Evolutionary proteomics of cotton

Guanjing Hu
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd
Part of the Evolution Commons, and the Genetics Commons

Recommended Citation
Hu, Guanjing, "Evolutionary proteomics of cotton" (2013). Graduate Theses and Dissertations. 13415.
https://lib.dr.iastate.edu/etd/13415

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

by

Guanjing Hu

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

DOCTOR OF PHILOSOPHY

Major: Genetics

Program of Study Committee:
Jonathan F. Wendel, Major Professor
Dennis Lavrov
Erik Vollbrecht
Guru Rao
John Nason

Iowa State University
Ames, Iowa
2013

Copyright © Guanjing Hu, 2013. All rights reserved.
I dedicate this work to:

My family,

my father Bangyao Hu,

my mother Jian Guo,

and also to my boyfriend, James Mitch Conner.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>ABSTRACT</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHAPTER 1  GENERAL INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Dissertation Organization</td>
<td>1</td>
</tr>
<tr>
<td>Description of Research Objectives</td>
<td>3</td>
</tr>
<tr>
<td>References</td>
<td>8</td>
</tr>
<tr>
<td>CHAPTER 2 GENOMICALLY BIASED ACCUMULATION OF SEED STORAGE PROTEINS IN ALLOPOLYPLOID COTTON</td>
<td>13</td>
</tr>
<tr>
<td>Abstract</td>
<td>13</td>
</tr>
<tr>
<td>Introduction</td>
<td>14</td>
</tr>
<tr>
<td>Materials and Methods</td>
<td>16</td>
</tr>
<tr>
<td>Results</td>
<td>21</td>
</tr>
<tr>
<td>Discussion</td>
<td>30</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>38</td>
</tr>
<tr>
<td>References</td>
<td>38</td>
</tr>
<tr>
<td>Figures and tables</td>
<td>48</td>
</tr>
<tr>
<td>Supporting information</td>
<td>59</td>
</tr>
<tr>
<td>CHAPTER 3 PROTEOMIC PROFILING OF DEVELOPING COTTON FIBERS FROM WILD AND DOMESTICATED GOSSYPIUM BARBADENSE</td>
<td>62</td>
</tr>
<tr>
<td>Abstract</td>
<td>62</td>
</tr>
<tr>
<td>Introduction</td>
<td>63</td>
</tr>
<tr>
<td>Materials and methods</td>
<td>65</td>
</tr>
<tr>
<td>Results</td>
<td>70</td>
</tr>
<tr>
<td>Discussion</td>
<td>76</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>87</td>
</tr>
<tr>
<td>References</td>
<td>88</td>
</tr>
<tr>
<td>Figures and tables</td>
<td>100</td>
</tr>
<tr>
<td>Supporting Information</td>
<td>108</td>
</tr>
<tr>
<td>CHAPTER 4 PROTEOMIC PROFILING OF FIBER DEVELOPMENT AND DOMESTICATION IN UPLAND COTTON (GOSSYPIUM HIRSUTUM L.)</td>
<td>113</td>
</tr>
<tr>
<td>Abstract</td>
<td>113</td>
</tr>
<tr>
<td>Introduction</td>
<td>114</td>
</tr>
<tr>
<td>Materials and methods</td>
<td>117</td>
</tr>
<tr>
<td>Results</td>
<td>121</td>
</tr>
</tbody>
</table>
Discussion ..................................................................................................................... 128
Acknowledgements ....................................................................................................... 137
References ..................................................................................................................... 137
Figures and tables .......................................................................................................... 144
Supporting Information ................................................................................................. 154

CHAPTER 5  GENE-EXPRESSION NOVELTY IN ALLOPOLYPLOID COTTON: A PROTEOMIC PERSPECTIVE ............................................................................................................. 156
Abstract ......................................................................................................................... 156
Introduction ..................................................................................................................... 157
Materials and Methods ................................................................................................. 161
Results ............................................................................................................................. 167
Discussion ....................................................................................................................... 174
Acknowledgements ........................................................................................................ 181
Reference ......................................................................................................................... 181
Figures and tables .......................................................................................................... 192

CHAPTER 6  GENERAL CONCLUSION ............................................................................. 209
ABSTRACT

Polyploidy has played a spectacular role in plant evolution. Through allopolyploidization, two or more divergent genomes are combined into a common nucleus, which is often followed by rapid as well as longer-term genomic and epigenetic responses. However, since mRNA abundance and protein amounts are poorly related, still little is known how the genomic changes associate with or affect the actual plant function and adaption. Within this context, this dissertation specifically explores the evolutionary history of cotton species on the proteomic level, involving natural polyploidization and human-mediated cotton fiber selection.

In the work presented here, cotton proteomes of mature seeds and developing fibers were examined to understand the genome-wide proteomic changes accompanying allopolyploidization. Interspecific comparisons of protein composition and expression level among diploid and polyploid cotton species revealed that the allopolyploid cotton display asymmetric proteomic divergence with respect to the diploid progenitors, which is detailed by describing important phenomena of non-additive expression, expression level dominance and homoeolog expression bias.

