Gene Expression Evolution Following Hybridization and Genome Duplication in Cotton (genus Gossypium)

Lex Flagel

Iowa State University

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

Part of the Ecology and Evolutionary Biology Commons

Recommended Citation
https://lib.dr.iastate.edu/etd/10897
Gene expression evolution following hybridization and genome duplication in cotton (genus *Gossypium*)

by

Lex Evan Flagel

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

DOCTOR OF PHILOSOPHY

Major: Genetics

Program of Study Committee:
Jonathan F. Wendel, Major Professor
Fredric J. Janzen
Dennis Lavrov
John Nason
Dan Nettleton

Iowa State University
Ames, Iowa
2009
I dedicate this work to:

My family,

my parents Janet and Kevin Flagel,

my brother Dack Flagel,

and also to my fiancée Suzanne McGaugh.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>vi</td>
</tr>
<tr>
<td>CHAPTER 1.</td>
<td>1</td>
</tr>
<tr>
<td>GENERAL INTRODUCTION</td>
<td></td>
</tr>
<tr>
<td>Dissertation Organization</td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>5</td>
</tr>
<tr>
<td>Objectives</td>
<td>17</td>
</tr>
<tr>
<td>CHAPTER 2.</td>
<td>19</td>
</tr>
<tr>
<td>DUPLICATE GENE EXPRESSION IN ALLOPOLYPLOID GOSSYPiUM</td>
<td></td>
</tr>
<tr>
<td>REVEALS TWO TEMPORALLY DISTINCT PHASES OF EXPRESSION EVOLUTION</td>
<td></td>
</tr>
<tr>
<td>Abstract</td>
<td>19</td>
</tr>
<tr>
<td>Introduction</td>
<td>20</td>
</tr>
<tr>
<td>Methods</td>
<td>22</td>
</tr>
<tr>
<td>Results</td>
<td>26</td>
</tr>
<tr>
<td>Discussion</td>
<td>29</td>
</tr>
<tr>
<td>Conclusion</td>
<td>35</td>
</tr>
<tr>
<td>References</td>
<td>37</td>
</tr>
<tr>
<td>Tables and Figures</td>
<td>43</td>
</tr>
<tr>
<td>Supplementary Material</td>
<td>48</td>
</tr>
</tbody>
</table>
CHAPTER 3. 55

EVOLUTIONARY RATE VARIATION, GENOMIC DOMINANCE, AND DUPLICATE GENE EXPRESSION EVOLUTION DURING ALLOTETRAPLOID COTTON SPECIATION

Abstract 55

Introduction 56

Materials and Methods 59

Results 63

Discussion 67

References 70

Tables and Figures 78

Supporting Information 83

CHAPTER 4. 85

RECIPROCAL SILENCING, TRANSCRIPTIONAL BIAS AND FUNCTIONAL DIVERGENCE OF HOMOELOGS IN POLYPLOID COTTON (GOSSYPIUM)

Abstract 85

Introduction 86

Materials and Methods 89

Results 93

Discussion 100

References 108

Tables and Figures 118

Supplementary Material 136
CHAPTER 5.  143

COORDINATED AND FINE-SCALE CONTROL OF HOMOELOGOUS GENE
EXPRESSION IN ALLOTETRAPLOID COTTON

Abstract  143
Introduction  144
Materials and Methods  145
Results and Discussion  147
References  150
Figure  154

CHAPTER 6.  156

GENERAL CONCLUSIONS

LITERATURE CITED  162

ACKNOWLEDGMENTS  170
ABSTRACT

Polyploidy, or whole genome duplication, is a common feature among plants, occurring in possibly more than half of all angiosperm species. Decades of research has revealed that polyploidy can have profound impacts on phenotypic, genomic, and epigenetic traits. Some changes associated with polyploidy arise immediately from the “genomic shock” caused by duplicating and combining genomes, while other changes result from long-term evolutionary processes that operate on duplicated and redundant genes and genomes. Within this context, this dissertation specifically explores allopolyploidy (polyploidy involving divergent genomes) and its effects on members of the cotton genus (*Gossypium*).

The work in this dissertation focuses on gene expression evolution, utilizing natural and synthetic *Gossypium* allopolyploids and a F₁ hybrid to characterize the expression changes contributed by various stages of the allopolyploidization process.

From these efforts, we find significant levels of expression evolution among the *Gossypium* species, which all include a maternal “A-genome” and a paternal “D-genome”. In the course of this research we have revealed several surprising results, for example that genes from the D-genome of the F₁ hybrid and allopolyploids are more often over-expressed relative to the A-genome, when they are compared to the ancestral condition, meaning that genomic merger and allopolyploidy in *Gossypium* has the net effect of creating D-genome expression biases. Furthermore, for some genes we find cases where this expression bias has gone to completion, leading to total silencing of expression from one of copies of the merged genomes. We also show evidence that *cis*-regulatory changes are a primary contributor to
expression differences between the A and D *Gossypium* genomes. Finally, by analyzing linked genes along two genomic locations we find that proximity can play a role in constraining expression evolution, though we also show that this proximity effect is not universally true.

Beyond these unexpected findings, we also capitalized on the well-understood phylogenetic framework of the *Gossypium* species to place expression evolution in a temporal context. We find that there are immediate effects associated with the genome merger, and that these effects may explain approximately one-quarter of the expression biases found among the natural allotetraploid species, or to put it another way, one-quarter of the expression alterations found in the 1-2 million year old allotetraploids happened immediately upon their formations. Furthermore, among all five natural allotetraploids, we find that expression biases tend to become more extreme, indicating that when maintained in duplicate, the expression profiles of *Gossypium* genes tend to diverge. Finally, between the five diversified natural *Gossypium* allotetraploids we find the most extreme expression evolution in *G. tomentosum*, a wild species endemic to the Hawaiian Islands, followed by *G. barbadense* and *G. hirsutum*, the two allotetraploids cottons domesticated for fiber production. These results indicate that natural histories may contribute to expression biases, including domestication and island colonization.
CHAPTER 1.

GENERAL INTRODUCTION

Dissertation Organization

This dissertation is separated into six Chapters. The first Chapter provides a general introduction to polyploidy, followed by an introduction to the evolutionary theory that predicts the types of changes that may be associated with polyploidy and empirical results that demonstrate that polyploidy is a major evolutionary force. Following this, I introduce the *Gossypium* system and its utility for experiments that answer basic questions at the heart of our evolutionary understanding of polyploidy. Finally, after laying out the background theory, current knowledge in the field, and the study system, I describe six primary objectives of this thesis, which will be answered in turn by the four ensuing research Chapters. Portions of this introductory chapter have been adapted from my contributions to two review papers, (Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF. 2008. *Annual Review of Genetics* 42: 443-461 and Flagel and Wendel. 2009. *New Phytologist* 183: 557-564).

Chapter 2 discusses our exploration of gene expression evolution following genome merger and genome duplication, by using a *Gossypium* microarray to analyze gene expression in a synthetic *Gossypium* F_1 hybrid and natural allotetraploid. This chapter was published in the journal *BMC Biology* in 2008. I undertook this research with Joshua Udall, Dan Nettleton, and Jonathan Wendel. Josh, Jonathan and I designed the experiment. Josh developed and troubleshooting the microarray platform used in this
Chapter 3 expands on Chapter 2, by analyzing the four additional allotetraploids, which can be used to place gene expression evolution following polyploidy in an evolutionary context. This Chapter is in preparation for a special edition of the journal New Phytologist, which will focus on recent findings among the polyploid plant research community. This paper is due for submission shortly, and is in a “nearly submission ready” draft form. Jonathan Wendel and I are the authors on this Chapter. We have worked closely together in developing this manuscript, I performed the research and analyzed the data and Jonathan and I have both had a hand in all other stages of the process.

Chapter 4 expands on our findings from Chapters 2 and 3 by analyzing gene expression a large number of tissues and developmental time-points. This study leveraged a mass-spectrometry based platform for measuring gene expression, which complements the microarray used in Chapters 2 and 3, in that it can be used to process a large number of biological samples for a modest number of genes very cost effectively. Using multiple species and data from multiple tissues and developmental time-points we are able to make a number of important comparisons, such expression in the same gene between different species, different genes within the same species, and the same gene among different tissues within the same species. This paper was published in the journal
In 2009, I undertook this research with Bhupendra Chaudhary, Bob Stupar, Joshua Udall, Neetu Verma, Nathan Springer and Jonathan Wendel. Bhupendra and I were co-first authors. Bob and Nathan designed the mass-spectrometry platform for use in maize and Josh, Bob, and I modified it to work in cotton. Bhupendra, Jonathan and I designed the research. Bhupendra, Neetu and I harvested the plant materials. Bob and I manage the mass-spectrometry workflow, and I analyzed the data with help from Bhupendra and Bob. Bhupendra and I drafted the manuscript with help from Jonathan and all authors offered helpful edits and comments.

Chapter 5 uses the same mass-spectrometry technique as Chapter 4, however the genes analyzed in this chapter were specifically chosen because that can be found on two sequenced portions of the *Gossypium* genome. Thus the expression patterns arising from this research can be tethered to the physical proximity of the genes in question. This research was performed with Liping Chen, Bhupendra Chaudhary, and Jonathan Wendel and was published in the *Journal of Heredity* in 2009 as a research note (a shorter contribution). Jonathan, Bhupendra, and I planned the project. Bhupendra, Liping and I performed the research, I analyzed the data and Jonathan and I drafted and edited the manuscript.

Chapter 6 provides a brief general conclusion and answers the six primary research objectives in light of the previous four original research Chapters (2-5), followed by Acknowledgments and the Works Cited by the Introduction and Conclusion. Finally,
each of the original research Chapters is presented in the format required by the specific journal it targeted.
Introduction

Polyploidy

Gene redundancy is a major feature of the genomic organization among plants. At the level of the chromosome, gene redundancy occurs as duplicated alleles in diploidy, which in many cases are re-duplicated by polyploidy (whole genome duplication). Estimates of the prevalence of polyploidy among the angiosperms (the flowering plants) range from 30-70% (Wendel, 2000; Soltis et al., 2004), which is a greater incidence than any other group of organisms. In addition to this high rate of contemporary polyploidy, genomic analyses have shown that modern angiosperm genomes were built from the remnants and resolutions of ancient genome duplications (paleopolyploidy). These ancient duplications can clearly be identified by the large gene families and segmentally duplicated chromosome regions they have left behind (Bowers et al., 2003; Yu et al., 2005; Barker et al., 2008; Tang et al., 2008; Soltis et al., 2009). Recognition of the evolutionary implications of genome duplications and the gene redundancy it creates first occurred over fifty years ago (Stebbins, 1950; Stephens, 1951). More recently, with the advent of genomic tools, a great amount of empirical and theoretical work has been conducted to gain an understanding of the evolution of duplicated genes following polyploidy within the angiosperms. The work collected in my dissertation sheds some light on aspects of duplicate gene evolution, specifically the rate and tempo of duplicate gene expression evolution in cotton, and some of the causes of this evolution.

Evolution via gene and genome duplication, background theory
A notable feature of duplication, when compared to other forms of mutation, is that it creates genetic redundancy. This redundancy has long been thought to foster evolutionary innovation, as the constraints of purifying selection are expected to be relaxed on duplicate loci thereby creating the opportunity for duplicates to explore new evolutionary terrain. Though this concept did not originate with Ohno (for a history see Taylor & Raes (2004)), it was broadly popularized in his book *Evolution by Gene Duplication* (1970). In Ohno’s classic formulation, if given sufficient time, one copy of a duplicate pair can acquire a beneficial mutation (*neofunctionalization*) resulting in retention of both divergent copies. Alternatively one copy can accumulate mutation(s) that render it nonfunctional and lead to mutational obliteration (*nonfunctionalization* in Ohno’s words, *pseudogenization* in modern terms), consequently maintaining the other copy through purifying selection. In recent years, with the recognition of the highly duplicate nature of many eukaryotic genomes, these concepts of evolution by duplication have been a source of great interest, leading to the development of a significant body of theory. Here I provide a brief overview of this theory and its connections to polyploidy.

One noteworthy contribution to the theory of evolution by duplication has been the realization that mutations may accumulate among duplicates in such a way that they *partition* aggregate ancestral functions such that both gene copies must be preserved to carry out their complementary ancestral roles. This process, which can arise in the absence of natural selection, has been termed *subfunctionalization* under the Duplication-Degeneration-Complementation (DDC) model of Force *et al.* (1999). It posits a mechanism that creates a stable safe-haven for preservation of both members of a
duplicate pair (Lynch & Force, 2000). Importantly, for DDC-subfunctionalization to occur, the ancestral gene must have had at least two necessary functions (here “functions” is broadly defined to include gene expression, for example). If expression in multiple cell lines, tissues, or organs is necessary for a given protein product, then a duplicate gene pair encoding this protein may experience expression DDC-subfunctionalization by regulatory rather than coding mutations. As we will see later, several of the findings of this dissertation support this viewpoint.

Another important contribution to the theory of evolution by duplication is the recognition that a single protein can perform multiple catalytic or structural functions. This has been famously demonstrated for structural eye crystallin proteins, which also have enzymatic functions when expressed outside the eye (Piatigorsky & Wistow, 1991). For these “shared genes” (shared in the sense of a single gene being employed by unrelated cellular processes), the selective optimization of one function may lead to a decline in another function, creating an adaptive conflict. Under this scenario, gene duplication and subsequent functional specialization between duplicates can provide a solution to the optimization problem. This has been termed Escape from Adaptive Conflict (EAC). EAC can generate observed patterns that may be misconstrued as either neofunctionalization or DDC-subfunctionalization (Des Marais & Rausher, 2008), the important distinction being that DDC-subfunctionalization may occur purely as a result of neutral mutations, whereas EAC requires positive natural selection on both copies of a duplicate gene pair (Conant & Wolfe, 2008; Des Marais & Rausher, 2008). In a similar vein, both neofunctionalization and EAC require positive natural selection, though for
neofunctionalization this selection need only influence one of the two duplicate genes. Consequently, the EAC model has the greatest number of conditions that must be met, and because of this it may occur less frequently than neofunctionalization or DDC-subfunctionalization.

Finally, an important contribution to evolution by duplication theory is the observation that duplications must maintain proper dosage balance among dosage-sensitive genes (Veitia, 2005; Veitia et al., 2008). If a duplication event produces a dosage imbalance in a finely-tuned gene network or protein complex it can lead to a reduction in the functional efficiency of these interactions. Under this scenario, selection against dosage imbalance would favor the return to a single-copy state. As an example, Thomas et al. (2006) proposed that following an ancient polyploidization event in Arabidopsis, chromosomal clusters of interacting dosage-sensitive genes were preferentially preserved during Arabidopsis's return to diploidy, while their homoeologous (genomic content that has been duplicated by polyploidy) clusters were shed. The dosage-balance hypothesis (Veitia et al., 2008) has resulted in an appreciation of the necessity of considering gene duplication in a more interdependent context: for any single locus, duplication may relax selection; however, this relaxed selection may be counterbalanced by other genomic dependencies, such as dosage-sensitive interactions with the products of other genes.

Collectively, the neofunctionalization, DDC-subfunctionalization, EAC, and dosage-balance models form a theoretical framework for understanding evolution
following gene duplication. With regard to polyploid species, wherein the entire
genomic complement has been duplicated, there is a tremendous opportunity for
duplicate gene evolution following the paths outlined by each of these models. Next we
will discuss experimental observations from polyploid plants, and their interpretations
with regard to the predictions of these models

Genomic consequences of polyploidy in plants

Some of the most illuminating studies of polyploid species have focused on
various aspects of genomic integrity. By examining various genome-wide structural
changes, this work has revealed a multitude of interactions that can occur between co-
resident genomes. Work along these lines has been performed in a diverse array of
model plant polyploids of ages ranging from newly formed to ancient, including old
allopolyploids (polyploids combining divergent genomes) from the *Triticum-Aegilops*
species complex (Ozkan *et al.*, 2001; Shaked *et al.*, 2001) and very recently formed
natural allopolyploids in *Spartina* (Salmon *et al.*, 2005) and *Tragopogon* (Tate *et al*.,
2006), and synthetic polyploids in the Brassicaceae (Song *et al.*, 1995, Madlung *et al*.,
2002; Madlung *et al.*, 2005). These studies have revealed that allopolyplody often leads
to unexpected and unexplained departures from predicted genomic additivity. Examples
include gene loss (Ozkan *et al.*, 2001; Chantret *et al.*, 2005; Tate *et al.*, 2006), widespread
modification of methylation patterns (Madlung *et al.*, 2002; Wang *et al.*, 2004; Salmon *et
al.*, 2005), and non-reciprocal chromosomal exchanges (Pires *et al.*, 2004; Udall *et al*.,
2005). From an evolutionary or ecological standpoint these phenomena may be viewed
as novel generators of potentially relevant genomic variation, as demonstrated in *Brassica* for several environmentally important phenotypic characters including flowering-time (Pires *et al.*, 2004), leaf morphology and seed set (Gaeta *et al.*, 2007). Though not directly related to the experiments carried out in this dissertation, many of these genomic perturbations can impact gene expression pathways, which are the primary topic of this dissertation, and the topic we will cover next.

*Gene expression variation arising from polyploidy.*

The alteration of gene expression patterns is a prominent cause of variation within and between species. Gene expression alterations may arise immediately in response to external or internal stimuli, often leading to rapid phenotypic change. Accordingly, it is thought that expression patterns may respond to selection faster than other mutational processes. Indeed, many groups studying diverse taxa have concluded that shifts in timing and intensity of gene expression patterns are among the primary sources, if not the primary source, of developmental novelty (King & Wilson, 1975; Frary *et al.*, 2000; Rifkin *et al.*, 2003; Clark *et al.*, 2004; Carroll, 2005; Clark *et al.*, 2006). Hybridization and polyploidy stimulate an array of changes that have genome-wide effects, which, not surprisingly, include massive alterations in gene expression. Next we will explore these changes from a micro- and macro-perspective, by focusing on gene-specific expression changes and genome-wide shifts in expression patterns associated with hybridization and polyploidy.
A genic perspective.

One of the more spectacular recent revelations with respect to gene expression in polyploids is that homoeologous genes (genes duplicated by polyploidy) commonly make unequal contributions to the transcriptome, as shown most thoroughly to date in cotton and in wheat (Adams et al., 2003; Mochida et al., 2003; Adams et al., 2004; Bottley et al., 2006; Hovav et al., 2008). Adams et al. (2003) demonstrated that 10 out of 40 homoeologs from the A- and D-genome of allotetraploid cotton exhibit biased expression, including several cases of reciprocal silencing among adjacent floral whorls. These and other biases in homoeologous expression can be described as expression subfunctionalization, or the partitioning of ancestral expression domains among duplicate genes (homoeologs in this case). This mode of subfunctionalization is of considerable evolutionary interest, as it may enhance the probability of duplicate gene retention (Lynch & Force, 2000) following polyploid formation. Also, expression subfunctionalization may act far more quickly than the classic conceptualization of subfunctionalization arising from mutations in coding or non-coding regions, which is governed by the slower process of stochastic accumulation of mutations. From an evolutionary perspective, it has been hypothesized that expression subfunctionalization may initially preserve a large number of homoeologous pairs from mutational decay, thus retaining additional raw material for subsequent evolutionary tinkering (Adams et al., 2003). However, it remains unclear if this process facilitated duplicate gene retention or evolutionarily relevant functional diversification. Evidence from Arabidopsis indicates that gene retention following polyploidy (the Arabidopsis lineage experienced its most recent polyploidy event approximately 20–40 million years ago) can, to some extent, be
explained by expression subfunctionalization (Blanc & Wolfe, 2004; Casneuf et al., 2006; Ganko et al., 2007). However, Casneuf et al. (2006) have shown that paleo-homoeologs in Arabidopsis show more highly correlated expression patterns (i.e. less subfunctionalization) than do other types of duplicates, indicating that expression subfunctionalization may have played only a weak role in preserving paleo-homoeologs. Other mechanisms, such as the retention of dosage sensitive genes (Thomas et al., 2006) and the buffering of critical functions through redundancy (Chapman et al., 2006) have also been suggested. Even without knowing the specific mechanism(s) involved, it is clear that there is some disconnect between the observation of high levels of expression subfunctionalization in recent homoeologs of cotton and wheat, and the apparent loss of this subfunctionalization among Arabidopsis paleo-homoeologs. It is difficult to resolve these two contradictory observations; however they may simply reflect lineage-specific differences between these species, or highlight our poor understanding of the processes that fractionate polyploid genomes and return them to a diploid state.

A genomic perspective.

Recent large-scale microarray studies in a range of polyploid plant species have confirmed that gene expression is radically altered by polyploidy (Hegarty et al., 2006; Wang et al., 2006; Stupar et al., 2007; Rapp et al., 2009). The magnitude of effects varies largely between species, but enough data now exist to reveal some general trends. By comparing global gene expression profiles in synthetic allotetraploids with their parental diploid genomic donors, work in Gossypium (cotton) (Rapp et al., 2009) and
*Arabidopsis* (Wang et al., 2006) has specifically addressed the transcriptional effects of combining differentiated genomes, with their divergent regulatory machinery, into a common nucleus. Wang et al. (2006) showed that a synthetic *Arabidopsis* allotetraploid, formed by combining *A. arenosa* with *A. thaliana*, exhibits strong expression dominance of the *A. arenosa* parent, coupled with suppression of the *A. thaliana* genome. The extent of this suppression is impressive; approximately 94% of the genes up-regulated in the *A. thaliana* parent relative to the *A. arenosa* parent are subsequently down-regulated (suppressed) to the level of the *A. arenosa* parent after allotetraploidy. In cotton, a similar story has unfolded. In comparing two synthetic allopolyploids with their parents, Rapp et al. (2009) found substantial dominance of the paternal expression phenotype. Both studies highlight an important and emerging property of allopolyploidy: with regard to expression, many genes do not behave as simple additive combinations of the parental genomes. Indeed, in cotton and *Arabidopsis*, genomic dominance appears to be quite common.

In addition to the effects of hybridization on gene expression during allopolyploid formation, ploidy level also plays an important role. This has been demonstrated by studies of gene expression in accessions of maize with increasing ploidy (Auger et al., 2005) and in *Senecio* (Hegarty et al., 2006). In both study systems, dosage balance (an even numbered ploidy level) was found to play a crucial role in establishing stability in gene expression. When compared to diploids, triploid individuals from both maize and *Senecio* were shown to exhibit radically different and novel expression profiles. Hegarty et al. (2006) took this one step further in *Senecio* by showing that a move from triploidy
to tetraploidy returns the transcription profile back to a state most similar to diploid individuals.

Together, these findings regarding hybridity and ploidy level point to an interplay between the “genomic shock” caused by hybridization and dosage imbalance during allopolyploid formation. As triploidy is often thought to be a necessary “bridge” during allopolyploid formation (Ramsey & Schemske, 1998), and because hybridization is involved in the incipient stages of any allopolyploidization event, it is likely that both dosage balance and the particularities of genomic combination during merger contribute to novel expression phenotypes. Much remains to be learned regarding both the underlying mechanisms of gene expression alteration during hybridization and genome doubling, as well as its short and long-term evolutionary consequences. In the upcoming chapters of this dissertation we will address some of the above issues, using cotton as a model system. In doing so, we will provide some insights on the large-scale changes in genomic expression patterns associated with allopolyploidy and hybridization (Chapters 2 and 3) as well as changes during development among individual genes (Chapter 4).

