Chapter 4, I then extend that to a more global view, looking at the whole transcriptomic co-expression network in fibers,
rather than a single protein family of interest. The fiber gene co-expression network analysis showed that, when expression is split into partitioned data sets, the topology from the aggregated network was generally preserved; only a few modules showed weak preservation. Despite this overall preservation, I did observe slightly higher D-homoeolog bias across the majority of modules. The module that exhibited the most extreme homoeolog bias, however, was highly biased towards the A-subgenome. This same module was strongly preserved in the A-subgenome, and weakly preserved in the D-subgenome. Here, using gene co-expression analyses, I find that the subgenomes of a polyploid generally share gene co-expression relationships; where these subgenomes deviate in network topology, strong homoeolog expression biases are found throughout the module.

In Chapter 5, the anthocyanin biosynthesis pathway was chosen as focus. Genes of the pathway were examined in the context of the gene co-expression network, and were found in only a few co-expression modules. Many of the genes were found in a single module, showing shared expression patterns. Further analyses tested whether the evolutionary rate of genes in the pathway might be associated with module assignment or correlated with connectivity. Both features were significantly associated with the gene in the pathway. Thus, genes that share co-expression patterns in polyploids may be evolving at similar rates.

Whole transcriptome gene co-expression networks were also studied in Chapter 5. Similar to Chapter 4, modules of the diploid progenitor species were preserved in polyploid species, whether their genes are partitioned or not; again, only a few modules showed weak or no support. This demonstrates that module topology is well preserved between diploid and polyploid species. Analyses of homoeolog biases and expression level in Chapter 5 show these phenomena can be module-specific, which supports the finding in the fiber gene expression data
set in Chapter 4, as well as accession specific, in the case of Cascot L-7. Here, I find that polyploidy preserves the general gene co-expression patterns of the diploids; however, homoeolog biases and non-additive expression still arise, and may be module specific. This supports the idea that the genes of these modules may be co-regulated within the genome, and that evolution of this expression regulation may be affecting their expression as a whole.

Further studies into expression may now be performed with the addition of another replicate of nature’s cotton polyploid experiment: *Gossypium stephensii*. In Chapter 3, we describe and name this species. Aside from the evolutionary interest and novelty in describing a new species, it may also bear traits of agronomic or ecological interest; indeed, it is already known to be quite salt resistant. Examining its gene co-expression and expression evolution, in comparison with other polyploid cottons, can confirm or refute hypotheses about the effects of polyploidy in cotton and other species.

3. **To understand the effects of selection on the duplicated gene co-expression network, using domestication as an exemplary selective force**

   In Chapter 2 and 4, I specifically focus on the effects of domestication on the fibers of polyploid cotton. In Chapter 2, the focus is on the profilin gene family, and targeted analysis reveals that homoeolog bias toward the D-genome has been caused by the process of domestication in several of those genes that are most closely connected to profilin. This shows one approach, of focusing on known genes of interest, and reveals how the gene co-expression network is altered in the course of domestication, with respect to profilin, in the form of homoeolog bias in the connected genes.

   In Chapter 4, I looked for changes wrought by domestication at a larger scale by examining whole network patterns in wild and domesticated fiber. Using the same module
preservation tools that I used to compare diploid and polyploid to compare wild and domesticated networks, I found that there were large changes between the two, and that these were amplified when that partitioned network was used. Overall D-homoeolog expression bias was found in both the wild and domesticated cotton fibers, although the strength varied between the two fibers. Using differential correlation and expression together, I identified a number of genes related to fiber development that have been altered by domestication, including transcription factors, xyloglucan transferase/hydrolases, reactive oxygen species sequestration proteins, and ethylene-related proteins. Here, I show how domestication has altered the duplicated transcriptional network via strong, human-mediated selection for long, spinnable fibers.