*Cotton as a model for studying duplicate genes*

Members of the cotton genus (*Gossypium*) provide a model system to study the evolution of duplicate gene expression because they contain five natural allotetraploid species (*G. barbadense, G. darwinii, G. hirsutum, G. mustelinum, and G. tomentosum*), which contain a full copy of the *Gossypium* A- and D-genomes, as well as excellent
models of the ancestral diploid progenitor species (A-genome, *G. arboreum*; D-genome, *G. raimondii*) (see Fig. 1; reproduced following Wendel and Cronn (2003)). An additional advantage of using these natural allopolyploid species as a model is that we have available a diploid synthetic F₁ hybrid (*G. arboreum ♀ X G. raimondii ♂*) and a synthetic allotetraploid (*G. arboreum ♀ X G. davidsonii ♂* (a D-genome diploid closely related to *G. raimondii*)). These synthetics offer the opportunity to study the immediate effects of the merger of the two parental genomes into a single diploid or tetraploid nucleus (and have been utilized for this purpose in Chapters 2-4). This collection of species and experimental synthetics provides a framework for studying the expression of duplicate genes from co-resident genomes on multiple temporal scales, ranging from the onset of hybridization, through the immediate impacts of genome duplication, and onto the longer-term evolutionary consequences now found in the natural allotetraploid species.

As described above, cotton provides an excellent framework for studying expression evolution following polyploidy. Additionally, because it is also a prominent commodity crop, significant investments have been made to develop genomic resources in cotton. With regard to this dissertation, the critical resources have been a large community effort to sequencing and assemble ESTs (Expressed Sequence Tags,
which are sequenced transcribed mRNAs) (Udall et al., 2006). This resource includes ESTs from both model diploid parents (A-genome: *G. arboreum*; D-genome: *G. raimondii*), as well as the allotetraploid *G. hirsutum*. Using this resource, the Wendel lab group has designed complementary microarray and mass-spectrometry platforms, each capable of measuring the proportion of A- and D-genome expression from a combined transcriptome. The two platforms offer contrasting capabilities. The microarray is excellent for measuring a large number of genes in parallel, but cannot be used to measure a large number of biological samples, as this would be prohibitively expensive. On the other hand, the mass-spectrometry platform can be used to measure genome-specific expression in a large number of biological samples, but for only a limited number of genes (~60) before becoming costly. As a consequence of these strengths and limitations, we used the two platforms to answer different questions. In Chapters 2 and 3 we apply the microarray platform to answering broad evolutionary questions about the evolution of duplicate gene expression, whereas, in Chapters 4 and 5 we have used the mass-spectrometry platform to ask targeted questions about specific patterns of homoeolog expression in a large number of tissues and developmental stages.
Objectives

The goal of this dissertation is to further develop our understanding of the evolution of gene expression following whole-genome duplication. This was accomplished using cotton as a model polyploid plant system and recently developed high-throughput technologies, to study gene expression. Because we have well-understood phylogenetic relationships among these cotton species (Fig. 1), an additional novel component of this dissertation was to bring a temporal dimension to bear on the process of gene expression evolution following whole genome duplication. This temporal component was added by studying gene expression within a synthetic hybrid and allotetraploid as well as across the five natural cotton alloployploid species. Within this framework these data have allowed us to address the following questions:

1. What is the temporal pace and scope of gene expression evolution following genome merger and duplication?
2. What is the extent of homoeologous expression evolution?
3. What amount of expression evolution can be attributed to hybridization, and what amount can be attributed to genome duplication and the diversification and evolution of polyploid species?
4. How do these patterns play out among different tissue types or among different developmental time-points with the same tissue?
5. How does genomic linkage (i.e. proximity among genes) influence expression evolution following polyploidy?
6. Finally, what does all of this tell us about the importance of expression evolution following gene duplication?
CHAPTER 2.
DUPLICATE GENE EXPRESSION IN ALLOPOLYPLOID GOSSYPIUM
REVEALS TWO TEMPORALLY DISTINCT PHASES OF EXPRESSION EVOLUTION

A paper published in BMC Biology in 2008 (BMC Biology 6: 16)
Lex Flagel, Joshua Udall, Dan Nettleton and Jonathan Wendel

ABSTRACT
Polyploidy has played a prominent role in shaping the genomic architecture of the angiosperms. Through allopolyploidization, several modern Gossypium (cotton) species contain two divergent, although largely redundant genomes. Owing to this redundancy, these genomes can play host to an array of evolutionary processes that act on duplicate genes. We compared homoeolog (genes duplicated by polyploidy) contributions to the transcriptome of a natural allopolyploid and a synthetic interspecific F1 hybrid, both derived from a merger between diploid species from the Gossypium A-genome and D-genome groups. Relative levels of A- and D-genome contributions to the petal transcriptome were determined for 1,383 gene pairs. This comparison permitted partitioning of homoeolog expression biases into those arising from genomic merger and those resulting from polyploidy. Within allopolyploid Gossypium, approximately 24% of the genes with biased (unequal contributions from the two homoeologous copies) expression patterns are inferred to have arisen as a consequence of genomic merger, indicating that a substantial fraction of homoeolog expression biases occur
instantaneously with hybridization. The remaining 76% of biased homoeologs reflect long-term evolutionary forces, such as duplicate gene neofunctionalization and subfunctionalization. Finally, we observed a greater number of genes biased toward the paternal D-genome and that expression biases have tended to increases during allopolyploid evolution. Our results indicate that allopolyploidization entails significant homoeolog expression modulation, both immediately as a consequence of genomic merger, and secondarily as a result of long-term evolutionary transformations in duplicate gene expression.

INTRODUCTION
A hallmark of angiosperm genome organization is gene redundancy. Redundant genome segments have been identified in the composition and architecture of modern-day angiosperm genomes suggesting one or more ancient genome duplication events (Wendel, 2000; Bowers et al., 2003; Lockton & Gaut, 2005). This has led to considerable interest in the evolution of the resulting duplicated genes. A key issue has been the identification of factors that enhance the retention of duplicate gene pairs and their potential for adaptive diversification or subfunctionalization (the partitioning of ancestral function). Mechanisms such as the maintenance of gene dosage and epistatic interactions (Birchler et al., 2005; Veitia, 2005) and epigenetically regulated expression subfunctionalization (Adams et al., 2003; Rodin & Riggs, 2003) have been implicated in aiding duplicate gene retention. These processes describe mechanisms of retention for ancient duplicate genes and naturally lead to questions about the evolutionary behavior of duplicate gene pairs in more recently formed polyploid species.
Members of the cotton genus provide a phylogenetic framework to study the evolution of duplicate gene expression in recent polyploids because five diverse allopolyploid species are thought to have diverged from a single allopolyploidization event (Wendel & Cronn, 2003), and models of the ancestral diploid progenitor species (denoted by A2 and D5) have been identified (Figure 1A). In addition, extensive genomic resources, such as comprehensive expressed sequence tag (EST) libraries (Udall et al., 2006a), microarray platforms (Udall et al., 2006b; Udall et al., 2007), and BAC libraries () have greatly extended research capabilities. Synthesis of an F1 hybrid, combining the A- and D-genome diploid model species, offers the opportunity to untangle the effects of genomic merger from those arising from genome doubling and subsequent evolutionary change. This phylogenetic framework facilitates the study of gene expression from co-resident genomes on two temporal scales, from the onset of hybridization to a longer-term evolutionary timeframe encompassed by the natural allotetraploid species.

Adams et al (Adams et al., 2003) demonstrated that homoeolog expression in allotetraploid cotton has been strongly influenced by developmentally regulated, organ-specific silencing, resulting in subfunctionalization of the aggregate ancestral expression profile. This subfunctionalization may occur immediately after polyploidization or may arise over a longer period of evolutionary resolution (Adams et al., 2004; Adams & Wendel, 2005). The net effect is a process that appears to impose a form of selective retention on both homoeologs. Thus, expression subfunctionalization leads to prolonged duplicate gene retention, which may in turn enhance the potential for spatial, temporal, or functional divergence of duplicated genes.
Here we employ a novel microarray technology, which uses homoeolog specific probe sets, to assess the relative contribution of 1,383 homoeologous gene pairs to the transcriptome of natural allopolyploid *Gossypium hirsutum* and a synthetic, diploid F1 hybrid (denoted as AD1 and F1, respectively). We show that the two genomes contribute unequally to the total transcriptome of the allopolyploid. By comparing these entities we demonstrate that, for a substantial fraction of the genome, homoeolog expression biases occur immediately with the onset of genomic merger. In addition, a greater number of homoeolog expression biases appear in allopolyploid cotton that likely were not instigated by genomic merger. These findings indicate that upon allopolyploid formation, homoeolog expression biases happen in two, distinct temporal phases.

**METHODS**

**Plant materials, experimental design, RNA isolation, and microarray preparation**

Three replicate blocks of four *Gossypium* accessions (A2 | D5 | A2 ♀ X D5 ♂ F1 | AD1; Table 1) were grown in the Pohl Conservatory at Iowa State University, Ames, IA. These four accessions include representatives of both diploid progenitor genomes (A- and D-genomes) of natural allopolyploid cotton, their synthetic F1 hybrid, and an allotetraploid, respectively (Wendel & Cronn, 2003) (Figure 1A). Petals from all four accessions were harvested on the day of anthesis and three biological replicates were generated by pooling tissues from a minimum of eight flowers obtained from three individuals, or alternatively from a minimum of three flowers from a single individual if multiple individuals were not available (applicable only to F1 hybrid). RNA extractions were performed following a modified hot borate procedure optimized for *Gossypium* (Wan & Wilkins, 1994). All
RNA samples were quantified and visually assessed for degradation and DNA contamination via a Bioanalyzer (Agilent Technologies, Santa Clara, CA). From each pair of A2 and D5 replicates, an equimolar RNA mix (1:1 mix) was made. RNA samples were sent to NimbleGen Systems (Madison, WI), for cDNA synthesis, labeling, and hybridization to 15 microarrays, following proprietary protocols.

**Homoeolog-specific microarray platform**

We have designed and implemented a novel microarray platform capable of measuring homoeolog-specific expression in *Gossypium* species (Figure 1B). The utility of this design has been demonstrated with our first-generation arrays (Udall *et al.*, 2006b), but rapid developments in the depth of cotton EST resources, EST assembly quality, and microarray probe density enabled us to create a second-generation platform, which was used in this study. A description of the microarray design can be found in Additional file 1. This second-generation platform features oligonucleotide probe-pairs near 35 bases in length differing by an A- or D-genome homoeolog-specific single nucleotide polymorphism (SNP) at their middle base (Figure 1C, box). Thus, the microarray platform has the ability to measure expression levels separately for each homoeolog, detected by the corresponding homoeolog-specific probe.

**Statistical analysis**
Raw data values for each microarray were natural log transformed, median centered, and scale normalized across all arrays prior to analysis. For each homoeolog probe pair the difference of natural logs of the A- and D-homoeolog-specific probe was calculated \( (\ln(A_{\text{probe}}) - \ln(D_{\text{probe}})) \); hereafter referred to as log ratio. Using this approach, positive values indicate an A-genome bias, whereas negative values indicate a D-genome bias. A linear model including effects for replication and genotype was fit to the log ratio data from each probe to identify the subset of probes that diagnostically detected homoeolog-specific expression. This was done by filtering for only those probes in which the log ratio for \( A_2 \) was significantly (FDR \( \leq 0.05 \); see (Benjamini & Hochberg, 1995)) and appreciably greater (fold change of at least 1.5) than the 1:1 mix of \( A_2 \) and \( D_5 \), and the 1:1 mix log ratio was significantly and appreciably greater than \( D_5 \) (\( A_2 > 1:1 \text{ mix} > D_5 \)). The resulting, empirically identified, probe sets can diagnose homoeolog-specific expression levels within transcriptionally mixed A- and D-genome hybrid and allopolyploid RNA samples.

Following the identification of all diagnostic probes, contig-level log ratio values were determined by calculating a robust average of the log ratio values from all diagnostic probe sets within a contig using Tukey’s Biweight method. A linear model including effects for replication and genotype was fit to this contig-level data, allowing the estimation of all possible contrasts between \( A_2 \), \( D_5 \), 1:1 mix, \( AD_1 \), and the \( F_1 \) hybrid. The contrasts between the \( AD_1 \) and \( F_1 \) samples and the 1:1 mix allow us to diagnose change relative to the \textit{in vitro} mid-parent value of the \( A_2 \) and \( D_5 \) diploids. In addition, these contrasts account for the specific hybridization kinetics of each probe, when faced with a genomically mixed transcript pool. This is useful, as it can factor out non-linear
competitive interactions that may occur as a result of the interaction between A- and D-genome transcripts.

Given the distributions of $p$-values from AD$_1$ versus 1:1 mix and F$_1$ versus 1:1 mix contrasts, we estimated the expected number of true null hypotheses, using the procedure of Nettleton et al (Nettleton et al., 2006). It was determined that approximately 495 and 884 genes were true nulls, and thus not statistically different in mean log ratio from the 1:1 mix, for AD$_1$ and F$_1$, respectively. Using these estimates from the AD$_1$ versus 1:1 mix and F$_1$ versus 1:1 mix contrasts, we selected an FDR threshold for significance (Storey & Tibshirani, 2003) of 0.15 to strike a reasonable balance between the expected number of false positives and false negatives. FDR significance thresholds of 0.05 and 0.10 were examined as well and can be found in Table S1A, B in Additional file 1.

Using the A$_2$ and D$_5$ diploids as a reference measure of pure A- or D-genome expression gives us the ability to discover cases of genome-specific gene silencing in both AD$_1$ and F$_1$. These putative cases of silencing can be detected as log ratio values that are greater than or equal to the A$_2$ diploid parent or less than or equal to the D$_5$ diploid parent. Using this definition of silencing, we were able to detect gene silencing in both the AD$_1$ and F$_1$ accessions.

Validation of microarray results with Sequenom quantitative mass-spectrometry

Validation of our microarray results was performed for 13 randomly selected homoeologous gene pairs using Sequenom quantitative mass-spectrometry following the
methods of Stupar and Springer (Stupar & Springer, 2006). Aliquots of RNA transcripts used for microarray hybridizations were analyzed for A- and D-genome contributions to the transcriptome for AD₁ and F₁ samples (the validation design can be found in Additional file 1). Briefly, the Sequenom technology amplifies A- and D-derived cDNA transcripts in parallel, and then quantifies relative homoeolog abundance based on matrix-assisted laser desorption/ionization time-of-flight mass-spectrometry. All Sequenom assays were conducted at the University of Minnesota Genotyping Facility.

RESULTS

Assessment of microarray quality

We analyzed the relative A- and D-genome contributions to the transcriptome of a synthetic F₁ hybrid and AD₁ allotetraploid cotton. This was done by comparing these mixed transcriptomes with the A₂ and D₅ model progenitors as well as with a 1:1 mix of A₂ and D₅ (Figure 1A). In total, 7,574 homoeolog-specific probe sets (around 33% of all possible) representing 1,383 unique EST contigs (hereafter referred to as genes) were identified as being reciprocally diagnostic with respect to identifying A- and D-genome specific expression in the F₁ hybrid and allotetraploid cotton. Thus, using conservative measures (false discovery rate (FDR) ≤ 0.05), we recovered 1,383 diagnostic genes, representing 2.6% (see (Rabinowicz et al., 2005)) to 4% (J Hawkins, personal communication) of the genic content of the cotton genome. As expected, a principal component analysis on the natural log differences of A- and D-genome expression distinguished among all accessions, placing the AD₁, F₁, and 1:1 mix values intermediate between A₂ and D₅ along the first axis (see Figure S1 in Additional file 1). This indicates
that the homoeolog-specific probes have performed as designed, and can be expected to yield useful estimates of A- and D-genome contributions to the transcriptome. Furthermore, quantitative mass-spectrometry validation of 12 homoeologous gene pairs from AD1 and 13 homoeologous gene pairs from F1 indicate that our findings regarding homoeolog-specific expression are reproducible (comparisons between platforms yielded $R^2$ values of 0.37 and 0.39 and $p$-values of 0.035 and 0.022, for AD1 and F1, respectively; see Figure S2 in Additional file 1).

Detection of genome expression biases in polyploid and F1 Gossypium

For each gene, a linear model was fit to the three replicate measures of relative A- and D-genome contributions. Using FDR corrected $p$-values (FDR ≤ 0.15) from this model, each gene from the AD1 and F1 samples was categorized as ‘A-biased’ (log ratio $((\ln(A_{probe}) - \ln(D_{probe}))$ statistically greater than 1:1 mix), ‘D-biased’ (log ratio statistically less than 1:1 mix) or ‘Equivalent’ (log ratio not statistically different from 1:1 mix); see Figure 2A. This categorization system is a rudimentary representation of the spectrum of homoeolog expression values, however, all categorizations presented here are based on known reference samples, which mitigates the effects of differential hybridization among homoeolog-specific probe pairs. In addition, this categorization is a statistical description of genome-specific transcript ratios and not a declaration of biological relevance (as pertaining to phenotype) of biases, which are unknown at present. Using this approach, many diagnostic gene pairs (29.9% (414 out of 1,383) of AD1 and 69.5% (961 out of 1,383) of F1) were inferred to be equivalently expressed in
petals. We infer that these gene pairs showed no statistically significant change in homoeologous (or allelic for the F1 hybrid) contribution to the transcriptome relative to the in vitro mid-parent value. Among those genes exhibiting biased expression, there was an approximately 1.3x and 2.5x overrepresentation of the D-genome biased genes in petal tissues of AD1 and F1, respectively (Figure 2A, B). In addition, we detected 46 AD1 and 6 F1 genes that appear to be A-genome silenced and 69 AD1 and 5 F1 genes that are D-genome silenced, indicating a significant increase in silencing in the AD1 allopolyploid in both the A- and D-genomes. For a limited sampling of genes, expression biases comparable to those above have been demonstrated previously in cotton (Adams et al., 2003; Adams et al., 2004; Adams & Wendel, 2005; Yang et al., 2006).

**Comparisons between hybridization (F1) and allopolyploidization (AD1)**

The comparison between the artificially synthesized F1 hybrid and the 1-2 MY old natural allopolyploid, G. hirsutum (AD1), allows us to assess the role genomic merger plays in the allopolyploidization process (Adams & Wendel, 2005; Hegarty et al., 2006). The inclusion of model A- and D-genome diploid progenitors facilitates inference of ancestral expression states and, hence, the directionality and pace of expression evolution (Figure 1A, B). An additional temporal dimension to the analysis concerns homoeolog-specific expression biases detected in the AD1 allopolyploid that were also detected in the F1 hybrid (Figure 2B, C). This is demonstrated by both the sizable set of shared genes found within all expression categories (Figure 2A) and the positive correlation (Pearson’s $r = 0.391; p$-value < $2.2 \times 10^{-16}$) between estimates of genomic contribution in the F1
hybrid compared with those from the allopolyploid (Figure 2C). Overall around 24% (235 out of 969) of the genes with an A- or D-genome expression bias in the polyploid are also found to be biased in the same direction in the F1 hybrid. This indicates that a significant portion of the expression evolution associated with allopolyploidization may have accompanied the initial genomic merger.

An additional directional trend in the data is a tendency for the allopolyploid genes to exhibit more extreme expression biases (Figure 2D). Both the A- and D-genome biased genes demonstrate a greater number of more extremely biased AD1 genes, when compared with the F1 (1 and 18 gene(s), respectively, for shared A- and D-biased sets). In addition, paired t-test for equality between AD1 and F1 values confirm that the differences in means between AD1 and F1 are significantly different for D-biased genes (AD1 mean = –0.45 and F1 mean = –0.36; p-value = 6.63 × 10⁻⁵), and marginally non-significant for A-biased genes (AD1 mean = 0.46 and F1 mean = 0.37; p-value = 0.07). Thus, for genes with immediate expression biases toward one parental *Gossypium* genome, stabilization and evolution of the allopolyploid genome preferentially continues to enhance this initial bias.

DISCUSSION

**Genomic merger and duplicate gene expression evolution**

It has long been thought that gene and genome duplication may serve as a key source of evolutionary innovation (Stephens, 1951; Ohno, 1970; Ohta, 1987; Walsh, 1995; Force *et al.*, 1999; Lynch & Conery, 2000). Recently, studies from a diverse array of organisms
have demonstrated that gene duplication stimulates a variety of evolutionary outcomes (Hughes & Hughes, 1993; Force et al., 1999; Lynch & Conery, 2000; Adams et al., 2003; Blanc & Wolfe, 2004a; Gu et al., 2004; Kellis et al., 2004; Gu et al., 2005; Casneuf et al., 2006; Duarte et al., 2006). These studies have demonstrated that following duplication, genes may evolve rapidly both at the sequence level and in their expression profile. It is thought that much of this change occurs as a result of the relaxation of purifying selection that occurs following duplication (Stephens, 1951; Ohno, 1970; Force et al., 1999; Lynch & Conery, 2000). During this period of relaxed selection, duplicate genes either find new roles (neofunctionalize), partition ancestral roles (subfunctionalize) or accumulate deleterious mutations and decay as pseudogenes. These processes are thought to occur on an evolutionary timescale measured in thousands to millions of years; for example, it has been estimated that the average half-life of duplicate gene pairs is of the order of 3 to 7 MY for mammals, invertebrates, and plants (Lynch & Conery, 2000). Here we have demonstrated that expression divergence among many genes duplicated by allopolyploidy (AD1) is already apparent at the stage of interspecific genomic merger between two genomes (F1). These genes, with conserved homoeologous biases between an ancient allotetraploid and modern F1 hybrid, represent the proportion of loci we might expect to have immediately experienced expression alteration at the time of allopolyploid origin 1 to 2 MYA. These data indicate that the critical parameter ‘time to subfunctionalization’ (Force et al., 1999; Lynch & Conery, 2000), may actually be zero for a significant fraction of the genome in allopolyploid plants. Thus, we conclude that during allopolyploidization, genomic merger per se plays a crucial and persistent role in determining subsequent evolutionary trajectories in homoeolog expression patterns.
In addition to the foregoing set of genes inferred to have experienced instantaneous expression alteration as a consequence of genomic merger, an even larger class of genes did not exhibit shared expression biases in the F1 hybrid and AD1 allopolyploid. Specifically, 76% of the genes that displayed biased expression in AD1 were not biased in the intergenomic F1. Reciprocally, about 44% (187 out of 422) of the genes with biased expression in the F1 were not biased in AD1. These differences of expression bias may reflect (1) additional expression evolution in allopolyploid cotton since the interspecific genomic merger via the mechanisms of neo-, sub-, and non-functionalization (Ohno, 1970; Force et al., 1999), (2) differences between the parents of the F1 hybrid and the actual diploid progenitors of AD1 (Figure 1A); that is, the extant diploids are good models but they are not the actual progenitors of allopolyploid Gossypium, or (3) elimination of initial genome specific biases during chromosome doubling or subsequent evolution of the natural AD1 allopolyploid.

It has been shown that genes belonging to some functional categories are retained, following duplication, at a higher than expected rate (Blanc & Wolfe, 2004a). As a corollary, it might be expected that gene function could also affect the likelihood of retention of expression bias. To explore this, we asked if genes from particular Gene Ontology (GO) (Ashburner et al., 2000) categories were over- or under-contributing to particular expression bias classes within the F1 hybrid and AD1 allopolyploid. Using the Blast2GO software (Conesa et al., 2005), only two GO categories were found to be significantly over-represented and none were under-represented (FDR ≤ 0.05; data not shown). Both significant GO categories were inclusive high-level biological processes (cofactor metabolic process (GO:0051186); and coenzyme metabolic process
(GO:0006732)), and were contained within the equivalently expressed genes from the F₁ hybrid. We had, however, only limited power (that is, small numbers of genes within GO categories) to detect distortions between the observed and expected frequencies of GO categories. Thus, within our subset of analyzed genes, gene classification does not appear to be a strong predictor of the direction or degree of genome-specific bias, although the strength of this conclusion may be limited by our current sample size.

Taken together, these data indicate that a significant proportion (around 24%) of duplicate gene expression evolution, ascribed to allopolyploid cotton, could have been generated immediately during allopolyploid formation by genetic and epigenetic factors associated with interspecific genomic merger (Birchler et al., 2005; Veitia, 2005; Chen & Ni, 2006). In addition, following allopolyploidy formation, subsequent duplicate gene evolution plays a large role in shaping homoeolog expression patterns. Thus, both immediate and long-term evolutionary processes contribute to homoeologous expression patterns. Based on this we speculate that expression-induced evolutionary novelty in allopolyploids occurs in two distinct modes: first, an immediate, massive, and saltational disruption of ancestral expression patterns accompanying the polyploidization process; and then a second, more gradual phase of expression evolution mediated by the mechanisms of duplicate gene evolution embodied in the traditional models (Ohno, 1970; Force et al., 1999) of the race between duplicate gene preservation and pseudogenization.

‘Instantaneous subfunctionalization’ could enhance duplicate gene retention
The signature of paleopolyploidy (ancient polyploidy) can be found in the genomes of all angiosperms (Vision et al., 2000; Wendel, 2000; Bowers et al., 2003; Blanc & Wolfe, 2004b; Lockton & Gaut, 2005; Cui et al., 2006). In addition, a high proportion (30% to 50%) of paleologs (duplicate gene pairs arising from a paleopolyploidy event) can be retained for millions of years (Force et al., 1999; Wendel, 2000; Blanc & Wolfe, 2004b).

Adams and Wendel (Adams & Wendel, 2005) have shown that A- and D-genome allelic pairs at the Adh locus display reciprocal silencing across multiple tissues in two Gossypium F1 hybrids. Thus, upon genomic merger ancestral gene expression domains are immediately partitioned and purifying selection is placed on both duplicate gene pairs, thereby increasing the probabilities of co-retention. To the extent that the results of Adams and Wendel (Adams & Wendel, 2005) are mirrored by the present analysis, we have demonstrated that, in petals, around 17% (235 out of 1,383; Figure 2A) of the homoeologous gene pairs studied could potentially fit this model, by having been found to be biased immediately in the F1 and by having that bias retained throughout allopolyploidy. If we extrapolate this finding to the entire Gossypium genome, it would indicate that, following polyploidization, a large number of homoeologs could be retained by ‘instantaneous subfunctionalization’, occurring solely from the initial effects of genomic merger. Furthermore, given that these biases appear to have been maintained for about 1 to 2 MY following polyploidization, this immediate form of expression bias may play an underappreciated role in the retention of duplicate genes following whole genome duplication (Adams et al., 2003).

**Tissue-specific expression dominance**
An intriguing aspect of the expression bias data is that for both natural AD1 allopolyplloid and the interspecific F1 hybrid, a greater number of genes exhibited a D-genome bias than the reverse (Figure 2A, B). This bias favors the paternal D-genome genome, and stands in contrast to the recently reported A-genome bias described for ovular tissue (Yang et al., 2006). To the best of the authors’ knowledge, this is the most extensive example of widespread paternal dominance. When considered in light of the results of Yang et al (Yang et al., 2006), our data suggest that neither Gossypium genome is globally dominant with respect to expression, but that instead, each genome may have local dominance in certain tissue types or developmental stages. This finding confirms previous results in Gossypium (Adams et al., 2003; Udall et al., 2006b) but differs from recent analysis of allotetraploid Arabidopsis (Wang et al., 2006). In the latter study, leaf and flower bud tissues from a synthetic Arabidopsis allotetraploid were shown to exhibit dominance favoring only its A. arenosa parent, with genome-wide suppression of the A. thaliana parental contribution. In the tissues that have been studied in Gossypium and Arabidopsis, it appears that both species demonstrate biased parental contributions to the transcriptome, however, in Gossypium these biases can favor either parental genome, whereas in Arabidopsis only the A. arenosa parent has demonstrated dominance. These findings reflect the importance and perhaps ad hoc nature of specific genomic combinations and their interactions during allopolyplloidization.

Among instantaneously subfunctionalized genes, genomic biases tend to become more extreme during subsequent allopolyplloid evolution
A notable observation in the present study is that genes showing biased expression patterns, tend to have more extreme biases in the AD₁ allopolyploid (Figure 2D), including a much larger number of silenced genes (115 total), when compared with the F₁ (11 total). One possible explanation for enhancement of genome-specific expression in allopolyploid cotton could be that immediately acting epigenetic effects become evolutionarily stabilized, either by natural selection or neutral processes. If this stabilization process is predisposed (through neutral or adaptive mechanisms) toward enhancing the initial expression bias, the result would be evolution toward a more extreme bias. This amplification of expression bias, which to our knowledge has not been described previously, may represent an additional factor underlying the genesis of phenotypic novelty in allopolyploid species.

CONCLUSIONS
These results extend previous findings of homoeolog expression biases in hybrid and allotetraploid cotton (Adams et al., 2003; Adams et al., 2004; Adams & Wendel, 2005; Udall et al., 2006b; Yang et al., 2006). By employing microarray technology to analyze a large number of genes, we describe the general phenomenon of genomic expression bias in both a modern synthetic F₁ hybrid and an ancient allotetraploid. Furthermore, for petal tissues, these biases favor the parental D-genome and have become more extreme in the allotetraploid when compared with the F₁ hybrid. By comparing homoeolog contributions to the transcriptome from the F₁ hybrid and AD₁ allotetraploid, it was possible determine the role of genomic merger in producing homoeolog expression biases. Given this comparison, we have shown that a significant fraction of the expression biases found in
the allotetraploid is likely initiated immediately by genomic merger. A still larger fraction of the expression biases is inferred to have arisen from long-term evolutionary processes, thus implicating two temporally distinct phases of expression evolution following allopolyplodization.

Authors' contributions
LF, JU, and JW designed the research. JU and JW designed the microarray platform. LF and JU performed the research. LF and DN analyzed the data. LF and JW drafted the manuscript. All authors participated in editing the manuscript and approved the final version.

Acknowledgements
We thank J Stewart and D Stelly for generation and contribution of the F<sub>1</sub> hybrid used in this study. B Stupar and N Springer offered invaluable advice and technical support regarding the Sequenom microarray platform. This research was supported by a grant from the National Science Foundation (to JW) and a grant from the US Department of Agriculture (to JW and JU), and by an Iowa State University Plant Sciences Institute Fellowship (to LF).
REFERENCES


Table 1 - Details of plant materials used in this study. Natural allotetraploid *Gossypium* evolved 1 to 2 MYA from diploid A- and D-genome progenitors, most similar to the modern species *G. arboreum* and *G. raimondii* (Senchina et al., 2003; Wendel & Cronn, 2003). The A-genome parent is the inferred cytoplasmic donor to *G. hirsutum* (Wendel, 1989; Wendel & Albert, 1992), and thus the F$_1$ cross was created in the same manner, with $A_2$ as the maternal parent.

<table>
<thead>
<tr>
<th>Species name</th>
<th>Genome designation</th>
<th>Accession</th>
<th>Ploidy level</th>
<th>Location of origin</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>G. arboreum</em></td>
<td>A$_2$</td>
<td>cv. AKA-8401</td>
<td>Diploid</td>
<td>Africa</td>
</tr>
<tr>
<td><em>G. raimondii</em></td>
<td>D$_5$</td>
<td>Accession</td>
<td>Diploid</td>
<td>Peru</td>
</tr>
<tr>
<td></td>
<td></td>
<td>unnamed</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>G. arboreum X</em></td>
<td>A$_2$ ♀ X D$_5$ ♂</td>
<td>Accession</td>
<td>Diploid</td>
<td>Synthetic</td>
</tr>
<tr>
<td><em>G. raimondii F$_1$</em></td>
<td></td>
<td>unnamed</td>
<td></td>
<td>hybrid</td>
</tr>
<tr>
<td><em>G. hirsutum</em></td>
<td>AD$_1$</td>
<td>cv. Maxxa</td>
<td>Allotetraploid</td>
<td>Mexico/Central America</td>
</tr>
</tbody>
</table>
Figure 1 - Phylogenetic context and inference of homoeologous expression evolution in *Gossypium*

(A) Phylogenetic relationships among the cotton accessions used in this study. An allopolyploid event between A- and D-genome diploid species (red star) created modern allopolyploid *Gossypium hirsutum* (AD$_1$). Using models of the ancestral genome donors (A$_2$ and D$_5$), an interspecific diploid hybrid (F$_1$) was created (blue star). Although not a perfect match, the model A- and D-genome donors are the best modern representatives of the diploids that underwent allopolyploidization to form AD$_1$ and, as such, provide the best available reconstruction this ancient event. (B) Possible expression phenotypes and associated evolutionary inference. The far left pie represents equal expression among model diploid progenitor species (denoted by A$_2$ and D$_5$). Given this starting condition, several expression states are possible following allopolyploidy or hybridization. Some potential outcomes are indicated by the five pies on the right (A$_i$ and D$_i$ denote co-resident genomes, either in the hybrid or allopolyploid). (C) Detection of conserved homoeolog-specific single nucleotide polymorphism (SNPs). Given an alignment of expressed sequence tag (EST) sequences from orthologous genes from both diploid and allopolyploid genomes, species- and genome-specific SNPs (all SNPs highlighted in gray) can be detected. The middle SNP is an example of a genome-specific SNP. With this conserved SNP, homoeolog- and allele-specific microarray probes can be generated (potential microarray probe region highlighted in blue), and used to assay expression in allopolyploid and hybrid species.
A. modern synthetic hybridization

B. possible gene expression states following genomic merger

C. equal expression between diploid species prior to genomic merger

- Complete A-genome bias (D-genome silencing)
- Partial A-genome bias
- No change
- Partial D-genome bias
- Complete D-genome bias (A-genome silencing)

Genetic sequences:

\( \alpha \)
\( \alpha_D \)
\( \alpha_t \)
\( \alpha_D \)

ATGCCGATTTAATCTGTCCATTTAGCCAATGATTCTCTAGTAATTAG
Figure 2 - Inferred contributions to the transcriptome by A- and D-genomes in a natural *Gossypium* allotetraploid and a synthetic diploid hybrid

(A) A- and D-genome contribution to the transcriptome for 1,383 homoeologous/allelic gene pairs. Each gene pair categorized based on a linear model analysis of three replicate measures of genomic contribution. ‘Shared genes’ are those with expression patterns that are conserved between allotetraploid *G. hirsutum* (AD₁) and the diploid F₁ hybrid (F₁).

(B) Diagrammatic representation of the conservation of genes with biased expression.

(C) Scatterplot comparing the homoeolog expression log ratios found in the natural allotetraploid AD₁ to the synthetic F₁ hybrid. Each point represents a single gene. The correlation ($r$) and best-fit line are indicated in red. This correlation has a $p$-value less than $2.2 \times 10^{-16}$, indicating that it is significantly different from zero.

(D) Kernel density estimates of the homoeolog expression log ratios for all 1,383 genes from the 1:1 mix (green line), F₁ hybrid (black line), and AD₁ allotetraploid (red line) cotton. This plot demonstrates an erosion of equal contribution from both genomes and a shift toward more extreme values in the allopolyploid when compared with the F₁ or 1:1 mix.
A. | Species | A-biased | D-biased | Equiv. | Total |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>F₁</td>
<td>121</td>
<td>301</td>
<td>961</td>
<td>1383</td>
</tr>
<tr>
<td>AD₁</td>
<td>423</td>
<td>546</td>
<td>414</td>
<td>1383</td>
</tr>
<tr>
<td>Shared genes</td>
<td>51</td>
<td>184</td>
<td>324</td>
<td>559</td>
</tr>
</tbody>
</table>

B. | AD₁ | F₁ |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>372</td>
<td>51</td>
</tr>
<tr>
<td>362</td>
<td>184</td>
</tr>
</tbody>
</table>

C. $r = 0.391$

D. | 1:1 mix | F₁ | ADr |

expression log ratio (ln(A/ren) - ln(D/ren))
Supplemental Material

SI Methods:

Microarray Design:

Homoeolog-specific probes were created by first assembling EST contigs from A2, D5, and AD1 (Table 1) libraries and then identifying homoeolog-specific SNPs within these contigs. These SNPs represent nucleotide differences between the A and D genome orthologs, and offer the possibility of diagnosing the genomic origin of transcripts in the allopolyploid nucleus. Additionally, when available, AD1 EST sequences confirmed the conservation of A- and D-genome-specific SNPs in the allopolyploid species. Using this strategy, 11,399 high-quality SNPs were identified, encompassing 2029 contigs. For each of these 11,399 SNPs, complimentary plus and minus strand A and D homoeolog-specific probes sets were designed, generating in total 22,798 probes sets, and 45,596 unique probes.

Mass-Spectrometry Validation Experimental Design and Methodology:

Cotton petal RNA samples were converted to cDNA and PCR amplified with multiplex primer sets, which targeted 13 genes from the homoeolog-specific microarray results. Each biological replicate was split into three technical replicates resulting in 9 total replicate measures for each species (3 bio. reps. X 3 tech. reps.). Amplified multiplex products were sent to the University of Minnesota for homoeolog-specific MALDI-TOF mass-spectrometry quantification using a Sequenom (San Diego, CA)
MassARRAY device. The mean value for each of the nine replicates was determined and compared to the estimates derived from the homoeolog-specific microarray (Supp. Fig. 3)
Supp. Table 1. A- and D-genome contribution to the transcriptome at FDR thresholds of 0.05 (\(A\)) and 0.1 (\(B\)). Each gene pair categorized based on a linear model analysis of three replicate measures of genomic contribution. “Shared genes” are those with expression patterns that are conserved between \(G. hirsutum\) and the diploid hybrid.

Supp. Table 1A: FDR threshold = 0.05

<table>
<thead>
<tr>
<th></th>
<th>A-bias</th>
<th>D-bias</th>
<th>Equiv.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(F_1)</td>
<td>37</td>
<td>76</td>
<td>1270</td>
<td>1383</td>
</tr>
<tr>
<td>(AD_1)</td>
<td>283</td>
<td>380</td>
<td>720</td>
<td>1383</td>
</tr>
<tr>
<td>shared</td>
<td>13</td>
<td>47</td>
<td>683</td>
<td>743</td>
</tr>
</tbody>
</table>

Supp. Table 1B: FDR threshold = 0.1

<table>
<thead>
<tr>
<th></th>
<th>A-bias</th>
<th>D-bias</th>
<th>Equiv.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(F_1)</td>
<td>76</td>
<td>186</td>
<td>1121</td>
<td>1383</td>
</tr>
<tr>
<td>(AD_1)</td>
<td>358</td>
<td>472</td>
<td>553</td>
<td>1383</td>
</tr>
<tr>
<td>shared</td>
<td>33</td>
<td>112</td>
<td>483</td>
<td>628</td>
</tr>
</tbody>
</table>
Supp. Figure 1. Principle Component Analysis of natural log differences between A- and D-genome specific probe expression levels. All three replicate samples of each genotype are represented. Character symbols for the five genomic samples are as follows: “A” = A₂, “D” = D₅, “F” = F₁ hybrid, “M” = 1:1 A₂:D₅ RNA mix, and “P” = AD₁ allotetraploid. The proportion of the total variance explained by each principle component is listed on the corresponding axis.
Supp. Figure 2. Validation of homoeolog expression results for AD₁ and F₁ accessions. (A) A comparison of results for 13 randomly chosen genes. All NimbleGen (microarray) values are expressed as the log ratio \( \ln(A_{\text{probe}}) - \ln(D_{\text{probe}}) \), whereas the Sequenom (mass-spectrometry) values are expressed as the proportion of the transcriptome contributed by the A-genome. Thus both metrics result in analogous interpretations of the different data types (i.e. for both technologies, larger values reflect greater A-genome contribution to the transcriptome, and smaller values reflect greater D-genome contribution). Scatter plots of validation results for AD₁ and F₁ (B) with their associated best-fit line, \( R^2 \) value, and \( p \)-value.

Supp. Figure 2A:

<table>
<thead>
<tr>
<th>contig</th>
<th>SNP position</th>
<th>AD₁ Sequenom % A</th>
<th>AD₁ NimbleGen ln(a) - ln(d)</th>
<th>A</th>
<th>F1 NimbleGen ln(a) - ln(d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>COTTON16_00001_062</td>
<td>1928</td>
<td>0.543</td>
<td>0.728</td>
<td>0.549</td>
<td>0.551</td>
</tr>
<tr>
<td>COTTON16_00024_03</td>
<td>2070</td>
<td>0.293</td>
<td>-0.867</td>
<td>0.097</td>
<td>-0.766</td>
</tr>
<tr>
<td>COTTON16_00076_06</td>
<td>860</td>
<td>NA</td>
<td>NA</td>
<td>0.449</td>
<td>-0.448</td>
</tr>
<tr>
<td>COTTON16_00174_02</td>
<td>802</td>
<td>0.734</td>
<td>0.562</td>
<td>0.408</td>
<td>-0.251</td>
</tr>
<tr>
<td>Code</td>
<td>Count</td>
<td>1st Value</td>
<td>2nd Value</td>
<td>3rd Value</td>
<td>4th Value</td>
</tr>
<tr>
<td>--------------------</td>
<td>-------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>COTTON16_00285_02</td>
<td>685</td>
<td>0.425</td>
<td>0.09</td>
<td>0.395</td>
<td>0.343</td>
</tr>
<tr>
<td>COTTON16_00690_02</td>
<td>916</td>
<td>0.469</td>
<td>0.725</td>
<td>0.384</td>
<td>-0.946</td>
</tr>
<tr>
<td>COTTON16_01391_01</td>
<td>705</td>
<td>0.607</td>
<td>0.01</td>
<td>0.554</td>
<td>0.059</td>
</tr>
<tr>
<td>COTTON16_07872_01</td>
<td>1017</td>
<td>0.311</td>
<td>-0.632</td>
<td>0.255</td>
<td>-0.405</td>
</tr>
<tr>
<td>COTTON16_07872_01</td>
<td>1185</td>
<td>0.313</td>
<td>-0.66</td>
<td>0.21</td>
<td>-0.433</td>
</tr>
<tr>
<td>COTTON16_09095_01</td>
<td>1544</td>
<td>0.504</td>
<td>0.59</td>
<td>0.586</td>
<td>0.691</td>
</tr>
<tr>
<td>COTTON16_21601_01</td>
<td>747</td>
<td>0.557</td>
<td>0.026</td>
<td>0.515</td>
<td>0.04</td>
</tr>
<tr>
<td>COTTON16_25466_01</td>
<td>1125</td>
<td>0.482</td>
<td>0.425</td>
<td>0.531</td>
<td>-0.412</td>
</tr>
<tr>
<td>COTTON16_32946_01</td>
<td>1145</td>
<td>0.702</td>
<td>-0.067</td>
<td>0.564</td>
<td>-0.107</td>
</tr>
</tbody>
</table>
Supp. Figure 2B:
CHAPTER 3.
EVOLUTIONARY RATE AND SCOPE OF DUPLICATE GENE EXPRESSION
EVOLUTION IN FIVE ALLOTETRAPLOID COTTON SPECIES

A paper in preparation for New Phytologist
Lex E. Flagel and Jonathan F. Wendel

Abstract

Here we describe the evolution of gene expression among a diversified cohort of five ancient allopolyploid species in the cotton genus (Gossypium) and compare this to a synthetic F₁ hybrid. Using this framework we are able to systematically analyze gene expression following a shared genome duplication within these allopolyploids giving us insights into the evolutionary importance of polyploidization. To perform this work we hybridized petal RNAs to a custom designed Gossypium microarray. This platform can measure total expression for ~42000 genes, and genome specific expression for ~1400 genes. Overall we find evidence for homoeolog (genes duplicated by polyploidy) expression bias favoring the allopolyploid D-genome over the A-genome in all species. Furthermore we find surprising levels of transgressive up- and down-regulation in the allopolyploids, but not in the F₁ hybrid. We detect significant levels of expression evolution among the five natural allopolyploid species. This expression evolution follows a shared pattern among the Gossypium allotetraploids, but shows sharp contrasts to the patterns found in the synthetic F₁ hybrid.
Introduction

Polyploidy, or whole genome duplication, is a prevalent feature among angiosperm species (Wendel, 2000; Comai, 2005; Leitch & Leitch, 2008). Beyond contemporary species, emerging genomic data has shed light on the ancient and recurrent history of polyploidy among the angiosperms (Barker et al., 2008; Tang et al., 2008; Soltis et al., 2009). Because polyploidy involves duplication of the entire genome, its effect on genomic organization can be extensive (Comai, 2005), including well documented cases of structural and epigenetic modifications (Shaked et al., 2001; Gaeta et al., 2007; Buggs et al., 2009; Ni et al., 2009; Tate et al., 2009), as well as changes in gene expression patterns (Bottley et al., 2006; Hegarty et al., 2006; Wang et al., 2006; Flagel et al., 2008; Hovav et al., 2008; Rapp et al., 2009). Furthermore, some of these genome-wide changes have been linked to phenotypic variation (Pires et al., 2004; Gaeta et al., 2007; Ni et al., 2009), indicating that polyploidy can be an important source of phenotypic evolution.

The establishment of a new allopolyploid species is not a trivial feat. First all viable allopolyploids must survive several immediate genomic challenges, including the merger of diversified genomes, the resolution of potentially conflicting developmental signals, and new or possibly accidental interactions with organellar genomes, in addition to overcoming the reproductive barriers associated with polyploidy (Wendel, 2000; Comai, 2005). Following this, and owing to their redundant genomic architecture, allopolyploid genomes then face several interesting and potentially dramatic evolutionary resolutions. These include the genomic decay of duplicate genes either in the form of genomic fragments loss (Shaked et al., 2001; Tate et al., 2009) or mutational obliteration (pseudogenization), the genomic partitioning of ancestral
functions (subfunctionalization; (Force et al., 1999)), or the possibility of a chance beneficial mutation conferring new functionality (neofunctionalization; (Ohno, 1970)). These outcomes most likely require evolutionary time-scales, are not mutually exclusive (Conant & Wolfe, 2008), and can be distorted by additional genomic disruptions, such as further hybridization and/or polyploidization leading to the accumulation of additional genomic content, yielding higher ploidies and additional genomic complexity (e.g. Spartina anglica, sugarcane (Saccharum officinarum), or wheat (Triticum aestivum)). In the absence of hybridization or additional rounds of polyploidization new polyploidy species can undergo divergence and speciation (e.g. cotton). As this special edition of New Phytologist demonstrates, the polyploid research community has made major inroads in studying the genomic consequences of polyploidy. Despite this progress many important questions remain unanswered. The study we present here addresses one of these questions, namely, how does the diversification of allopolyploid species during evolutionary time-scales impact gene expression among co-resident genomes in cotton?

One to two million years ago allopolyploidization within the genus Gossypium (cotton) resulted in novel allotetraploid species containing a full complement of the Old World A- and New World D-diploid cotton genomes (Senchina et al., 2003; Wendel & Cronn, 2003). Since that time, species containing this favorable genomic combination have spread throughout the tropical portions of the New World and have diversified into five extant allotetraploid species (Wendel & Cronn, 2003), though a sixth species, G. ekmanianum, was recently proposed (Krapovickas & Seijo, 2008). The presence of shared allopolyploid-specific nucleotide polymorphisms within these species indicates they likely evolved from a common polyploidy event and as a consequence have left a traceable phylogenetic history which was been revealed in previous studies (Wendel et al., 1994; Small et al., 1998) (Fig. 1a).
The evolutionary framework provided by the five natural *Gossypium* allotetraploids offers an excellent opportunity to study replicated evolutionary trajectories following the combination of diversified genomes. In addition to their compelling natural history, two allotetraploid cottons, *G. hirsutum* and *G. barbadense*, are the primary contributors of natural fiber for use in the apparel industry, making it agriculturally and economically important to understand their evolutionary history. The study of these allopolyploids has benefited from considerable genomic resources, including a sizable EST collection (Udall *et al.*, 2006a), with ESTs from both model diploid parents (A-genome: *G. arboreum*; D-genome: *G. raimondii*; Fig. 1a) as well as the allotetraploid *G. hirsutum*. This genomic resource has been used to create a novel microarray platform, which can be used to explore global gene expression levels among ~42000 genes using probes targeted at conserved genic regions of the A and D cotton genomes, and homoeologous (genes duplicated by polyploidy) expression levels for ~1400 genes using pairs of probes differentiated by a genome specific SNP (Udall *et al.*, 2006b; Flagel *et al.*, 2008). Using this microarray platform several key findings have been made regarding polyploidy in *Gossypium*. With most relevance to the present study, we have previously shown that both genomic merger and allopolyploid evolution play an important role in homoeolog expression evolution (Flagel *et al.*, 2008), and that homoeologs expression is biased in favor of the D-genome in *G. hirsutum* in both petal and fiber tissues (Flagel *et al.*, 2008; Hovav *et al.*, 2008). Following these initial findings regarding homoeologous expression, continued work with this microarray platform has highlighted a form of genomic expression dominance, whereby the allotetraploid assumes an expression state of the D-genome parent significantly more often than it does the A-genome parent, regardless of whether that state is up- or down-regulation (Rapp *et al.*, 2009). Beyond these studies in *Gossypium*, work in allopolyploid wheat (Bottley *et al.*, 2006; Bottley &
Koebner, 2008; Pumphrey *et al.*, 2009) and *Tragopogon* (Tate *et al.*, 2006) has further demonstrated a considerable frequency of biases in the genomic contribution among homoeologs, and work in synthetic *Arabidopsis* allotetraploids has shown global down regulation of the *A. thaliana* genome in favor of the *A. arenosa* genome (Wang *et al.*, 2006), which could be considered another form of genomic dominance. Together these observations are beginning to confirm the notion that the genomic disruptions associated with allopolyploidy (i.e. genomic hybridization and duplication) may contribute considerably to gene expression evolution within established and nascent polyploids (Osborn *et al.*, 2003; Chen, 2007; Paun *et al.*, 2007; Doyle *et al.*, 2008). Here, we extend the scope of earlier findings by demonstrating significant levels of expression evolution among a diversified collection of natural allopolyploid species, and by showing that portions of this evolutionary pattern appear to have advanced shared across the five natural alltetraploid cotton species.

**Materials and Methods**

*Plant materials, RNA extraction and microarray preparation*

Replicates of four *Gossypium* allotetraploid species (*G. barbadense*, *G. darwinii*, *G. mustelinum*, and *G. tomentosum*; Table 1) were grown in the Pohl Conservatory at Iowa State University. Petal tissues were harvested from these accessions on the day of anthesis at full petal expansion. All petal tissues were snap frozen in liquid nitrogen and stored at -80°C. Prior to RNA extraction, petal tissue from multiple flowers (>3) of a single plant were pooled to form three replicates, and subjected to RNA extraction following a modified hot borate procedure (Wan & Wilkins, 1994). Following extraction, RNA samples were run on a Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) to assess degradation. Finally total RNA extracts
were sent to Roche NimbleGen (Madison, WI, USA) for labeling and hybridization to a custom *Gossypium* microarray platform (microarray design details found in Flagel *et al.* (2008)). Briefly, this microarray features two classes of probes, including 7574 ~35-mer pairs of A- and D-genome specific probes (each containing a genome specific SNP at their central base; targeting 1383 contigs), which have previously demonstrated diagnosticity in assessing levels of A- and D-genome expression within an A- by D-genome F1 hybrid (*G. arboreum* X *G. raimondii*) and allopolyploid *G. hirsutum* (Flagel *et al.*, 2008), as well as 297206 ~60-mer generic probes (conserved between the A- and D-genome; targeting 42459 contigs), which have been utilized to detect global expression, without homoeolog specificity (Chaudhary *et al.*, 2008; Rapp *et al.*, 2009). Thus, this microarray platform makes it possible to measure total expression for about 80% of the estimated genic content of the cotton genome (Rabinowicz *et al.*, 2005), and for a smaller subset of genes, the platform can also detect the proportions of A- and D-genome contribution.

**Statistical analysis**

All raw microarray data from the four natural allotetraploid species were extracted into two working files, one for the ~35-mer genome specific probes and one for the ~60-mer generic probes. These data sets were then combined with expression data from an earlier study (Flagel *et al.*, 2008), which included *G. hirsutum* (the fifth natural allotetraploid cotton species), as well as the F1 hybrid mentioned above and an equimolar mix of RNA pools from the model diploid progenitors (*G. arboreum* (A-genome) and *G. raimondii* (D-genome)). Following this merger, both the genome specific and generic data sets were normalized and subjected to statistical
analysis separately, as they represent dissimilar probe types each addressing different aspects of
gene expression.

For the 7574 diagnostic genome specific probe pairs (see Flagel et al. (2008) for detail
regarding diagnostic probe selection) all raw values were natural log transformed and quantile
normalized. Following this the expression values of each pair were converted to the difference
between the A- and D-genome probe natural logs ($\ln(A_{\text{probe}}) - \ln(D_{\text{probe}})$; hereafter referred to as
log ratio). Next these log ratio values were reduced to the 1383 contigs they represent by
calculating a robust average of all probe pairs for each contig using Tukey's Biweight method.
Finally contig-level expression differences were determined using a linear model which included
genotype and replication effects. This model was used to contrast the five natural allotetraploid
species and the F1 hybrid to the parental mix. $P$-values derived from this contrast were corrected
for multiple testing using the method of Storey and Tibshirani (2003). Significance was assessed
from the resultant $q$-values using a false discovery rate (FDR) threshold of $q \leq 0.15$. This
threshold was arrived at by first estimating the number of true nulls using the method described
by Nettleton et al. (2006) (Table 2), then selecting the $q$-value of 0.15 as a good compromise
between the expected number of false positives and false negatives for all contrasts. The results
from the $q$-value thresholds $q \leq 0.05$ and $q \leq 0.1$ can also be found in Supporting Information
Table S1.

The analysis of expression from the 297206 ~60-mer generic probes has been previously
described by Rapp et al. (2009) and follows a general outline similar to that above. The
expression values natural log transformed, quantile normalized and reduced to 42459 contigs
using Tukey's Biweight method. Following this expression differences were detected after
fitting a linear model which included genotype and replication effects. \(P\)-values from these contrasts were converted to \(q\)-values using the method of Storey and Tibshirani (2003), and a threshold of \(q \leq 0.05\) was used to assess significance to allow direct comparison to the results of Rapp et al. (2009).

**Validation**

We validated our microarray estimates of homeolog expression ratio expression for 14 genes using a sensitive SNP specific Sequenom (San Diego, CA, USA) mass-spectrometry platform that was initially described for use in maize by Stupar & Springer (2006), and has a proven utility for estimating homoeologous expression ratios (Chaudhary et al., 2008) and for validation of our custom *Gossypium* microarray (Flagel et al., 2008; Hovav et al., 2008). Using this platform we compared homoeolog expression ratios between the microarray and mass-spectrometry platforms for *G. barbadense*, *G. darwinii*, *G. mustelinum*, and *G. tomentosum* (Supporting Information Fig. S1); the *G. hirsutum* and F\(_1\) hybrid microarray expression estimates have been previously validated (Flagel et al., 2008). The validations show significant correlations between the microarray and mass-spectrometry estimates for *G. darwinii*, *G. mustelinum*, and *G. tomentosum* (Pearson’s \(r = 0.525, 0.535,\) and 0.54; \(p\)-value = 0.053, 0.048, and 0.046, respectively), and a non-significant, though moderate correlation for *G. barbadense* (Pearson’s \(r = 0.366, p\)-value = 0.19). Despite the non-significant correlation for *G. barbadense*, these results confirm the quality of our microarray data when we take into account the major technological differences between microarray and mass-spectrometry platforms and a considerable history of validated results for this platforms when applied

**Microarray data deposition**

All original microarray data files can be found on the NCBI GEO website under the accessions XXXXX-XXXXX in compliance with MIAME standards.

**Results**

*Comparision of homoeolog expression biases between allotetraploid cottons*

Using *Gossypium* petals as a model tissue type, we have assessed the ratio of homoeolgous contribution to the transcriptome among 1383 duplicate gene pairs. After applying a FDR threshold of 0.15 for significance testing, we have tabulated the A-biased (significantly more A-genome expression than the 1:1 parental mix), D-biased (significantly more D-genome expression than the 1:1 parental mix), and equivalently expressed genes for each of the five allotetraploid *Gossypium* species and a synthetic F1 hybrid (Table 2). The 1:1 parental RNA mix represents a best approximation of the anticipated expression state within the allotetraploids and F1 in the absence of gene expression evolution. From our results it is clear that all species show considerable deviations from this parental mix, with each species showing a substantial number of genes with both A- and D-genome biases. As was the case in our previous study (Flagel *et al.*, 2008), the F1 hybrid shows less biases overall than do any of the five allotetraploids. Also, among the allotetraploids there is considerable variation, with *G. mustelinum*, a wild species and the most basal of the *Gossypium* allotetraploids (Fig. 1a), showing the least divergence and *G. tomentosum* (a wild Hawaiian Island endemic) and *G. barbadense* (a domesticated South
American cultivar) showing the greatest levels of homoeolog expression biases (Table 2). Also consistent with our previous studies of petal and fiber tissues (Flagel et al., 2008; Hovav et al., 2008), all allotetraploids and the F₁ hybrid show a greater number of paternal D-biased genes than maternal A-biased genes.

Because the *Gossypium* allotetraploids have a known phylogenetic history (Fig. 1a), it is possible to visualize these homoeologous expression changes on the phylogeny. To do this we treated the expression log ratio values as quantitative characters and used them to estimate the species level phylogeny of the *Gossypium* allotetraploids using the contml program from the PHYLIP package (Felsenstein, 2005). The resulting “homoeolog expression” phylogeny (Fig. 1b) has a similar topology to the known phylogeny (note the polytomy at the base of Fig. 1b compared to Fig. 1a). The branch lengths found on this “expression tree” are proportional to the levels of expression deviation from a common ancestor. From this representation it is clear that *G. tomentosum* has experienced the greatest amount of total expression evolution. This is because *G. tomentosum* has a large number of A- and D-biases (Table 2), and in addition many of these biases are quite extreme, as indicated by the total branch length in Fig. 1b, which is a function of the total deviation from an equivalent expression ratio. Furthermore, *G. barbadense*, which has similar numbers of biased genes when compared to *G. tomentosum* (Table 2), has less overall deviation from its common ancestor with *G. darwinii* (a wild Galapagos Islands endemic) than we might expect. This effect is likely because many *G. barbadense* homoeologs are expressed in a biased manner that is statistically significant, though they do not deviate from equivalence to the degree found in *G. tomentosum*. For all species the distribution of homoeolog expression levels can be found in Fig. 2. These histograms represent the expression log ratios for all 1383 genes. From the profiles of these histograms we can see that there are significant
differences in the level of deviation from equivalent expression in each of the species, with *G. tomentosum* having a very broad profile relative to all other species, consistent with its high level of homoeolog expression divergence, and the F₁ hybrid having the most narrow profile, consistent with the least deviation from equivalent genomic expression. Also from these histograms we can visualize a trend toward D-genome bias, as all species have a greater density of values below zero than above. However, it is also clear that this D-genome bias is largely a quantitative phenomenon, and is not caused, for example, by a large number of genes with an extreme D-bias, but rather by an overall accumulation of many small D-biases.

*Global categorization of expression profiles and genomic dominance among allotetraploid cottons*

Beyond examining homoeologous expression for 1383 genes we also compared generic expression states for 42459 genes between each of the allopolyploids and their A- and D-genome parents. The probes used to measure expression among these genes are generic to either the A- and D-genome, meaning they can only measure the cumulative output of both homoeologs, and cannot detect homoeolog specific expression as in the previous section. Within an allopolyploid these generic probes can, however, be used to detect expression evolution in the form of non-additive expression states (allotetraploid expression not equivalent to the average expression of the parental species), such as parental dominance and transgressive up- or down-regulation (Wang *et al.*, 2006; Rapp *et al.*, 2009). Using this method of comparison, Rapp *et al.* (2009) showed that this type of expression data can be parsed into twelve informative categories of expression evolution, to which they gave the Roman numeral designations seen across the top of Fig. 3. These include two forms of additive expression (I and XII; Fig. 3), which represent the
null hypothesis, as well as genomic dominance (II, IV, IX and XI), and transgressive up- (V, VI, and VIII) and down-regulation (III, VII, and X). For each of these twelve evolutionarily informative states we have tallied the gene counts from among the 42459 genes assess by our microarray along with a tally of genes that showed statistically equivalent expression among the A- and D-genome parents and the allotetraploid or F₁ hybrid (“No Change”; Fig. 3).

From the data collected in Fig. 3 we can see that the levels of additive expression (I and XII) are approximately equal among all species and that A- and D-genomic dominance (IV and IX vs. II and XI) is approximately equal among the allopolyploids. On the other hand, the F₁ hybrid shows about double the level of D-genome up- and down-expression dominance (II and XI) when compared to the reciprocal forms of A-genome dominance (II and XI), while all of the natural allotetraploids show far more transgressive up- (V, VI, and VIII) and down-regulation (III, VII, and X), often greater than the F₁ hybrid by approximately a factor of ten. These intriguing patterns would appear to indicate that genomic merger (F₁ hybrid) causes an immediate D-genome bias. The natural allotetraploids may have also experienced this immediate bias upon their formation, but as they matured the level overrepresentation of D-genome bias diminished, and a massive amount of transgressive expression evolution occurred. Finally within each of these categories there is some variation between the allopolyploids, though this variation is smaller than that between any of the allopolyploids and the F₁ hybrid, and is likely constrained to an extent by a shared evolutionary history (Fig. 1a).
Discussion

The role of genome merger and duplication in bring about gene expression evolution

Previous analyses in *Gossypium* have shown that genome merger, genome duplication, and subsequent duplicate gene evolution can play a role in altering homoeologous expression profiles (Flagel *et al.*, 2008; Chaudhary *et al.*, 2009). These studies used *G. hirsutum* as the only allotetraploid representative, and our present study we provide additional support for these findings by showing that all *Gossypium* allotetraploids have significant levels of homoeologous expression bias (much more so than the F1 hybrid; Table 2 and Fig. 2) and that these biases favor the D-genome. Because these characteristics are found throughout the allotetraploid phylogeny we can speculate that they either; 1) occurred after allopolyploid formation but prior to speciation, or 2) evolved recurrently after speciation in each allotetraploid lineage. Either alternative is interesting; the first would indicate that these patterns of gene expression evolved relatively quickly in the original ancestral allotetraploid, while the second would indicate a repeating evolution of a this particular expression patterns, possibly due to some intrinsic characteristics of the A- by D-genome combination. Though intriguing, we must point out that this is merely speculation, as we have at the present no data with which to substantiate either alternative, and furthermore the alternatives are not mutually exclusive, thus the observed expression patterns may be the product of evolution occurring both before and after allotetraploid speciation. In any case, our studies on the four remaining allotetraploid species, and there confirmation of our initial findings in *G. hirsutum*, provide strong support for our earlier hypotheses of the enhancement of homoeolog biases and increased D-genome bias following allopolyplody in *Gossypium* (Flagel *et al.*, 2008).
Among the five natural allotetraploids used in this study, two species were represented by elite cultivars from a domesticated background (G. barbadense cv. Pima S7 and G. hirsutum cv. Maxxa), while the other three species, G. mustelinum, G. darwinii, and G. tomentosum, are wild, the latter two being island endemics, and G. mustelinum restricted to a small native range in northeastern Brazil (Wendelet al., 1994; Wendel & Cronn, 2003). Interestingly, though both domesticates show significant levels of homoeologous expression bias (Table 2 and Fig. 1b), neither are as strongly biased as the wild species G. tomentosum. Among the domesticates, it is possible that some alteration in expression is the byproduct artificial selection during domestication. However, it is important to note that our study focuses on petal tissues, the phenotypes of which are unlikely to have been under selection during domestication and subsequent crop improvement. Also, both domestication and island colonization are likely to cause a genetic bottleneck, an event that may trigger the release of epigenetic variation (Rapp & Wendel, 2005), potentially contributing to the varied expression patterns and phenotypes found among the G. barbadense, G. hirsutum, G. tomentosum, and G. darwinii. Interestingly, however, G. darwinii, the other island endemic, has less biased expression patterns than does G. tomentosum, which may indicate that there are additional contributing factors and that there is a idiosyncratic nature to homoeologous expression evolution. Despite these variations between species, we must also highlight the fact that there is also substantial conservation among biased genes. Overall the number of shared A- and D-biased genes from among the five allotetraploids ranges from between 29-60% in each species (Table 2), which is a considerable and may be a reasonable representation of the proportion of ancestral biases inherited and maintained by all species.
Temporal changes in expression evolution in cotton

Our findings highlight a key temporal component of expression evolution among natural Gossypium allotetraploids. The synthetic F₁ hybrid used in this study and the synthetic A/D allopolyploid used by Rapp *et al.* (2009) both show strong evidence for genomic dominance, whereby the D-genome parental expression state is taken in strong preference over the A-genome parental state (D-dominance: II = 4888 and 5719; XI = 4629 and 5257 for the F₁ hybrid and synthetic allotetraploid used in Rapp *et al.* (2009), respectively, versus A-dominance IV = 2264 and 663; IX = 1951 and 119 for the F₁ hybrid and synthetic allotetraploid used in Rapp *et al.* (2009), respectively). For the for the F₁ hybrid this D-dominance effect can also be observed at the homoeolog level, as there are more than twice as many D-genome biases as A-genome biases (334 versus 153; Table 2). The overrepresentation of these D-genome biases is largely reversed among all five natural allopolyploids, both at the homeolog level (Table 2) and among total gene expression profiles (Fig. 3). Over evolutionary time it appears that the allotetraploids begin to assume roughly equivalent numbers of A- and D-dominant states. In constrast to this trend, however, transgressive up- and down-regulation is far more frequent among the allopolyploids than it is among the F₁ hybrid and synthetic allotetraploid used in Rapp *et al.* (2009) (the values from Rapp *et al.*(2009) are as follows: transgressive-up: V = 81, VI = 238, and VIII = 102; transgressive-down: III = 27, VII = 23, and X = 19). From these results we can conclude that the instantaneous effect of genomic merger among the Gossypium A- and D-genomes is to create a significant level of D-genome dominance, regardless of ploidy level, whereas, over an evolutionary times scale, all five natural allotetraploid species have alleviated this D-genome control while at the same time exploring a large number of transgressive expression states. Because the diploid A- and D-genome species
used in this study are not the exact parents of the 1-2 million year old allotetraploids we cannot say definitively that the differences described above are not the result of a different ancestry. However, a significant body of evidence indicates that *G. arboreum* and *G. raimondii* are the best extant models for the parents of allotetraploid cotton (reviewed by: Wendel & Cronn (2003)) and that these diploid species are highly similar to the corresponding allopolyploid genomes at the sequence level (Senchina *et al.* 2003; Grover *et al.*, 2004; Grover *et al.*, 2007). Therefore, it is likely that our temporal findings are genuine and are unlikely to have arisen as an artifact the discrepancies between the model diploid progenitors used in this study and the actually parents of the *Gossypium* allotetraploids.

**Acknowledgments**

We gratefully acknowledge the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service (2005-35301-15700 to J.A.U and J.F.W.) and National Science Foundation Plant Genome Research Program (0638418 to J.F.W.) for their support. We also thank Nathan Springer and Bob Stupar for their help in developing the Sequenom platform used in validating our microarray results and the University of Minnesota BioMedical Genomics Center for processing all Sequenom assays. James McD. Stewart and David Stelly kindly generated and shared the F1 hybrid used in this study. Finally, we thank Dan Nettleton for statistical guidance.

**References**

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel


shock after interspecific hybridization in *Senecio* is ameliorated by genome duplication. *Current
Biology* 16: 1652-1659.

duplicated genes during development and evolution of a single cell in a polyploid


Leitch AR, Leitch IJ. 2008. Genomic plasticity and the diversity of polyploid

Nettleton D, Hwang JTG, Caldo RA, Wise RP. 2006. Estimating the number of true null
hypotheses from a histogram of *p* values. *Journal of Agricultural, Biological, and Environmental


Table 1:

*Gossypium* species used in this study.

<table>
<thead>
<tr>
<th>Species Name</th>
<th>Genome Designation</th>
<th>Accession</th>
<th>Ploidy level</th>
<th>Location of Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>G. arboreum</em></td>
<td>A\textsubscript{2}</td>
<td>cv. AKA-8410</td>
<td>diploid</td>
<td>Africa</td>
</tr>
<tr>
<td><em>G. raimondii</em></td>
<td>D\textsubscript{5}</td>
<td>Accession Unnamed</td>
<td>diploid</td>
<td>South America</td>
</tr>
<tr>
<td><em>G. hirsutum</em></td>
<td>AD\textsubscript{1}</td>
<td>cv. Maxxa</td>
<td>allotetraploid</td>
<td>Mexico/Central America</td>
</tr>
<tr>
<td><em>G. barbadense</em></td>
<td>AD\textsubscript{2}</td>
<td>cv. Pima S7</td>
<td>allotetraploid</td>
<td>South America</td>
</tr>
<tr>
<td><em>G. tomentosum</em></td>
<td>AD\textsubscript{3}</td>
<td>WT936</td>
<td>allotetraploid</td>
<td>Hawaii</td>
</tr>
<tr>
<td><em>G. mustelinum</em></td>
<td>AD\textsubscript{4}</td>
<td>Accession Unnamed</td>
<td>allotetraploid</td>
<td>South America</td>
</tr>
<tr>
<td><em>G. darwinii</em></td>
<td>AD\textsubscript{5}</td>
<td>Accession Unnamed</td>
<td>allotetraploid</td>
<td>Galapagos Islands</td>
</tr>
<tr>
<td><em>G. arboreum</em> X <em>G. raimondii</em></td>
<td>A\textsubscript{2}	extsuperscript{♀} \times D\textsubscript{5}	extsuperscript{♂}</td>
<td>Accession Unnamed</td>
<td>diploid</td>
<td>synthetic hybrid</td>
</tr>
</tbody>
</table>
Table 2:

Categorization of A- and D-genome biases and equivalent contribution to the transcriptome for 1383 homoeologous/allelic gene pairs including the estimate of true nulls (Est. True H₀; compare to the “Equivalent” category), and the intersection of biased gene lists for all species and for only the five allotetraploid species.

<table>
<thead>
<tr>
<th>Accession</th>
<th>A-biased</th>
<th>D-biased</th>
<th>Equivalent</th>
<th>Total</th>
<th>Est. True H₀</th>
</tr>
</thead>
<tbody>
<tr>
<td>F₁</td>
<td>153</td>
<td>334</td>
<td>896</td>
<td>1383</td>
<td>905</td>
</tr>
<tr>
<td>G. hirsutum</td>
<td>455</td>
<td>570</td>
<td>358</td>
<td>1383</td>
<td>504</td>
</tr>
<tr>
<td>G. tomentosum</td>
<td>552</td>
<td>666</td>
<td>165</td>
<td>1383</td>
<td>352</td>
</tr>
<tr>
<td>G. barbadense</td>
<td>486</td>
<td>720</td>
<td>177</td>
<td>1383</td>
<td>391</td>
</tr>
<tr>
<td>G. darwinii</td>
<td>373</td>
<td>441</td>
<td>569</td>
<td>1383</td>
<td>591</td>
</tr>
<tr>
<td>G. mustelinum</td>
<td>292</td>
<td>370</td>
<td>721</td>
<td>1383</td>
<td>730</td>
</tr>
<tr>
<td>all species intersection</td>
<td>30</td>
<td>79</td>
<td>8</td>
<td>117</td>
<td>NA</td>
</tr>
<tr>
<td>only allotet. intersection</td>
<td>176</td>
<td>208</td>
<td>12</td>
<td>396</td>
<td>NA</td>
</tr>
</tbody>
</table>
Fig. 1:

_Gossypium_ allotetraploid phylogeny and “expression phylogram”. (a) The phylogeny of the five _Gossypium_ allotetraploids, including an image of their flowers at maturity. (b) A phylogeny of the same species, where branch length represents the extent of homoeologous expression divergence among 1383 genes.
Fig. 2:

Histograms of homoeologous expression among a synthetic F1 hybrid and five *Gossypium* allotetraploids. All homoeolog expression values are expressed as log ratios (ln(Aprobe) – ln(Dprobe)), thus positive values indicate greater A-genome expression and negative values indicate greater D-genome expression.
Fig. 3:

F₁ hybrid and *Gossypium* allotetraploids expression states relative to their model maternal (*G. arboreum*) and paternal (*G. ramondii*) progenitors. Each expression category is labeled with a Roman numeral which corresponds to Rapp. *et al.* (2009), and includes a cartoon depiction of the category, where maternal (♀) and paternal (♂) states are on the edges and the polyploid or F₁ hybrid (labeled “Allo”) state is in the middle. Expression values on the same horizontal line indicate statistically equivalent expression, while expression values on higher or lower vertical lines represent statistically significant up- and down-regulation, respectively. Finally, under each species name we record the number genes which show statistical equivalence among both progenitors and the species in question.

### Counts of Allotetraploid/Parental Generic Expression Patterns

<table>
<thead>
<tr>
<th>Categories</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
<th>VII</th>
<th>VIII</th>
<th>IX</th>
<th>X</th>
<th>XI</th>
<th>XII</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>G. hirsutum</em></td>
<td>1581</td>
<td>3240</td>
<td>621</td>
<td>3496</td>
<td>553</td>
<td>523</td>
<td>1855</td>
<td>1563</td>
<td>4444</td>
<td>497</td>
<td>4016</td>
<td>1747</td>
</tr>
<tr>
<td>(No Change = 18323)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>G. hpestans</em></td>
<td>2067</td>
<td>3750</td>
<td>1097</td>
<td>3453</td>
<td>781</td>
<td>803</td>
<td>3502</td>
<td>2371</td>
<td>3320</td>
<td>733</td>
<td>3902</td>
<td>2153</td>
</tr>
<tr>
<td>(No Change = 14527)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>G. tomentosum</em></td>
<td>2064</td>
<td>2927</td>
<td>1489</td>
<td>3578</td>
<td>1030</td>
<td>1290</td>
<td>4387</td>
<td>3666</td>
<td>3857</td>
<td>1036</td>
<td>3403</td>
<td>2176</td>
</tr>
<tr>
<td>(No Change = 11856)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>G. mustelinum</em></td>
<td>1827</td>
<td>3498</td>
<td>978</td>
<td>3601</td>
<td>825</td>
<td>781</td>
<td>3281</td>
<td>2371</td>
<td>3844</td>
<td>800</td>
<td>3863</td>
<td>2024</td>
</tr>
<tr>
<td>(No Change = 14793)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>G. darwinii</em></td>
<td>2069</td>
<td>3050</td>
<td>1274</td>
<td>4010</td>
<td>924</td>
<td>991</td>
<td>4056</td>
<td>3199</td>
<td>3958</td>
<td>937</td>
<td>3217</td>
<td>2213</td>
</tr>
<tr>
<td>(No Change = 19561)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F₁</td>
<td>1581</td>
<td>4888</td>
<td>248</td>
<td>2264</td>
<td>69</td>
<td>168</td>
<td>452</td>
<td>302</td>
<td>1951</td>
<td>60</td>
<td>4629</td>
<td>1951</td>
</tr>
<tr>
<td>(No Change = 23878)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supporting Information Fig. S1:

Validation of homoeolog expression results for *G. mustelinum*, *G. tomentosum*, *G. barbadense*, and *G. darwinii*. We present a comparison of expression estimates from 14 randomly chosen genes. All NimbleGen (microarray) values are expressed as the log ratio \( \ln(A_{\text{probe}}) - \ln(D_{\text{probe}}) \), whereas the Sequenom (mass-spectrometry) values are expressed as the proportion of the transcriptome contributed by the A-genome. Thus both metrics result in analogous interpretations (ie. for both technologies, larger values reflect greater A-genome contribution to the transcriptome, and smaller values reflect greater D-genome contribution). Scatter plots of validation results include the best-fit line, Pearson’s \( r \) value, and \( p \)-value.
CHAPTER 4.

RECIPROCAL SILENCING, TRANSCRIPTIONAL BIAS AND FUNCTIONAL DIVERGENCE OF HOMOELOGS IN POLYPLOID COTTON (GOSSYPIUM)

A paper published in Genetics in 2009 (Genetics 182:503-517)

* These authors contributed equally to the work

ABSTRACT

Polyploidy is an important force in the evolution of flowering plants. Genomic merger and doubling induce an extensive array of genomic effects, including immediate and long-term alterations in the expression of duplicate genes (‘homoeologs’). Here we employed a novel high-resolution, genome-specific, mass-spectrometry technology and a well-established phylogenetic framework to investigate relative expression levels of each homoeolog for 63 gene pairs in 24 tissues in naturally occurring allopolyploid cotton (Gossypium L.), a synthetic allopolyploid of the same genomic composition, and models of the diploid progenitor species. Results from a total of 2,177 successful expression assays permitted us to determine the extent of expression evolution accompanying genomic merger of divergent diploid parents, genome doubling, and genomic coevolution in a common nucleus subsequent to polyploid formation. We demonstrate that 40% of homoeologs are transcriptionally biased in at least one stage of cotton development, that genome merger per se has a large effect on relative expression of homoeologs, and that the
majority of these alterations are caused by cis-regulatory divergence between the diploid progenitors. We describe the scope of transcriptional subfunctionalization and 15 cases of probable neofunctionalization among 8 tissues. To our knowledge, this study represents the first characterization of transcriptional neofunctionalization in an allopolyploid. These results provide a novel temporal perspective on expression evolution of duplicate genomes and add to our understanding of the importance of polyploidy in plants.

**INTRODUCTION**

Duplicate genes are widespread in genomes of almost all eukaryotes. Among flowering plants, polyploidy (whole genome duplication) is a primary source of duplicate genes (Soltis & Soltis, 1999; Wendel, 2000; Bowers et al., 2003; Lockton & Gaut, 2005). All flowering plants are either contemporary polyploids or harbor the evolutionary signature of paleopolyploidy (ancient polyploidy) in their genomes. Polyploidy may have influenced flowering plant diversification, as it provides raw material for the evolution of novelty by relaxing purifying selection on duplicate genes (Stephens, 1951; Ohno, 1970; Lynch & Conery, 2000; Wendel, 2000). Through genic redundancy, polyploids may be subject to an array of evolutionary processes, including subfunctionalization (evolution of partitioned ancestral functions among duplicate genes) and neofunctionalization (evolution of novel functions among duplicate genes).

Subfunctionalization and neofunctionalization have been demonstrated in several species (Force et al., 1999; Adams et al., 2003; Duarte et al., 2006; Cusack & Wolfe, 2007; Liu & Adams, 2007; Teshima & Innan, 2008). From an evolutionary perspective, both processes can lead to the preservation of the two members of a duplicate gene pair (Ohno, 1970; Lynch & Force, 2000). Because duplicate genes tend to be lost rapidly through mutational processes
(Lynch & Conery, 2000), subfunctionalization is thought to be most important shortly after gene duplication. As the age of the duplicate pair increases, neofunctionalization becomes increasingly likely (Ohno, 1970). Further linking these two processes, it has been suggested that subfunctionalization could serve as a preservational transition state leading to neofunctionalization (Rastogi & Liberles, 2005). Thus following polyploidy, both subfunctionalization and neofunctionalization may make significant contributions to duplicate gene retention and functional diversification.

In addition to subfunctionalization and neofunctionalization, allopolyploid plants also generate diversity through rapid genomic changes at various levels, including chromosomal lesions and intergenomic exchanges, as in wheat (Shaked et al., 2001), Brassica (Song et al., 1995; Pires et al., 2004; Udall et al., 2004), and Arabidopsis (Madlung et al., 2002), epigenetic modifications (Lee & Chen, 2001; Madlung et al., 2002; Wang et al., 2004; Salmon et al., 2005; Gaeta et al., 2007) and gene expression changes (Adams et al., 2003; Wang et al., 2004; Bottley et al., 2006; Adams, 2007; Flagel et al., 2008). It is thought that these changes result from “genomic shock” caused by the joint effects of genome merger and genome doubling during allopolyploid formation (Adams et al., 2004; Hegarty et al., 2006; Flagel et al., 2008). Additionally, allopolyploidy entails combining homoeologous regulatory variation and may lead to expression variation through interacting cis- and trans-regulatory factors, as has been shown for allelic variation (Wittkopp et al., 2004; Stupar & Springer, 2006; Swanson-Wagner et al., 2006). Collectively, these results demonstrate that both genomic and genic evolutionary processes play a role in allopolyploid evolution.

The cotton genus (Gossypium) is a useful system to study the extent of genomic changes that accompany genome merger and allopolyploidization (Wendel & Cronn, 2003).
Allotetraploid cottons were formed by the merger of two diploid species originating, respectively, from the cotton A- and D-genome groups. This event took place 1-2 million years ago (Percy & Wendel, 1990; Wendel & Albert, 1992; Seelanan et al., 1997; Cronn et al., 2002; Senchina et al., 2003) (Figure 1A). The modern diploid species *G. arboreum* (A-genome) and *G. raimondii* (D-genome) are extant diploids most similar to the ancestral A- and D-genome diploids involved in the formation of natural allotetraploids (Percy & Wendel, 1990; Wendel & Albert, 1992; Seelanan et al., 1997; Cronn et al., 2002; Senchina et al., 2003) (Figure 1A).

Following formation, the allotetraploid lineage diverged into five extant species. Furthermore, F₁ hybrids and allotetraploids synthetically derived from A- and D-genome species mergers are also available (Figure 1A and Table 1). These synthetic accessions have proved particularly useful in teasing apart the effects of genome merger and genome doubling during the formation of the natural allopolyploid (Adams et al., 2004; Adams & Wendel, 2005a; Flagel et al., 2008). Although these studies and others (Comai et al., 2000; Adams et al., 2004; Soltis et al., 2004; Hegarty et al., 2006; Tate et al., 2006; Chen, 2007) have provided insights into the formation and immediate genetic consequences of polyploidy, there is still much to be learned about stabilization and evolution of polyploid genomes following formation.

In the present study we employ a genome-specific, mass-spectrometry technology to study relative levels of allelic and homoeologous (gene pairs duplicated by polyploidy) gene expression in diploid and allopolyploid cotton. By contrasting allelic and homoeologous gene expression in cotton species within an appropriate phylogenetic framework (Figure 1A), we have detected expression patterns consistent with subfunctionalization and neofunctionalization (Figure 1B). Because the cotton accessions selected for the present study represent three successive stages in allopolyploid evolution, i.e., genomic merger of divergent parents, genome
doubling, and finally genomic coevolution in a common nucleus, we were able to determine the extent of expression evolution accompanying each stage.

**MATERIALS AND METHODS**

**Maintenance of cotton germplasm and tissue collection**

*Seedling tissues.* Seeds of two diploid cottons, *G. arboreum* (A2) and *G. raimondii* (D5), and a natural (*G. hirsutum* L. cv. Maxxa) and synthetic (2(A2 x D3)) allotetraploid cotton (*Table 1*), were sown and grown in steamed potting mix in the Pohl Conservatory at Iowa State University at 24°C day / 20°C night with a photoperiod of 16h light / 8h dark. The synthetic allotetraploid cotton was formed by colchicine-doubling the hybrid resulting from a cross between A2 and the D-genome species *G. davidsonii* (D3). Three biological replicates were planted for each species and seedling stage tissues were sampled at 10 days post emergence. Additionally, a sterile F1 hybrid (A2 x D5) population has been maintain through vegetative propagation, and was also sampled for some tissues. The above accessions include representatives of both diploid progenitor genomes (A- and D-genomes), their synthetic F1 hybrid and synthetic allotetraploid, and a natural allopolyploid cotton (Wendel & Cronn, 2003) (*Figure 1A*). All seedlings were sampled between 9 AM to 10 AM to minimize circadian effects, and tissues were flash-frozen in liquid nitrogen and stored at -80°C prior to RNA isolation.

*Vegetative and floral tissues.* Seedlings were grown for 3-5 weeks before transfer to larger pots and maintained at 32°C and a photoperiod of 16h light / 8h dark. After the emergence of the fifth leaf, the first, third and fifth leaves were harvested from all five taxa on the same day and flash-frozen immediately and stored. Petioles were sampled from the fifth leaf of each biological
replicate, and midrib and lamina tissues were harvested from young and newly emerged leaves at the same time. After 3-4 months, flowers from all species, except D₅, were harvested on 0 dpa (days post anthesis; 0 dpa is the day the flower opened). Juvenile plants from D₅ were grown separately under a shade regime for approximately one month, a treatment necessary to induce flowering. Fully opened flowers were collected between 9 AM and 11 AM to mitigate circadian effects. All flower tissues were manually excised and immediately flash-frozen in liquid nitrogen.

*Fiber.* Plants from all five taxa were grown in three replicates in the Horticulture Greenhouse at Iowa State University and flowers were harvested for four different stages of fiber development (5, 10, 20, and 25 dpa). For each replicate and developmental time-point, ovules were excised and immediately frozen in liquid nitrogen. Ovules were visually inspected for cell damage and fibers were inspected for contaminating tissue.

**Isolation of total RNA and sample platform preparation**

*RNA isolation.* All 24 tissues (**Table 2**) from the five taxa and three biological replicates were collected in 1.7 ml microfuge tubes. RNAs were extracted from all seedling, vegetative and floral tissues using a modified Qiagen RNA extraction protocol according to the manufacturer’s instructions (Qiagen Inc., Valencia, CA) with modifications as follows: Tissues were ground in fresh XT buffer (Wan & Wilkins, 1994) in microfuge tubes with plastic pestles and incubated at 42°C for 1.5 hr. Then 2M KCl was added and the sample was incubated on ice for 1 hr. After incubation, the samples were transferred to Qiashredder columns supplied with the Qiagen Plant RNeasy kit and all subsequent steps followed this kit’s protocol.
RNAs were extracted from fibers at each developmental time point using a liquid nitrogen/glass bead shearing approach following a lithium chloride hot borate protocol (Hovav et al., 2007; Taliercio & Boykin, 2007). Purified RNA samples were quantified using a NanoDrop spectrophotometer (NanoDrop Technologies, Wilmington, DE) and assayed for degradation using a BioAnalyzer (Agilent, Palo Alto, CA).

**cDNA preparation.** A total of 307 tissue samples were used for RNA isolations, each yielding approximately 5 µg of total RNA. All RNA samples were treated with DNase following the manufacturer’s protocol (New England Biolabs (DNase I, M0303S)), and assayed for genomic DNA contamination by PCR amplification with primers flanking intron eight of a *Gossypium* RNA helicase with high similarity to the gene At4G00660 in *Arabidopsis* (GenBank Accession NM_179204). Following DNase treatment, cDNAs were synthesized using Superscript III reverse transcriptase, according to the manufacturer’s instructions (Invitrogen, Carlsbad, CA). In tissues with surplus RNA yields, cDNAs were also synthesized from equal RNA mixes of A2 and D5 accessions. These mixes served as an *in vitro* model for mid-parent expression within the allopolyploid and hybrid accessions.

**Probe selection for multiplex PCR.** MALDI-TOF mass-spectrometry assays for genome-specific expression were designed for the Sequenom (San Diego, CA) MassARRAY platform. Genes for this platform were selected from 1231 cotton ESTs contigs (Udall et al., 2006), derived from A2, D5, and AD1 accessions. These contigs were inferred to represent homoeologous relationships in AD1 based on comparisons to orthologous sequences from the A- and D-genome diploids. This led to the identification of genome-specific SNPs, which were processed using the Sequenom probe selection software. From these results we selected four multiplexes, each including 29 genes.
**Genome-specific expression assays.** For each multiplex, forward and reverse primers from all 29 genes were pooled and used to amplify each cDNA sample using the manufacturer’s specifications (Sequenom Inc., San Diego, CA). Amplified cDNAs were visualized on agarose gels to confirm amplification, and loaded on a 384 well plate in three technical replicates. Mass spectrometry quantification of genome-specific expression ratios was performed at the University of Minnesota Genotyping facility.

**Data processing, filtering, and analysis**

*Identification of diagnostic assays.* All expression data recovered from the MassARRAY process were first filtered based on internal measures of assay quality, which included removing all assays flagged as “Bad Spectra”, or having a frequency of uncertainty > 0.2 or an unused extension primer frequency > 0.5. Next all genes were filtered based on assays of A₂ and D₅ DNA samples, which were mixed in known ratios (4:1, 2:1, 1:1, 1:2, and 1:4), and used to standardize the genome-specific quantification procedure for each gene (Stupar & Springer, 2006). All genes were required to display a strong correlation ($R^2 > 0.9$) between the expected and observed A₂:D₅ DNA ratios. Additionally, DNAs from Maxxa, the synthetic polyploid, and the F₁ hybrid were also assayed as controls for lineage-specific SNPs, which could potentially arise in these accessions, with the expectation that good assays would yield approximately 1:1 A-to D-genome values. Maxxa, synthetic, and F₁ hybrid assays were excluded if these DNA control values exceeded the expected 1:1 ratio by ± 25%. Following filtering, a maximum of 9 replicates (3 biological X 3 technical) could potentially be recovered for each assay. Each of these replicate pools represents one gene by tissue, and evaluates the proportion of A- and D-genome contribution to the transcriptome. These values were stored as the percent D-genome
contribution to the transcriptome (% D) and outlier replicates were identified and removed if they deviated from the median of the replicate pool by +/- 50 % D. Next the mean % D and standard deviation of the remaining values was recorded and used for all subsequent analyses, this complete data set can be found as supplemental file (See Supplemental Material.).

Statistical contrasts of genome-specific expression ratios. Contrasts of A- and D-genome expression ratios were made using a t-test. P-values were then converted to q-values using the method of Storey and Tibshirani (2003), and individual contrasts were considered equivalent when q > 0.05.

RESULTS

Assessment of Sequenom MassARRAY performance

Using a single nucleotide polymorphism (SNP)-based Sequenom MassARRAY technology, we simultaneously assayed the A- and D-genome contribution to the transcriptome for 63 gene pairs (Supplemental Table 1). These A- and D-genome gene pairs are hereafter termed ‘homoeologous’ in the polyploid genotypes and ‘allelic’ in the diploid F1 hybrid genotypes (note however, in the F1 hybrid chromosomes from the A- and D-genome chromosome pairing is limited (Endrizzi et al., 1985)). The MassARRAY technology has previously been shown to be effective in determining the relative allelic transcript levels in hybrid maize (Stupar & Springer, 2006). Assays of cotton A- and D-genome expression were made possible by the availability of A- and D-genome-specific SNPs obtained from cotton EST contig assemblies (Udall et al., 2006), which included transcripts from the diploid members of the A- and D-genome (A2 and D5) and an allotetraploid (AD1). The presence of a genome-
specific SNP alters the molecular weight such that the MassARRAY platform can distinguish either variant from a mixed transcript pool and estimate relative abundance.

Expression assays were filtered using a rigorous quality-control protocol (see Materials and Methods), yielding the total number of successful assays summarized in Table 2. The percentage of successful assays varied among tissues from a maximum of ~73% in petioles to a minimum of 5% in pollen and hypocotyl tissues (Table 2 and Supplemental Figure 1). Among 63 genes and 24 tissue types examined, 660 and 646 gene by tissue combinations were successful in the natural (‘Maxxa’ hereafter) and synthetic (‘synthetic’ hereafter) allopolyploids.

Patterns of genome-specific gene expression in cotton tissues

The primary goal of the present study was to quantify genome-specific expression among a sampling of cotton tissues and developmental conditions in an evolutionary context. This was accomplished by assaying 24 tissues or developmental stages, which fit into four general categories: seedling, vegetative, and floral tissues, as well as developing fibers (Table 2). For each of these categories, genome-specific expression values were extracted for the mix, F1, and the synthetic and natural (Maxxa) allopolyploids and binned into five groups, using the percent D-genome expression as a metric (0-20% D, 20-40% D, 40-60% D, 60-80% D, and 80-100% D) (Figure 2). In the F1, synthetic, and Maxxa, biases indicate differential gene expression between the A- and D-genome transcripts within the same nucleus, whereas in the mix, which pools two
biologically different species, a bias reflects differential gene expression between the A2 and D5 parents.

Overall, F1, synthetic, and Maxxa show an A-genome bias in seedling and vegetative tissues, but in floral tissues the mix shows a D-bias whereas the F1 and Maxxa show an A-genome biases and the synthetic is nearly equivalent (Figure 2). For ‘floral’ samples the mix is represented by only ovary wall and ovule tissues, though both individually support a D-bias. Fiber expression in the mix and Maxxa show a substantial level of A-genome bias, whereas the synthetic is less A-genome biased (Figure 2; note that the F1 hybrid between A2 and D5 is sterile and hence fibers could not be studied). These expression patterns are interesting, as they highlight previous observations (Adams et al., 2003; Adams et al., 2004; Adams & Wendel, 2005a; Udall et al., 2006; Yang et al., 2006; Flagel et al., 2008; Hovav et al., 2008b) that neither the A- or D-genome is globally dominant with regard to genome-specific expression.

These general trends describe the overall patterns of expression states for this sampling of genes and tissues. At the individual gene level there is considerable variation. An interesting example is the gene CO131164 (a putative phytochrome-associated protein), which shows highly variable expression among tissues and accessions. In Maxxa this gene demonstrates nearly complete A-genome expression in anthers and complete D-genome expression in ovary wall (Figure 3A), indicative of developmentally regulated reciprocal silencing of alternative homoeologs in different parts of the same flower (cf. Adams et al. (2003)). Additionally, shortly after fiber initiation (5 dpa), CO131164 is strongly A-genome biased in the synthetic, though Maxxa shows approximately equivalent expression (Figure 3A). Another illustrative gene is CO130747 (a putative CBL-interacting protein kinase), which shows significant differences in tissue-specific homoeolog expression between Maxxa and the synthetic during many
developmental stages (Figure 3B). The synthetic is more A-genome biased in seedling, vegetative and floral stages, including almost total A-genome expression in roots, petioles, the calyx and all four developmental stages of fiber. In contrast, Maxxa is only strongly A-genome biased in 5 and 10 dpa fibers. A third example gene illustrated (Figure 3C) is DW008528 (similar to a putative protein with unknown function in Arabidopsis thaliana), for which we observed equivalent A- and D-genome homoeolog expression in all tissues for the synthetic and the F1, but considerable expression variation for vegetative tissues in Maxxa. In all fiber stages studied, both Maxxa and the synthetic show nearly equal expression of homoeologs.

**Genome-specific expression biases during genome merger and doubling**

The accessions studied were selected to provide insight into the various stages involved in allopolyploid speciation, including diploid divergence, genome merger, genome doubling, and subsequent evolution and stabilization. To assess homoeolog transcriptional alteration accompanying each of these stages, we identified all gene × tissue combinations shared by all four accessions (mix, F1, synthetic, and Maxxa), as well as those just shared by the F1, synthetic and Maxxa, and finally just by the synthetic and Maxxa. For each of these groups, we assigned all gene × tissue relationships as either equivalent (“=”; q-value > 0.05) or nonequivalent (“≠”; q-value ≤ 0.05). Specific examples of several expression patterns and their biological interpretation can be found in Figure 4.

As summarized in Table 3, when comparing all four accessions, the category that induced the most expression alteration was genome merger (implicated in 23 + 16 + 9 + 9 = 57 gene × tissue events) followed by change due to polyploid evolution (implicated in 9 + 16 + 6 + 9 = 40 gene × tissue events). From these results, it is clear that genome merger and polyploidy
evolution (subsequent to formation) have the greatest effect on homoeologous gene expression, though diploid divergence and genome doubling are implicated in 11 and 30 gene × tissue events, respectively. For genes lacking data from the mix sample, more homoeolog expression changes occurred due to polyploid evolution than polyploidy alone, corroborating the foregoing result. Alternatively, some of the above observations could be due to the divergence between the model diploid progenitors used in this study and the actual ancient parents of natural allotopolyploid cotton. Similar findings have been reported in cotton and other polyploid systems regarding the relative importance of genome merger (Adams & Wendel, 2005a; Wang et al., 2005; Hegarty et al., 2006; Flagel et al., 2008) and genome doubling (Stupar et al., 2007). However, to our knowledge, this is the first study wherein the specific effects of each of these four components (divergence, merger, polyploidy, and polyploidy evolution) have been disentangled.

**Tissue-specific subfunctionalization and gene silencing**

To address the prevalence of subfunctionalization between homoeologous genomes, we searched for patterns of highly differential homoeolog expression biases between tissues from the F1, synthetic and Maxxa (see Figure 1B). We did not detect any cases of complete reciprocal homoeolog silencing (here silencing is operationally defined as the absence of detectable transcript) among the 63 genes assayed, but the most subfunctionalized genes and their respective tissues are listed in Table 4. In Maxxa, the most striking example is the gene CO131164, where the A-genome homoeolog has been silenced in the ovary wall, but the reverse is observed in anthers, where the A-genome homoeolog accounts for 93% of homoeologous expression. Other genes showed similar patterns of subfunctionalization in various tissues.
Interestingly, among those genes displaying the largest degree of expression subfunctionalization, it appears that reproductive tissues such as anthers, style/stigma, staminal tube and ovary wall are often involved (these tissues comprise 12 of 18 tissues in Table 4). This observation mirrors similar findings from (Adams et al., 2003).

In addition to subfunctionalization, hybrid and polyploid plants also display genome-specific silencing biases. For each genotype, the percentage of completely silenced genes varied from a maximum of ~ 6% D-homoeolog silencing in Maxxa to a minimum ~ 0.3% A-homoeolog silencing in the synthetic (Table 5). In most cases, complete silencing remains in each subsequent stage along the pathway to allopolyploidy. For example, the gene CAO23634 (a putative S-formylglutathione hydrolase), is D-silenced in petioles and apical shoot meristems in the F1, and this silencing remains in the synthetic and Maxxa polyploids (Figure 5). However, there are also counterexamples, such as the gene CO130747. This gene is D-silenced in petioles of the synthetic but the D-genome is once again expressed in Maxxa (Figure 5). Another example is gene CO131164, where there is silencing of the A2 diploid in ovary walls, but this gene is expressed in the F1 and synthetic, and then once again the A-genome is silenced in Maxxa (Figure 5).

**Tissue-specific transcriptional neofunctionalization**

Neofunctionalization may be detected in our framework by first indentifying all gene × tissue assays that lack expression of either the A- or D-genome ortholog in the mix (i.e. not expressed in the A2 or D5 parent) and which gain expression in the F1, synthetic, or Maxxa (Figure 1B). It is important to note that the pattern above can arise *de novo*, as a totally novel form of expression, or as a product of the reactivation of a lost ancestral expression regime, and
our experiment cannot distinguish between these two forms of transcriptional neofunctionalization.

Using the criteria above, a total of 15 genes across 8 different tissues exhibit transcriptional neofunctionalization. Additionally, by observing the range of expression values for the 1:1 parental mixtures in all available genes (Supplemental Figure 2), it appears unlikely that these cases of neofunctionalization are a product of an inaccurate mix. Among the neofunctionalized genes, 10 showed substantial contributions from both genomes in the F1, synthetic and Maxxa, reinforcing the presence of gene expression neofunctionalization (Table 6). Genes CO108066 (a putative glyceraldehyde-3-phosphate dehydrogenase), and CO076921 (a putative vacuolar ATP synthase catalytic subunit) show lack of expression of either the A or D orthologs, respectively, in leaf lamina, but both homoeologs are expressed in the same tissue in the F1, synthetic, or Maxxa (Table 6). In addition, in those cases where neofunctionalization has occurred, it has been maintained in all genomically merged samples (F1, synthetic and Maxxa; Supplemental Figure 3). Overall, the nonfunctional alleles were usually from the diploid A-genome (11 of 14 cases), indicative of the potential for a genome-of-origin bias for neofunctionalization in cotton, albeit for a relatively small sampling of genes.

_Evolution of cis- and trans-regulatory variations in cotton_

Expression variation can originate via either cis- or trans-regulatory evolution, or both. By comparing genome-specific expression between the mix and F1 it is possible to partition expression variation into cis and trans origins, using the procedures described by Wittkopp et al. (2004) and Stupar and Springer (2006). Our analysis of cis- and trans-acting regulation in cotton includes 30 genes in leaf lamina and 38 genes in the petiole (Figure 6). Among both leaf
lamina and petiole tissues the most prevalent type of regulatory divergence is cis-regulatory evolution (50% and 39% in lamina and petiole, respectively) followed by a combination of cis and trans factors. This result is similar to other studies regarding the prevalence of these modes of regulatory evolution (Wittkopp et al., 2004; Stupar & Springer, 2006; Zhuang & Adams, 2007; Springer & Stupar, 2007a). Additionally this result gives an indication that some of the expression changes attributed to genome merger (Table 3) are likely caused by cis-regulatory divergence between the A- and D-genomes.

**DISCUSSION**

**Homoeologous contributions to the transcriptome**

We used a mass-spectrometry based SNP detection technique to measure allele- and homoeolog-specific contributions to the transcriptome of diploid and allopolyploid cotton accessions that were selected to be informative with respect to the evolutionary stages involved in allopolyploid speciation and subsequent evolution (Figure 1A). Although the representative progenitor diploid species used in the present study (A2, D3, and D5) are not the actual parents of natural allopolyploid cotton, which formed 1-2 million years ago, a substantial body of evidence indicates that they represent close approximations (reviewed in Wendel and Cronn 2003). Furthermore, to evaluate differences between D3 and D5 (and as a corollary species-specific biases associated with the 2(A2 x D3) synthetic allotetraploid) we compared expression between these species from 18 randomly selected genes in petiole tissues and 17 genes in leaf tissues. These comparisons were made relative to a common A2 reference sample and were conducted using the Sequenom platform following the procedures outlined in the Materials and Methods. These experiments show that D3 and D5 are similar in their expression, having an average
expression difference of 15.5% among the 35 comparisons (Supplemental Figure 4). For comparison, the average variation between biological replicates within D3 and D5 was 12.5%, meaning that within species variation was ~ 81% of the level of the difference between D3 and D5. These results indicate that species-specific differences between D3 and D5 are small.

Contrasting genome-specific expression in these accessions allowed us to allocate expression alterations to the stages of genome merger, genome doubling, and subsequent evolution within the allopolyploid lineage, while revealing examples of subfunctionalization and neofunctionalization (Figure 1B).

To substantiate the MassARRAY-based interpretations, we validated these estimates of genome-specific expression through comparisons to expression data generated by a genome-specific microarray platform (Udall et al., 2006). These validations were conducted for both petals (Flagel et al., 2008) and fibers from several developmental stages (Hovav et al., 2008a), and demonstrate significantly positive correlations.

Allopolyploidy entails the merger of two diploid genomes, which may contribute either equally or disproportionately to the transcriptome. Data presented here demonstrates that genomically biased expression in cotton is a common phenomenon, occurring in vegetative and floral tissues, and also in single-celled fibers, consistent with previous studies using other genes and analytical methods (Adams et al., 2003; Adams et al., 2004; Adams & Wendel, 2005a; Udall et al., 2006; Yang et al., 2006; Flagel et al., 2008; Hovav et al., 2008a). In the present study, among 49 homoeologous genes sampled in Maxxa, ~ 40% exhibit biased expression towards the A- or D-homoeolog, in all tissues examined (Figure 2). Furthermore, the extent of genome-
specific bias varies substantially among tissues, from nearly equal expression to complete silencing (here again, silencing refers to an absence of detectable transcript). The accumulated results from this study and others noted above indicate that among hybrid and allopolyploid cotton both the A- and D-genome contribute unequally to the transcript pool, but that neither genome displays an overall expression preference. This result differs from natural and synthetic allotetraploids in *Arabidopsis*, which show a global down-regulation of the *A. thaliana* genome in favor of the *A. arenosa* genome (Wang *et al.*, 2006; Chen *et al.*, 2008).

Although genomic preference was not detected at a global scale, relative transcript abundance from individual genes varied greatly. Genome-specific silencing was observed in 4 genes in the F1 hybrid, 5 genes in the synthetic, and 11 genes in Maxxa, noting that this differed widely among tissue types for many of those genes (Table 5). These results indicate that silencing is most prevalent in the natural allopolyploid, following 1-2 MY of allopolyploid evolution. Furthermore, in Maxxa, silencing is more prevalent among D-genome homoeologs than among A-genome homoeologs (Table 5). Both of these findings regarding the enhancement of silencing in Maxxa and a greater level of D-genome silencing mirror the findings of Flagel *et al.* (2008), though their study was limited to only petal tissues. Though the phenotypic effects of homoeolog silencing in cotton are unknown, it is possible that tissue-specific homoeolog silencing has had an impact on the evolution of allotetraploid cotton. For the *AdhA* gene in *G. hirsutum*, Liu and Adams (2007) have shown that homoeologous expression biases can occur as a response to abiotic stress. These findings of altered homoeologous expression patterns in response to genomic stress may hint at the adaptive potential of polyploidy. In this vein, our findings shed additional light on the extensive breadth and diversity of homoeolog expression patterns in natural allotetraploid cotton.
Distinguishing the effects of genome merger, genome doubling, and polyploid evolution on gene expression

By partitioning genome-specific expression changes within a selected framework of cotton accessions (Figure 1A), we were able to determine that genome merger has the largest impact on biased expression of homoeologs along the pathway to polyploidy in cotton (Table 3). Allelic expression differences, detectable immediately in the F1 hybrid, likely arise as a result of the merger of the divergent regulatory machinery of the A- and D-genomes within cotton. As many expression biases are shared with ancient allopolyploid cotton, the early establishment of expression patterns may play a role in gene expression evolution during the formation and subsequent evolution of natural cotton allopolyploids (Adams, 2007; Chen, 2007). Similar results have been previously noted in cotton (Adams & Wendel, 2005a; Flagel et al., 2008), as well as Senecio (Hegarty et al., 2006) and Brassica (Albertin et al., 2006). These authors all found that a considerable portion of gene expression alteration took place at the F1 hybrid stage when compared to re-synthesized allopolyploids. In Senecio and Brassica the effect of genome merger was, in fact, found to contribute a majority of the observed expression changes. Hegarty et al. (2006) classified this result as an example of “genomic shock”, a phenomenon which has been often observed in plant hybrids, but remains poorly understood at the molecular level. Some insight may derive from estimating the relative roles of cis- and trans-regulation within the F1 (Figure 6), and in this respect our data indicate that cis evolutionary factors (those arising from A- and D-genome cis-regulatory divergence), appear to be most prevalent. Taken together, these data indicate that reuniting divergent cis-regulatory domains may be a major component of “genomic shock” as it pertains to cotton hybrids and allopolyploids.
Following genomic merger, we found that allopolyploid evolution was the next most prevalent contributor to expression evolution (Table 3). This result is interesting as it implicates a significant role for the action of long-term evolutionary processes, such as sub- and neofunctionalization. Furthermore, changes that occur via allopolyploid evolution are more prevalent than those occurring via genome duplication alone (40 vs. 30 gene × tissue events; Table 3). This result indicates that genomic duplication alone may play a less significant role in altering homoeologous gene expression states in cotton, possibly affecting only those homoeologs with dosage-regulated expression (Osborn et al., 2003).

**Mechanisms of functional divergence and retention of homoeologs following allopolyploidy**

Tissue-specific and developmental expression variation between co-resident genomes may occur via several mechanisms, including altered regulatory interactions, epigenetic modifications, and gene dosage changes (Comai et al., 2000; Birchler et al., 2003; Osborn et al., 2003; Riddle & Birchler, 2003; Adams & Wendel, 2005b). At present, we lack an explanation of the underlying mechanisms of allelic and homoeologous gene expression biases, though our results indicates that both short (genome merger) and long-term (duplicate gene evolution) evolutionary processes play a role in determining homoeolog expression states in allopolyploid cotton. Recent work in allotetraploid Arabidopsis has shown that genome-specific methylation may play a crucial role in establishing homoeolog expression patterns (Chen et al., 2008). Using RNAi to silence met1, a cytosine methyltransferase, CHEN et al. (2008) demonstrated that many previously identified cases of genome-specific gene silencing were caused by or connected to methylation. Though these results may offer a promising mechanistic explanation of our findings of genome-specific biased expression and silencing, changes in methylation do not
appear to accompany allopolyploidy in cotton (Liu et al., 2001). This difference between *Arabidopsis* and cotton indicates that there may be no single unifying factor that governs genome-specific expression biases in allopolyploid plant species; instead genome-specific expression evolution may occur via a unique and *ad hoc* mixture of genetic and epigenetic regulatory mechanisms within different species.

Following allopolyploid establishment, several mechanisms may affect the fate of homoeologous genes (Leitch & Bennett, 1997; Matzke et al., 1999; Wendel, 2000; Levy & Feldman, 2002; Liu & Wendel, 2002; Soltis et al., 2003; Comai, 2005; Chen & Ni, 2006). One model of homoeologous gene retention is subfunctionalization, which is the partitioning of ancestral function and/or expression domains between duplicated genes, such that both copies continue to be necessary (Ohno, 1970; Force et al., 1999; Lynch & Force, 2000). Various studies of subfunctionalization, including MADS-box genes in *Arabidopsis* (Duarte et al., 2006), germin genes in barley (Federico et al., 2006), ZMM1 and ZAG1 genes in maize (Mena et al., 1996), and the *AdhA* gene in cotton (Adams et al., 2003), have shown that expression subfunctionalization occurs in plants. Here we show that instantaneous expression subfunctionalization may occur immediately following genomic merger (Table 4). Because of this, the preservational forces of subfunctionalization may be immediately initiated for a significant number of genes within allopolyploid cotton, as previously suggested (Adams et al., 2003; Adams & Wendel, 2005a; Flagel et al., 2008). Recent genomic analyses comparing homoeologous regions in *G. hirsutum* lend support to this claim, as homoeologous gene loss appears to be rare (Grover et al. 2004; Grover et al. 2007).

During allopolyploid evolution, duplicate genes not subject to subfunctionalization may still be retained if one copy evolves a novel function via neofunctionalization (Force et al., 1999;
Several studies have identified neofunctionalization among duplicate genes in diploid plants, including lectins in legumes (Van Damme et al., 2007), MADS-box genes in Physalis (He & Saedler, 2005) and Arabidopsis (Duarte et al., 2006), LEAFY paralogs in Idahoa scapigera (Brassicaceae) (Sliwinski et al., 2007), and diterpene synthase paralogs in conifers (Keeling et al., 2008). Expression neofunctionalization was also detected in the present study, which makes this the first example of neofunctionalization in an allopolyploid, as far as we are aware. We found 15 genes in 8 different tissues where expression was undetectable in one of the parental diploids but appeared in the F1, synthetic, and Maxxa. This pattern, which indicates an expansion of ancestral expression domains, is consistent with expression neofunctionalization.

In addition to the processes described above, cis- and trans-regulatory changes provide insight into the evolution of regulatory networks in cotton. We observed that most variation in gene expression following genome merger is the result of cis-regulatory variation. This finding suggests a mechanism for additive expression patterns detected for many genes in a microarray study of the F1 hybrid (Flagel et al., 2008). Additionally, cis-regulatory variation has been found to be a prevalent mechanism for generating expression differences in F1 maize hybrids (Stupar & Springer, 2006; Swanson-Wagner et al., 2006). While cis-regulatory evolution may be more common, it is also possible that trans-regulatory effects may affect gene expression, and even profoundly so. For example, reactivation of a silenced gene copy in a hybrid background, due to a trans-effect, may generate novel expression cascades that have evolutionary consequences. Mechanistic studies that determine the exact nature of important cis changes would be of tremendous help in advancing our understanding of underpinnings of the observation of a prevalence of cis-regulatory in the divergence in hybrid and allopolyploid plants.
Evolutionary consequences of homoeologous gene expression in cotton

Recurrent polyploidization has played a significant role in adding genetic variation to the genomes of plant species. It has been demonstrated that a most duplicate genes are lost quickly on evolutionary times scales (Lynch & Conery, 2000; Kellis et al., 2004; Thomas et al., 2006). Despite these rapid losses some homoeologous genes are retained, and various explanations have put forth to explain this retention, including dosage sensitivity (Thomas et al., 2006) and gene function (Blanc & Wolfe, 2004). For example, among the retained homoeologs in A. thaliana, transcription factors and signal transduction genes have been preferentially retained, whereas genes performing enzymatic functions have not (Blanc & Wolfe, 2004). It has also been suggested that alteration in duplicate gene expression patterns may enhance retention (Adams et al., 2003; Flagel et al., 2008). In cotton, this form of duplicate gene retention may be facilitated by expression subfunctionalization and neofunctionalization. These forms of divergence can occur rapidly after polyploidization; indeed we show here that many changes occur immediately in synthetic F1 hybrids and allopolyploids. From an evolutionary perspective, this immediate form of expression divergence can enhance expression variation and phenotypic diversification in the short-term with the long-term consequence of homoeolog retention. Together these processes may add to genetic and phenotypic variation with a species, thus enhancing the future potential for natural selection to lead to adaptive evolution.

Acknowledgements

This project was supported by the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service, grant number 2005-35301-15700. B.C. received
financial assistance from the Department of Biotechnology, Government of India. L.F. received financial assistance through a graduate fellowship from the Plant Sciences Institute at Iowa State University. Additionally the authors thank two anonymous reviewers for their helpful comments.

REFERENCES


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


Table 1. Details of plant materials used. The natural allotetraploid (*G. hirsutum*) was derived from hybridization, 1-2 MYA, between diploid A- and D-genome species most similar to the modern species *G. arboreum* and *G. raimondii*. The cytoplasmic donor of *G. hirsutum* is its A-genome parent and thus the F1 cross was created in the same direction, with A2 as the maternal parent. The synthetic allotetraploid was created by crossing A2 and D3 diploid parents followed by genome doubling through colchicine treatment.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Genome Designation</th>
<th>Accession</th>
<th>Ploidy Level</th>
<th>Location of origin</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>G. arboreum</em></td>
<td>A2</td>
<td>AKA-8401</td>
<td>Diploid</td>
<td>Africa</td>
</tr>
<tr>
<td><em>G. raimondii</em></td>
<td>D5</td>
<td>Jfw</td>
<td>Diploid</td>
<td>Peru</td>
</tr>
<tr>
<td><em>G. arboreum</em> X <em>G. raimondii</em></td>
<td>(A2 x D3)</td>
<td>NA</td>
<td>Diploid</td>
<td>Laboratory</td>
</tr>
<tr>
<td><em>G. hirsutum</em></td>
<td>AD1</td>
<td>cv. Maxxa</td>
<td>Tetraploid</td>
<td>Mexico/Cen. Am</td>
</tr>
<tr>
<td><em>G. arboreum</em> X <em>G. davidsonii</em></td>
<td>2(A2 x D3)</td>
<td>NA</td>
<td>Tetraploid</td>
<td>Laboratory</td>
</tr>
</tbody>
</table>
Table 2. Number of successful genome-specific assays calculated for all genes in all tissues.

<table>
<thead>
<tr>
<th>Tissues</th>
<th>Number of genes successfully assayed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mix</td>
</tr>
<tr>
<td>Seedling Stage</td>
<td></td>
</tr>
<tr>
<td>Primary root</td>
<td>41</td>
</tr>
<tr>
<td>Hypocotyl</td>
<td>NA</td>
</tr>
<tr>
<td>Cotyledon</td>
<td>45</td>
</tr>
<tr>
<td>Vegetative Stage</td>
<td></td>
</tr>
<tr>
<td>1&lt;sup&gt;st&lt;/sup&gt; leaf</td>
<td>NA</td>
</tr>
<tr>
<td>3&lt;sup&gt;rd&lt;/sup&gt; leaf</td>
<td>NA</td>
</tr>
<tr>
<td>5&lt;sup&gt;th&lt;/sup&gt; leaf</td>
<td>NA</td>
</tr>
<tr>
<td>Petiole</td>
<td>50</td>
</tr>
<tr>
<td>Apical shoot</td>
<td>NA</td>
</tr>
<tr>
<td>Leaf midrib</td>
<td>NA</td>
</tr>
<tr>
<td>Leaf lamina</td>
<td>40</td>
</tr>
<tr>
<td>Floral Stage</td>
<td></td>
</tr>
<tr>
<td>Pedicel</td>
<td>NA</td>
</tr>
<tr>
<td>Bract</td>
<td>NA</td>
</tr>
<tr>
<td>Calyx</td>
<td>NA</td>
</tr>
<tr>
<td>Petal</td>
<td>NA</td>
</tr>
<tr>
<td>Anther</td>
<td>NA</td>
</tr>
<tr>
<td>Stamina tube</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Pollen</td>
</tr>
<tr>
<td>--------------</td>
<td>--------</td>
</tr>
<tr>
<td>Style and stigma</td>
<td>NA</td>
</tr>
<tr>
<td>Ovary wall (0 dpa)</td>
<td>30</td>
</tr>
<tr>
<td>Ovule (0 dpa)</td>
<td>37</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fiber</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiber (5 dpa)</td>
<td>25</td>
<td>NA</td>
<td>21</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Fiber (10 dpa)</td>
<td>32</td>
<td>NA</td>
<td>26</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>Fiber (20 dpa)</td>
<td>37</td>
<td>NA</td>
<td>30</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Fiber (25 dpa)</td>
<td>41</td>
<td>NA</td>
<td>34</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>378</strong></td>
<td><strong>493</strong></td>
<td><strong>646</strong></td>
<td><strong>660</strong></td>
<td></td>
</tr>
</tbody>
</table>

NA = tissue not available
Table 3. Distribution of expression states among the mix, F1, synthetic, and Maxxa and their biological interpretation. The first nine rows compare the distribution of expression categories for all available gene × tissue combinations among all four taxa. The next four rows compare just F1, synthetic, and Maxxa, and the last two only synthetic and Maxxa.

<table>
<thead>
<tr>
<th>Genotype comparison</th>
<th>gene × tissue combinations showing pattern (% of total)</th>
<th>Biological description</th>
</tr>
</thead>
<tbody>
<tr>
<td>mix = F1 = synthetic = Maxxa with equal A-D expression</td>
<td>6 (6.4)</td>
<td>no change</td>
</tr>
<tr>
<td>mix = F1 = synthetic = Maxxa with unequal A-D expression</td>
<td>11 (11.7)</td>
<td>change due to A-D divergence</td>
</tr>
<tr>
<td>mix ≠ F1 = synthetic = Maxxa</td>
<td>23 (24.4)</td>
<td>change due to genome merger</td>
</tr>
<tr>
<td>mix = F1 ≠ synthetic = Maxxa</td>
<td>5 (5.3)</td>
<td>change due to polyploidy alone</td>
</tr>
<tr>
<td>mix = F1 = synthetic ≠ Maxxa</td>
<td>9 (9.6)</td>
<td>change due to polyploid evolution</td>
</tr>
<tr>
<td>mix ≠ F1 ≠ synthetic ≠ Maxxa</td>
<td>16 (17)</td>
<td>change due to all sources</td>
</tr>
<tr>
<td>mix ≠ F1 ≠ synthetic ≠ Maxxa</td>
<td>6 (6.4)</td>
<td>change due to polyploidy and polyploid evolution</td>
</tr>
<tr>
<td>Total</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>Count (Percentage)</td>
<td>Change Type</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>--------------------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>$F_1 = \text{synthetic} = \text{Maxxa}$</td>
<td>104 (34.9)</td>
<td>no change</td>
</tr>
<tr>
<td>$F_1 \neq \text{synthetic} = \text{Maxxa}$</td>
<td>57 (19.1)</td>
<td>change due to polyploidy alone</td>
</tr>
<tr>
<td>$F_1 = \text{synthetic} \neq \text{Maxxa}$</td>
<td>67 (22.5)</td>
<td>change due to polyploid evolution</td>
</tr>
<tr>
<td>$F_1 \neq \text{synthetic} \neq \text{Maxxa}$</td>
<td>70 (23.4)</td>
<td>change due to all sources</td>
</tr>
<tr>
<td>Total</td>
<td>298</td>
<td></td>
</tr>
<tr>
<td>synthetic $= \text{Maxxa}$</td>
<td>275 (50.6)</td>
<td>no change</td>
</tr>
<tr>
<td>synthetic $\neq \text{Maxxa}$</td>
<td>269 (49.4)</td>
<td>change due to polyploid evolution</td>
</tr>
<tr>
<td>Total</td>
<td>544</td>
<td></td>
</tr>
</tbody>
</table>
Table 4. Proportional transcript contribution of A and D homoeologs in different tissues.

Each gene (listed by GenBank accession) demonstrates nearly complete expression subfunctionalization among the two tissues shown.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>GenBank accession</th>
<th>Strongly A-biased tissue</th>
<th>A-biased tissue expression (A, D)</th>
<th>Strongly D-biased tissue</th>
<th>D-biased tissue expression (A, D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maxxa</td>
<td>CO131164</td>
<td>Anther</td>
<td>93, 7</td>
<td>Ovary wall</td>
<td>0, 100</td>
</tr>
<tr>
<td></td>
<td>CO080701</td>
<td>Pollen</td>
<td>94, 6</td>
<td>10 dpa fiber</td>
<td>14, 86</td>
</tr>
<tr>
<td></td>
<td>CO077994</td>
<td>10 dpa fiber</td>
<td>77, 23</td>
<td>Ovary wall</td>
<td>0, 100</td>
</tr>
<tr>
<td></td>
<td>DW008528</td>
<td>Ovary wall</td>
<td>100, 0</td>
<td>Anther</td>
<td>25, 75</td>
</tr>
<tr>
<td></td>
<td>CO082621</td>
<td>Anther</td>
<td>71, 29</td>
<td>Ovary wall</td>
<td>0, 100</td>
</tr>
<tr>
<td>Synthetic</td>
<td>CO124958</td>
<td>1st Leaf</td>
<td>100, 0</td>
<td>Style/Stigma</td>
<td>25, 75</td>
</tr>
<tr>
<td></td>
<td>CO098920</td>
<td>Cotyledon</td>
<td>91, 9</td>
<td>Anther</td>
<td>22, 78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F₁ hybrid</td>
<td>CO121715</td>
<td>Staminal tube</td>
<td>100, 0</td>
<td>Leaf lamina</td>
<td>25, 75</td>
</tr>
<tr>
<td></td>
<td>CAO23634</td>
<td>Leaf lamina</td>
<td>100, 0</td>
<td>Staminal tube</td>
<td>39, 61</td>
</tr>
</tbody>
</table>
Table 5. Distribution of tissue-specific homoeologous gene silencing events

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Number of gene × tissue assayed</th>
<th>Gene × tissue combinations with D silencing</th>
<th>Gene × tissue combinations with A silencing</th>
<th>% D silenced</th>
<th>% A silenced</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>493</td>
<td>7 (4 genes)</td>
<td>6 (3 genes)</td>
<td>1.41</td>
<td>1.21</td>
</tr>
<tr>
<td>Synthetic</td>
<td>646</td>
<td>16 (5 genes)</td>
<td>2 (2 genes)</td>
<td>2.48</td>
<td>0.31</td>
</tr>
<tr>
<td>Maxxa</td>
<td>660</td>
<td>42 (11 genes)</td>
<td>9 (5 genes)</td>
<td>6.36</td>
<td>1.36</td>
</tr>
</tbody>
</table>
Table 6. Expression neofunctionalization. Each gene exhibited differential expression between the diploids (shown by the 1:1 parental mix), but expression from both the A- and D-genomes in the F₁, synthetic and natural allopolyploid (Maxxa).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CAO62858</td>
<td>Root</td>
<td></td>
<td>0, 100</td>
<td>NA</td>
<td>32, 68</td>
<td>41, 59</td>
</tr>
<tr>
<td></td>
<td>Ovule</td>
<td></td>
<td>0, 100</td>
<td>30, 70</td>
<td>25, 75</td>
<td>32, 68</td>
</tr>
<tr>
<td>CO111212</td>
<td>Leaf petiole</td>
<td></td>
<td>0, 100</td>
<td>--</td>
<td>27, 73</td>
<td>33, 67</td>
</tr>
<tr>
<td></td>
<td>5 dpa Fiber</td>
<td></td>
<td>0, 100</td>
<td>NA</td>
<td>34, 66</td>
<td>49, 51</td>
</tr>
<tr>
<td></td>
<td>10 dpa Fiber</td>
<td></td>
<td>0, 100</td>
<td>NA</td>
<td>29, 71</td>
<td>42, 58</td>
</tr>
<tr>
<td>CO108066</td>
<td>Leaf lamina</td>
<td></td>
<td>0, 100</td>
<td>49, 51</td>
<td>66, 34</td>
<td>63, 37</td>
</tr>
<tr>
<td></td>
<td>Ovule</td>
<td></td>
<td>0, 100</td>
<td>59, 41</td>
<td>68, 32</td>
<td>34, 66</td>
</tr>
<tr>
<td>CO076921</td>
<td>Leaf lamina</td>
<td></td>
<td>100, 0</td>
<td>61, 39</td>
<td>93, 07</td>
<td>57, 43</td>
</tr>
<tr>
<td>AAP41846</td>
<td>Ovule</td>
<td></td>
<td>0, 100</td>
<td>53, 47</td>
<td>50, 50</td>
<td>48, 52</td>
</tr>
<tr>
<td>CO081422</td>
<td>Ovule</td>
<td></td>
<td>0, 100</td>
<td>48, 52</td>
<td>40, 60</td>
<td>--</td>
</tr>
<tr>
<td>CAO71171</td>
<td>Ovule</td>
<td></td>
<td>0, 100</td>
<td>61, 59</td>
<td>40, 60</td>
<td>47, 53</td>
</tr>
<tr>
<td>AAK69758</td>
<td>Ovule</td>
<td></td>
<td>0, 100</td>
<td>49, 51</td>
<td>22, 78</td>
<td>38, 62</td>
</tr>
<tr>
<td>CO077994</td>
<td>20 dpa Fiber</td>
<td></td>
<td>100, 0</td>
<td>NA</td>
<td>63, 37</td>
<td>58, 42</td>
</tr>
<tr>
<td>CO093729</td>
<td>25 dpa Fiber</td>
<td></td>
<td>100, 0</td>
<td>NA</td>
<td>72, 28</td>
<td>86, 14</td>
</tr>
</tbody>
</table>

NA= tissue not available

(--) = value could not be determined
Figure 1. Phylogenetic framework and detection of subfunctionalization and neofunctionalization among homoeologs. (A) Phylogenetic history of diploid and allopolyploid cotton (*Gossypium*). Allopolyploidy occurred ~1-2 MYA by hybridization between A- and D-genome diploid species, most similar to the modern species *G. arboreum* and *G. raimondii*. The modern F1 hybrid and synthetic allopolyploid, both derived from A- and D-genome diploid species, mimic the stages of genome merger and genome duplication during allopolyploid formation. (B) After genome merger regulatory changes may cause allelic/homoeologous gene expression patterns to diverge. These patterns can result in subfunctionalization, the partitioning of ancestral expression, or neofunctionalition, operationally defined here as the development of novel expression patterns relative to that of the ancestor. The latter was detected by comparing ancestral expression (1:1 mix) to the expression found in the F1, synthetic, and Maxxa.
A

![Diagram of diploid species divergence](image)

B

**detecting subfunctionalization**

<table>
<thead>
<tr>
<th>Tissue</th>
<th>A and D expression in Mix</th>
<th>A and D expression in Allopolyploid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>A subfunctionalization</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>A subfunctionalization</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>no evolutionary change</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>partial D subfunctionalization</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>complete D subfunctionalization</td>
</tr>
</tbody>
</table>

**detecting neofunctionalization**

<table>
<thead>
<tr>
<th>Tissue</th>
<th>A and D expression in Mix</th>
<th>A and D expression in Allopolyploid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>only A diploid expression</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>D-genome expression reactivated (or neofunctionalized) following allopolyploidy</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>only D diploid expression</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>A-genome expression reactivated (or neofunctionalized) following allopolyploidy</td>
</tr>
</tbody>
</table>
Figure 2. Distribution of genome-specific expression states among accessions and within different tissue categories. Each panel represents a tissue category and shows histograms for the mix F₁, synthetic, and Maxxa. The expression categories correspond to the following values: Strongly A biased (0-20% D expression); A biased (20-40% D expression); Equivalent (40-60% D expression); D biased (60-80% D expression); Strongly D biased (80-100% D expression). The y-axes indicate the number of gene by tissue combinations that fell under each category. (NA = tissue type not available)
Figure 3. Tissue-specific and genome-specific gene expression among three gene pairs in A₂, D₅, F₁ hybrid, synthetic and Maxxa. Tissues are arrayed along the x-axis while the proportion of D-genome expression is on the y-axis. The lines linking tissues do not imply a strict order of plant development; instead they serve as a viewing aid. (A) gene CO131164 (a putative phytochrome-associated protein), (B) gene CO130747 (a putative CBL-interacting protein kinase), and (C) gene DW008528 (a protein of unknown function).
Figure 4. Examples of tissue-specific expression alteration arising from parental divergence, genomic merger, polyploidy, and polyploidy evolution. Shown are the proportions of D-genome (y-axis) homoeolog expression, including the associated standard deviation. (A) Four representative genes from petioles, which exhibit statistically equivalent ratios in all accessions, indicating little expression evolution since divergence between the A-and D-genome parents. (B) Four representative genes from leaf lamina, each showing equivalent expression among the F₁ hybrid, synthetic and Maxxa, which is not equivalent to the mix, indicating an expression change resulting from genomic merger. (C) Four representative genes from petioles (gene CAO71171, CO131379 and CAO49511) and leaf lamina (gene CO131379) showing equal genome-specific expression values in the mix and F₁ hybrid, which differ from the synthetic and Maxxa, suggesting that the change occurred as a result of genome doubling. (D) Four genes in petioles showing no change among mix, F₁ and synthetic, but a new expression pattern in Maxxa, indicating a change in homoeolog-specific expression during the evolution (~1-2 million years) of allopolyploid cotton.
Figure 5. Examples of tissue-specific expression partitioning. The y-axis represents the proportional transcript contribution from A- and D-homoeologs.
Figure 6. Plots of A- and D-genome parental mix (mix) vs. F₁ hybrid (F₁) for leaf lamina (A) and petiole (B). In principle, genome-specific expression differences initiated by cis-regulatory divergence are expected to share this difference between both the mix and the F₁, and will accordingly fall on a 1:1 diagonal when plotted against one another (red points), whereas trans-regulatory divergence will equilibrate genome-specific expression when co-resident in the F₁ nucleus and instead fall on equivalently expressed horizontal line for the F₁ only (blue points). Genes that fall along neither of these lines are inferred to be regulated by a combination of cis- and trans-factors (Wittkopp et al., 2004; Stupar & Springer, 2006) (green points). Finally, genes with divergence only in the F₁ (purple points) or no expression divergence (grey points) offer no insight into cis or trans expression evolution.
Supplemental Materials

**Supplemental Figure 1.** Number of informative homoeologous gene-specific assays for each tissue. The tissues were selected for the present study from four different stages of cotton development representing seedling, vegetative, floral and fiber development and from F1 hybrid, Synthetic allopolyploid and natural allopolyploid. A total of 63 genes were assayed in different tissues by genotype combination. The y-axis represents the number of successful assays for any particular tissue.
Supplemental Figure 2. Expression values of Mix samples in 8 tissues with relative contribution of A-and D-homoeologs. The x-axis represents the number of homoeologous genes in which the Mix values could be detected in each tissue, and y-axis represents the relative expression values from D-homoeolog. The expression values in Mix vary between 0 to 1 confirming the precision of physical mixture of RNAs from two parental species.
Supplemental Figure 3. Relative homoeolog-specific gene expression for each of all 9 neo-functionalized genes in other tissues of F1, Synthetic and Maxxa. The y-axis represents the relative expression of D-homoeolog. Each panel is shown with three genes representing (left to right) (A) gene CAO62858, CO111212 and CO108066 (B) gene CO076921, AAP41846 and CAO71171 (C) gene AAK69758, CO077994 and CO093729.
Supplemental Figure 4. Comparison of expression divergence between *G. davidsonii* (D3) and *G. raimondii* (D5) relative to a common *G. arboreum* (A2) reference sample. On the plot both leaf (green circles) and petiole (red circles) are represented, as well as the best fit line (solid) and a hypothetical line representing a perfect 1 to 1 fit (dashed line) between these species.
<table>
<thead>
<tr>
<th>Contig</th>
<th>SNP_position</th>
<th>Reference_GeneBank_Accession</th>
<th>Gene_description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cotton16_00001_034</td>
<td>832</td>
<td>CO112436</td>
<td>putative small GTP.binding protein</td>
</tr>
<tr>
<td>Cotton16_00001_056</td>
<td>1492</td>
<td>CO130933</td>
<td>clathrin adaptor medium chain protein MU1B, putative</td>
</tr>
<tr>
<td>Cotton16_00001_062</td>
<td>1928</td>
<td>CO124017</td>
<td>vacuolar proton.ATPase subunit.like protein</td>
</tr>
<tr>
<td>Cotton16_00001_192</td>
<td>2122</td>
<td>CO071793</td>
<td>acetyl.CoA carboxylase</td>
</tr>
<tr>
<td>Cotton16_00001_452</td>
<td>1635</td>
<td>CO131164</td>
<td>phytochrome.associated protein 1</td>
</tr>
<tr>
<td>Cotton16_00004_01</td>
<td>779</td>
<td>CO108066</td>
<td>glyceraldehyde.3.phosphate dehydrogenase</td>
</tr>
<tr>
<td>Cotton16_00001_12</td>
<td>1315</td>
<td>CAO71171</td>
<td>hydroxyproline.rich glycoprotein.like protein</td>
</tr>
<tr>
<td>Cotton16_00013_06</td>
<td>1353</td>
<td>CO094037</td>
<td>putative cystathionine gamma.synthase</td>
</tr>
<tr>
<td>Cotton16_00024_03</td>
<td>2070</td>
<td>CO105110</td>
<td>calmodulin.like domain protein kinase</td>
</tr>
<tr>
<td>Cotton16_00025_07</td>
<td>221</td>
<td>CO081422</td>
<td>Os05g0455600 [Oryza sativa ]</td>
</tr>
<tr>
<td>Cotton16_00056_02</td>
<td>720</td>
<td>AAP41846</td>
<td>cysteine protease</td>
</tr>
<tr>
<td>Cotton16_00069_04</td>
<td>828</td>
<td>CO076921</td>
<td>VATA_GOSHI Vacuolar ATP synthase catalytic subunit A</td>
</tr>
<tr>
<td>Cotton16_00071_01</td>
<td>540</td>
<td>CO121715</td>
<td>AUX/IAA protein</td>
</tr>
<tr>
<td>Cotton16_00075_03</td>
<td>432</td>
<td>CO117833</td>
<td>aldehyde dehydrogenase family 7 member A1</td>
</tr>
<tr>
<td>Cotton16_00076_06</td>
<td>860</td>
<td>CO102987</td>
<td>ethylene transcription factor</td>
</tr>
<tr>
<td>Cotton16_00156_04</td>
<td>492</td>
<td>CO105072</td>
<td>putative beta.1.3.glucanase</td>
</tr>
<tr>
<td>Cotton16_00173_04</td>
<td>1494</td>
<td>CO122994</td>
<td>TBA3_ELEIN Tubulin alpha.3 chain</td>
</tr>
<tr>
<td>Cotton16_00174_02</td>
<td>802</td>
<td>CO111918</td>
<td>unknown protein</td>
</tr>
<tr>
<td>Cotton16_00197_01</td>
<td>52</td>
<td>CO090037</td>
<td>TPIC_SPIOL Triosephosphate isomerase</td>
</tr>
<tr>
<td>Cotton16_00285_02</td>
<td>685</td>
<td>CO131379</td>
<td>serine acetyltransferase 7</td>
</tr>
<tr>
<td>Cotton16_00373_02</td>
<td>1079</td>
<td>CO125131</td>
<td>ADT1_GOSHI ADP,ATP carrier protein 1</td>
</tr>
<tr>
<td>Cotton16_00479_01</td>
<td>592</td>
<td>CO093729</td>
<td>ATP binding</td>
</tr>
<tr>
<td>Cotton16_00690_02</td>
<td>916</td>
<td>CO118820</td>
<td>Beta.COP.like protein</td>
</tr>
<tr>
<td>Cotton16_00727_02</td>
<td>140</td>
<td>CA062858</td>
<td>nucleoporin family protein</td>
</tr>
<tr>
<td>Cotton16_00922_01</td>
<td>949</td>
<td>CO124958</td>
<td>phosphoinositide.specific phospholipase C P13</td>
</tr>
<tr>
<td>Cotton16_01040_02</td>
<td>1017</td>
<td>CO111355</td>
<td>TBB5_GOSHI Tubulin beta.5 chain (Beta.5 tubulin)</td>
</tr>
<tr>
<td>Accession</td>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_01121_02</td>
<td>632 CO129884 2.phosphoglycerate kinase.related</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_01189_01</td>
<td>1178 CO130747 CBL.interacting protein kinase 16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_01218_01</td>
<td>872 CO096546 LEJ2 (LOSS OF THE TIMING OF ET AND JA BIOSYNTHESIS 2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_01391_01</td>
<td>705 DT461656 phosphorybosyl anthranilate transferase 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_01436_02</td>
<td>169 CO102224 t.complex polypeptide 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_01499_01</td>
<td>472 CO110993 phosphate.responsive 1 family protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_01704_01</td>
<td>1415 DW008528 putative protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_01766_02</td>
<td>410 CO106047 oligosaccharide transporter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_01818_02</td>
<td>625 CO085186 tetratricopeptide repeat (TPR).containing protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_02074_02</td>
<td>678 CO081139 transcription factor Hap5a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_02786_01</td>
<td>891 CO084845 1,4.alpha.glucan branching enzyme</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_03680_01</td>
<td>309 CO085733 universal stress protein (USP) family protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_04501_01</td>
<td>681 CAO49511 EMB1417 (EMBryo DEFECTIVE 1417)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_06427_01</td>
<td>504 CO080453 unknown protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_07872_01</td>
<td>1017 CO118336 TBA4_GOSHI Tubulin alpha.4 chain (Alpha.4 tubulin)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_09095_01</td>
<td>1294 CO111212 Ubiquitin.associated protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_09331_01</td>
<td>537 CO102525 leucine.rich repeat family protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_14442_01</td>
<td>395 CAO23634 S.formylglutathione hydrolase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_15666_01</td>
<td>268 CO077994 putative c.myc binding protein MM.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_17428_01</td>
<td>832 ABA95925 Amidase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_19029_01</td>
<td>267 CO075025 acyl carrier protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_19620_01</td>
<td>376 CO080701 Tubby; Di.trans.poly.cis.decaprenylcistransferase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_19657_01</td>
<td>224 NP_196352 4SNC.Tudor domain protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_21601_01</td>
<td>747 CO101293 unnamed protein product</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_21697_01</td>
<td>393 CO080172 putative lateral suppressor region D protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_22170_01</td>
<td>513 CO098920 GTP.binding protein GB2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_24663_01</td>
<td>633 CO073158 protein kinase family protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_25466_01</td>
<td>1125 CO129204 Calmodulin.binding transcription activator 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_26306_01</td>
<td>942 CO089285 putative secretory carrier.associated membrane protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton16_27501_01</td>
<td>1173 CO122743 predicted proline.rich protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene Name</td>
<td>Accession</td>
<td>Function</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------</td>
<td>-------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>CO087191</td>
<td>1145</td>
<td>6-phosphogluconate dehydrogenase</td>
<td></td>
</tr>
<tr>
<td>AAK69758</td>
<td>940</td>
<td>Unknown protein</td>
<td></td>
</tr>
<tr>
<td>CO082621</td>
<td>1149</td>
<td>Putative protein kinase</td>
<td></td>
</tr>
<tr>
<td>CO072998</td>
<td>334</td>
<td>Calcineurin B-like protein</td>
<td></td>
</tr>
<tr>
<td>CO131697</td>
<td>812</td>
<td>Mutant cincinnata</td>
<td></td>
</tr>
<tr>
<td>CO115511</td>
<td>1103</td>
<td>Ethylene signaling protein</td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 5.

COORDINATED AND FINE-SCALE CONTROL OF HOMOELOGOUS GENE EXPRESSION IN ALLOTETRAPLOID COTTON


Lex E. Flagel, Liping Chen, Bhupendra Chaudhary, Jonathan F. Wendel

ABSTRACT

Within polyploid plant species it has been demonstrated that homoeologous genes (genes duplicated by polyploidy) often display dynamic expression patterns. To determine if chromosomal location plays a role in establishing these expression patterns we analyzed the relative levels of homoeolog expression among linked genes from two locations in the cotton genome. Genes from the region containing the AdhA gene show coordinated expression across several tissues, whereas genes from the region containing CesA do not. These results indicate that changes in homoeolog expression may be constrained by linkage in some genomic regions, while in other regions homoeolog expression is largely decoupled from physical proximity. Furthermore these results suggest that both large- and small-scale regulatory mechanisms may control homoeolog expression patterns.
INTRODUCTION

Polyploidy is a common and phylogenetically widespread phenomenon in angiosperms. Many polyploids are chromosomal allopolyploids, typically descendants of a polyploidy event that included interspecific hybridization (Wendel & Doyle, 2005). As such, allopolyploids maintain two or more divergent genomes within a common nucleus. This sharing of a common nucleus creates the potential for a variety of intergenomic interactions, including homoeologous recombination (Udall et al., 2005; Gaeta et al., 2007), genomic deletions (Ozkan et al., 2001; Kashkush et al., 2002), and modifications of homoeologous gene expression (Adams et al., 2003; Mochida et al., 2003; Bottley et al., 2006; Flagel et al., 2008; Hovav et al., 2008). These findings have led to speculation that interactions among co-resident genomes play a key role in the development of novel phenotypes in allopolyploid species (Wendel, 2000; Osborn et al., 2003; Chen, 2007).

One of the more impressive discoveries regarding allopolyploids concerns the scale and scope of unequal contributions to the transcriptome made by homoeologous genes. Adams et al. (2003) showed that among 40 homoeologous gene pairs in cotton, 10 exhibit biased expression patterns, including developmentally regulated expression ratio variation among homoeologs and reciprocal silencing of alternative homoeologs in different floral whorls. Similar findings were reported in wheat (Mochida et al., 2003; Bottley et al., 2006). More recently, Flagel et al. (2008) and Hovav et al. (2008) used microarray technology to show that homoeolog expression biases are common throughout the transcriptome and can vary temporally and developmentally, even within the single-celled trichomes of Gossypium (cotton “fibers”).
Despite these advances into understanding the phenomenon of homoeolog expression variation, the responsible molecular mechanisms remain largely unexplored and unknown. Although both cis and trans factors are implicated (Chaudhary et al., 2008), little is understood regarding the roles of genetic and epigenetic factors in creating homoeolog expression biases, nor how these expression changes affect adjacent genes. With respect to the latter, preliminary evidence from allotetraploid Arabidopsis indicates that homoeolog expression may be quite variable, even among closely spaced genes on a single BAC (Lee & Chen, 2001). To explore this suggestion further and in a different system, we quantitatively assayed homoeolog expression ratios for genes found on two well-characterized regions of the allotetraploid cotton (G. hirsutum) genome. By collecting homoeolog expression data from a wide range of tissues we sought to determine the effect of physical proximity on homoeolog expression. If closely linked genes show correlated homoeolog expression biases, one might infer that the biases likely result from the action of a single cis-acting factor such as a chromatin modification or a shared enhancer affecting a large genomic region. Conversely, if linked genes show uncorrelated expression patterns, one would conclude that multiple fine-scale factors may be at work, including localized divergence among homoeologous cis-regulatory regions or epigenetic marks.

MATERIALS AND METHODS

Gossypium has emerged as an excellent system for studying patterns of homoeologous gene expression. This genus contains five extant allotetraploid species that were formed during a single polyploidization event approximately 1-2 million years ago
(Senchina et al., 2003). Allotetraploids (‘AD genome’) combine diploid genomes from the ‘A’ and ‘D’ species groups, are chromosomally diploidized (i.e., form bivalents at meiosis), and have a genome size that is approximately additive with respect to those of its diploid progenitors (Wendel & Cronn, 2003). Beyond their well-understood origin and phylogeny, there also are ample genetic and genomic resources for cotton. Of importance to this study are annotated BAC sequences from *G. hirsutum* for both homoeologous genomes around the regions containing the *Alcohol dehydrogenase A* (*AdhA*; GenBank accessions EF457753 and EF457754) and the *Cellulose synthase A* genes (*CesA*; GenBank accessions AY632359 and AY632360) (Grover et al., 2004; Grover et al., 2007). By aligning genes on these BACs, we were able to identify SNP markers specific to the homoeologous A- and D-genomes for 3 genes from the *AdhA* BACs and 8 genes from the *CesA* BACs (Fig. 1A). These genome-specific SNPs were used as targets for the MassARRAY (Sequenom) mass-spectrometry platform, which can quantitatively assay relative homoeolog transcript abundance as described below. Our goal was to ascertain the relative ratios of homoeolog expression in a diverse panel of tissues, and in particular to evaluate the degree to which gene expression among homoeologs is correlated with genomic proximity.

Following protocols described elsewhere (Chaudhary et al., 2008), we extracted RNAs and prepared cDNAs libraries from 18 tissue types (Fig. 1B). These cDNA libraries were then subjected to MassARRAY mass-spectrometry, which utilizes differences in mass between genome-specific SNP variants to quantify the relative abundance of homoeologous transcripts. The resulting homoeolog expression ratio data were then filtered based on internal MassARRAY technical controls to remove assays flagged as “Bad Spectra”, or
having a frequency of uncertainty \(> 0.2\) or an unused extension primer frequency \(> 0.5\). Next, assays were required to respond correctly to range of genomic DNA mixtures \((4:1, 2:1, 1:1, 1:2, \text{ and } 1:4)\) from the A- and D-genome diploids, \textit{G. arboreum} and \textit{G. raimondii}. Assays which did not show a strong correlation \((R^2 > 0.9)\) with their expected ratios were removed. Additionally, genomic DNAs from \textit{G. hirsutum}, the allotetraploid species used in this study, were also assayed as a control for allotetraploid lineage-specific SNPs. In the absence of a lineage-specific SNP the allotetraploid genomic DNA should yield approximately 1:1 A- to D-genome values. Any assays beyond this expected 1:1 ratio (± 25%) were individually excluded. Finally, only those assays with 3 or more replicates (from a maximum of 9 replicates) are summarized in Fig. 1B.

**RESULTS AND DISCUSSION**

Several interesting patterns emerge from the homoeolog expression data. First, for nearly all tissues (except anther, leaf lamina, and calyx), genes along the \textit{AdhA} BAC show a consistent V-shaped pattern, in which \textit{AdhA} and \textit{PDI} consistently maintain a bias toward greater D-genome transcription than does \textit{FAD-Ox}. The maintenance of this pattern is striking and consistent, despite the appreciable differences in homoeolog ratios among tissues (e.g. root vs. 10 dpa fiber). These findings from the \textit{AdhA} region indicate that homoeologous gene expression can be coordinated among adjacent genes. This coordination involves the maintenance of a specific pattern of homoeolog expression bias across many tissues, suggesting the operation of a mechanism that is developmentally persistent and widespread with regard to tissue type.
To our knowledge this is the first time that large-scale (≈ 40 kb) coordinated control of expression for linked homoelogous genes across several tissues has been demonstrated. Given the breadth of tissues examined, it is reasonable to speculate that this pattern is caused by a developmentally stable, coordinated chromatin modification that exercises control over homoelogous gene expression in most tissues. This regional effect likely dominates specific gene effects in the AdhA region.

In contrast to coordinated gene expression in the AdhA region, a different story emerges for genes along the CesA BAC. These genes show no clear pattern of coordinated expression (Fig 1B). Instead, genes in this genomic location suggest fine-scale control of homoelog expression, including, in several cases, apparently complete gene silencing. For example, in the staminal tube the D-genome contributes nearly all of the transcripts of Hypo1, whereas its closest neighbor, G Prot B, is predominantly expressed by the A-genome. In some cases genes only a few kilobases apart show very different expression states. For example, in the leaf midrib LeuRR2 expression is approximately equivalently for both genomes, but Hypo2, which is less than 4 kilobases away, is represented almost entirely by the A-genome homoelog. Furthermore, we performed statistical analyses on the CesA region to determine if physical distance, gene order, and gene orientation had an effect on pairwise correlations of homoelogous gene expression across tissues. The significance of these physical factors were assessed using permutation tests with physical distance measured in base pairs and tested for both the A- and D-genome, gene order measured as the number of intervening genes, and gene orientation encoded as either parallel (→ → or ← ←) or opposing (← → or → ←). These distance data sets were subjected to 1000 random
permutations and compared to the actual correlation. All physical factors were found to be non-significantly correlated with expression for the CesA region (physical distance: A-genome (Pearson’s $r = 0.17$, $P = 0.73$), D-genome ($r = 0.15$, $P = 0.74$); gene order ($r = 0.14$, $P = 0.74$); and gene orientation ($r = -0.27$, $P = 0.13$)). Thus, within this region our results reveal the presence of localized and intricate regulatory control among homoeologous genomes. This suggests a reduced role for large-scale factors such as heterochromatin related effects, instead suggesting locally operating factors such as cis-regulatory differences or alteration in localized epigenetic states.

The foregoing explorations of homoeologous gene expression in a diverse panel of tissues reveal two distinct patterns, enhancing our understanding of the phenomenon of homoeolog expression bias and providing clues into its genesis. Results from the AdhA region demonstrate that genomic locale may exercise a profound effect on homoeolog expression in a persistent and developmentally widespread manner. This finding is supported by genomic expression analyses in diploids, which often show that angiosperm genomes contain many short tracts of genes (generally less than 10) with correlated expression patterns (Williams & Bowles, 2004; Ren et al., 2005; Zhan et al., 2006; Ren et al., 2007; Quesada et al., 2008). Though this phenomenon has not been demonstrated in Gossypium, it is possible that expression in the AdhA region for both the A- and D-genomes falls into one of these co-expressed tracts. Our data add to this body of research by showing evidence of apparent coordination between homoeologous genomes. In contrast, the CesA region demonstrates that closely linked genes can deviate significantly in their homoeologous contribution to the transcriptome. These results imply that widespread patterns of
homoeolog expression biases in allotetraploid cotton and wheat (Adams et al., 2003; Mochida et al., 2003; Bottley et al., 2006; Flagel et al., 2008; Hovav et al., 2008) are likely the product of both fine-scale local regulation as well as more far-reaching chromosomal factors.

REFERENCES


Udall JA, Quijada PA, Osborn TC. 2005. Detection of chromosomal rearrangements derived from homeologous recombination in four mapping populations of *Brassica napus L.* *Genetics* 169: 967-979.


Zhan S, Horrocks J, Lukens LN. 2006. Islands of co-expressed neighbouring genes in

Arabidopsis thaliana suggest higher-order chromosome domains. Plant J. 45: 347-357.
Figure 1 – Representations of AdhA and CesA genomic regions from G. hirsutum and summary of relative homoeolog contribution to the transcriptome for neighboring genes in multiple tissues. A) Arrows represent genes for which expression data were generated from the AdhA and CesA BACs. The direction of the arrow indicates the direction of transcription for each gene, and the number of bases between genes is given for both the A- and D-genomes of the allopolyploid. The order and direction of transcription is conserved between the A- and D-genomes for all genes used in this study. Asterisks above the CesA BAC mark intervals that contain a predicted gene which could not be assayed due to lack of genome-specific SNPs. B) Mean expression results for the AdhA and CesA BACs for each tissue. For each panel, genes are arrayed in physical order (x-axis), and homoeolog expression ratio is shown on the y-axis (as proportion of the transcriptome contributed by the D-genome). Some low-quality assays were excluded from the CesA data. Abbreviations: AdhA: Alcohol dehydrogenase A; CesA: Cellulose synthase A; dpa: days post anthesis; FAD-Ox: FAD-dependent oxidoreductase; G Prot B: G-protein beta; Hypo: expressed hypothetical protein; LeuRR: Leucine-rich repeat protein; PDI: Protein disulfide isomerase; Perm: Permease; PPR: Pentatricopeptide repeat protein.
CHAPTER 6.
GENERAL CONCLUSIONS

In this conclusion I will address each of the six primary research objectives outlined in the introduction of this dissertation, in light of our experimental findings, as detailed in Chapters 2-5.

1. What is the temporal pace and scope of gene expression evolution following genome merger and duplication?

Chapters 2-4 address this point. It is clear from these Chapters, that as a general rule, expression biases accumulate over evolutionary time. Also, among the five natural allotetraploids there is a trend to this accumulation, in that it tends to enhance transgressive up- and down-regulation while at the same time diminish the overwhelming D-genome expression dominance that can be found in synthetic diploid and allotetraploid cottons (Rapp et. al (2009) and Chapter 3). This said, there are also significant differences among the five natural allotetraploid species. *Gossypium tomentosum* clearly shows extreme homoeologous expression biases (Chapter 3), and *G. barbadense* and *G. hirsutum* are not far behind. However, when we look to the global patterns of expression for the 42000 genes for which we have microarray probes, we can see that all five species are more similar (Chapter 3), indicating that the patterns found within the homoeologous expression biases are either 1) particular to only the 1400 genes we can measure, or 2) do not transmit overall changes in total transcript accumulation, but instead just alter the homoeologous make-up of this
transcript pool. Future research, possibly using next-generation sequencing as a tool, may be able to better pin-point the cause of this discrepancy between homoeologous and total expression among the *Gossypium* allopolyploids.

2. What is the extent of homoeologous expression evolution?

The findings from Chapters 2 and 3, which both assay about 1400 homoeologous pairs, clearly show that homoeologous expression frequently departs from the parental expression regimes. If we consider these 1400 homoeologs to be a fair cross-section of the cotton genome, which may contain 45000+ homoeologs, we can estimate, that for petal tissues, homoeologous expression evolution (as in non-additive expression when compared to the 1:1 parental mix) has taken place in approximately 59-89% of all genes (or tens of thousands of genes, when extrapolated to the genome as a whole). This is of course, only in one tissue, thus the extent of homoeologous expression changes could be truly striking, when we account for all tissues, cell types, and developmental and environmental conditions. We look at multiple tissues and developmental stages in Chapter 4, and indeed find that the evolution of homoeologous expression in these tissues is consistent with our findings from petals, and that homoeologous expression evolution occurs in all tissues thus far analyzed, even single celled seed trichomes (cotton fibers).

With regard to this question, it is important to keep in mind that our diagnosis of expression evolution is statistically defined, rather than biologically defined. This is out of necessity, as we do not yet have the capacity to study the evolutionary consequences of
expression changes for dozens of genes in cotton, let alone the thousands of genes that make up the genome. Important future work will be determining the extent to which expression changes result in phenotypic evolution. I suspect, given the extent of homoeologous expression evolution, it will be determined that numerous differences between the allotetraploid *Gossypium* species are attributable to expression alterations following polyploidy.

3. What amount of expression evolution can be attributed to hybridization, and what amount can be attributed to genome duplication and the diversification and evolution of polyploid species?

We address this question most directly in Chapters 2 and 4. In Chapter 2 we find that approximately one-quarter of the expression changes found in the natural allopolyploid could be explained purely as a consequence of genomic merger, because they are shared between the allopolyploid and a synthetic F₁ hybrid. In Chapter 4 we elaborate on this point, showing that the contributions of genomic divergence, genome merger, genome duplication, and long-term polyploidy evolution can be divided up as follows among 63 homoeologs assayed in 25 different tissues and developmental time-points (derived from Chapter 4, Table 3):

1. genome merger (40%) (changes found in synthetic F₁ hybrid)
2. polyploid evolution (28%) (changes found in natural allotetraploid *G. hirsutum*)
3. genome doubling (21%) (changes found in synthetic allotetraploid)
4. diploid divergence (7%) (changes between A- and D- diploids)

5. no change (4%)

This sets a general ranking and shows the importance of each of these phenomena associated with allopolyploidization. It will be interesting to see if these same general rankings hold true in other allopolyploid species, which may have different levels of genomic divergence and shorter or longer periods since polyploid formation.

4. How do these patterns play out among different tissue types or among different developmental time-points with the same tissue?

This question is addressed by Chapter 4. Among the different tissues assayed, we see that there are consistent differences between those grouped as seedling, vegetative, floral, and fiber tissues (Chapter 4, Fig. 2). These differences are not extremely strong, all tissue types appear to have roughly equivalent A- and D-genome homoeolog expression patterns. We do note, however, that reproductive tissues (ovule, stamen, ovary wall) seem to show the most cases of expression sub- and neo-functionalization (Chapter 4, Tables 4 & 6)

5. How does genomic linkage (i.e. proximity among genes) influence expression evolution following polyploidy?

Chapter 5 shows that in one region of the genome (near the CesA gene), homoeologous gene expression is unconstrained by linkage in an large number of tissues, meaning that each of these genes could presumably evolve new expression patterns
independently. On the other hand, genes in the AdhA region of the genome show a distinctive pattern of expression in numerous tissues, and are likely to be co-expressed to an extent, meaning that expression evolution among these genes may not be independent. These data, are, of course, quite small, and reflect our current knowledge of the cotton genome. As continued genomic resources develop in cotton, it will be interesting to determine more specifically the proportion of genes that are constrained by linkage versus those that are largely independent.

6. Finally, what does all of this tell us about the importance of expression evolution following gene duplication?

The accumulated results of this dissertation suggest that dynamic expression changes may occur immediately upon allopolyploid formation. Because of the genomic scale and potential phenotypic effects of gene expression change following polyploidy, it is likely that expression alteration following polyploidy will prove to be a significant source of evolutionary novelty among plants. We show that this expression change likely due in part to functional redundancy removing purifying selection from one copy (long-term evolutionary processes), and in part due to the “genomic shock” caused by genomic merger and duplication (immediate saltational evolutionary processes). Furthermore, we show evidence that this expression evolution is tempered by a host of factors, including tissue type, linkage, and cis regulatory divergence between genomes. Furthermore, and possibly most crucially, changes in gene expression that result in tissue-specific subfunctionalization (which we show evidence for in Chapter 4) can have a preserving effect on duplicate genes,
by requiring that each be maintained. This factor could be an important overlooked pathway to novel evolutionary changes within plants.
LITERATURE CITED


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


ACKNOWLEDGEMENTS

I first thank my parents Kevin and Janet Flagel, who are both ardent supporters of higher education and have both encouraged each step of my progression from undergraduate to graduate, and on to post-doctoral pursuits. In addition to my parents, I also thank my brother, friends, and especially my fiancée, Suzanne McGaugh, for filling my life with happiness, ultimately making my academic pursuits possible.

I must also thank my academic advisor and mentor Jonathan Wendel. Jonathan has made tremendous contributions to the polyploid research community and has a sterling track record when it comes to producing top-quality scholars. I am deeply indebted to him for taking me in, funding me, and allowing me academic freedom while still supplying a safety-net. It would be hard for me to list all the skills and traits I’ve learned or absorbed from working with Jonathan, but the two that stand out most and that I will use forever are leadership and scientific scholarship and writing. I truly believe that Jonathan does a fabulous job of preparing his students in these two areas, and it is reflected in their success over the years.

I am also deeply indebted to the co-authors on each of these studies. In many cases, much of the work presented here could not have been accomplished without their help and expertise. I especially want to thank Josh Udall, whose pioneering work in cotton genomics laid the ground work for each of these studies. Were it not for Josh, my thesis certainly would not have been possible. Beyond my co-authors I also thank my lab mates, who provided support throughout this process, and the EEOB staff, who have handled countless things behind the scenes.
I also thank my committee members—Fred Janzen, Dennis Lavrov, John Nason, and Dan Nettleton—for their time, insight, and for lending their expertise. I hope their investment in my education has given them some satisfaction and insights, I can certainly say that the reverse is true for me. I must also thank Dean Adams, for numerous interesting and informative discussions about statistics.

I thank Charlie Brummer, who hosted me on my first rotation at Iowa State, and who introduced me to polyploidy in alfalfa, which in turn lead me to the Wendel lab. Though I didn’t end up working with Charlie, I am indebted to him for helping me find my academic interest and for selflessly promoting the Wendel lab.

Finally, this work would not have occurred without financial support from the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service (2005-35301-15700 to Josh Udall and Jonathan Wendel) and the National Science Foundation Plant Genome Research Program (0638418 to Jonathan Wendel), and the generous support from the Iowa State Plant Sciences Institute, for awarding me a graduate fellowship.