To understand the genome-wide expression changes associated with domestication, the fiber proteomes represented by four representative developmental stages were studied for two important crop species - *G. hirsutum* and *G. barbadense*, using paired wild and domesticated accessions. By contrasting the fiber proteomes and developmental dynamics between wild and domesticated accessions for each species, we
were able to characterize the key features of global protein pattern change corresponding to the domestication processes. Furthermore, we identified a number of proteins differentially expressed during fiber development and altered by domestication, as candidate proteins and metabolic processes for functional analyses that may yield insight into domestication and future cotton improvement.
CHAPTER 1

GENERAL INTRODUCTION

Dissertation Organization

This dissertation is organized into six chapters. The first chapter provides a general introduction to polyploidy, followed by illustrating the important role of proteomics in evolutionary studies. After introducing the *Gossypium* system and its advantages for evolutionary proteomics, I describe four primary objectives of this thesis.

Chapter 2 describes our first exploration of gene expression at the proteome level in response to allopolyploidization. Using two-dimensional gel electrophoresis (2-DE), we conducted a comparative analysis of cotton seed proteomes from the allopolyploid *Gossypium hirsutum* (AD genome) and its model A-genome and D-genome diploid progenitors. This chapter was published in the journal *Genetics* in 2011. I undertook this research with Norma L. Houston, Dharminder Pathak, Linnea Schmidt, Jay J. Thelen and Jonathan F. Wendel. Jonathan and I designed the research. I conducted protein extraction and 2-DE experiments with the help from Norma and Jay, who also managed the mass spectrometry workflow and protein identification. Dharminder Pathak and Linnea Schmidt helped me with genetic analysis of seed storage genes. I drafted the manuscript with help from Jonathan and all authors offered helpful edits and comments.

Chapter 3 presents a comparative analysis of fiber development in wild and domesticated *G. barbadense*, using a mass spectrometry-based iTRAQ proteomic technology for protein identification and quantitative analysis. This chapter was
published in the journal *New Phytologist* in 2013. I undertook this research with Jin Koh, Mi-Jeong Yoo, Kara Grupp, Sixue Chen, and Jonathan F. Wendel. Jonathan and I designed the research. Kara helped me to collect plant tissues. Jin and Sixue performed the iTRAQ experiments. I analyzed the data with the help from Jin and Mi-Jeong. I drafted the manuscript with help from Jonathan and all authors offered helpful edits and comments.

Chapter 4 uses wild and domesticated *G. hirsutum* as a parallel model to study proteomic variation with respect to fiber development and crop domestication. In addition to the utilization of iTRAQ technique, as in Chapter 3, 2-DE analysis was included to provide a comparative and more comprehensive characterization of the fiber proteome. This work is being prepared for submission to the journal *Molecular Biology and Evolution*, and it is in a “nearly submission ready” draft form. This research was performed with Jin Koh, Dharminder Pathak, Sixue Chen, and Jonathan F. Wendel. Jonathan and I planned the project. Jin and Sixue performed the iTRAQ experiments. Dharminder helped me with 2-DE experiments. I analyzed the data and drafted the manuscript with the help from Jonathan.

Chapter 5 expands our understanding of fiber proteomes from Chapters 3 and 4, by analyzing proteomic divergence between allopolyploid cotton, using both *G. hirsutum* and *G. barbadense*, and their diploid A- and D-genome progenitors. This study also expands on Chapter 1 by studying the proteomic consequences of allopolyploidization with another tissue type, and the one that is of most concern to the cotton industry – cotton fibers. Paralleled application of iTRAQ and 2-DE methods was
performed, as in Chapter 4. This work is being prepared for submission to the journal *Heredity* shortly. This research was performed with Jin Koh, Dharminder Pathak, Sixue Chen, and Jonathan F. Wendel. Jonathan and I planned the project. Jin and Sixue performed the iTRAQ experiments. Dharminder helped me with 2-DE experiments. I analyzed the data and drafted the manuscript with the help from Jonathan.

Chapter 6 provides a brief general conclusion and discusses the four primary research objectives in light of the four original research Chapters.

**Description of Research Objectives**

**Polyploidy - a major force in plant evolution**

Genome doubling, or polyploidy, is a major process of speciation and genome evolution, particularly in plants (Soltis & Soltis, 2000; Wendel, 2000; Osborn et al., 2003; Adams & Wendel, 2005b; Adams & Wendel, 2005c; Chen & Ni, 2006; Chen, 2007; Soltis et al., 2009). Many important crops have long been recognized as polyploids: watermelon, strawberries, apples, potato and alfalfa possess multiple chromosome sets of a single genome (autopolyploidy); other crops, including wheat, canola, tobacco, peanut, and cotton, have experienced genome merging from two or more different progenitor species (allopolyploidy). All species once considered as typical diploids, such as *Arabidopsis* and rice, are now known to be ancient polyploids (paleopolyploids), followed by diploidization, a process that leads to massive gene loss and restoration of normal bivalent pairing (Bowers et al., 2003; Ilic et al., 2003; Blanc & Wolfe, 2004; Cheng et al., 2013). It has also become clear that all flowering plants have
undergone several rounds of polyploidy during their genomic history (Jiao et al., 2011). However, the question regarding why polyploids are so common and successful is still not well understood. As early as 1951, Stebbins proposed that multiple genomes in the same nucleus provide increased allelic diversity and heterozygosity, which would lead to novel phenotypic variation and high stress tolerance, and hence a higher adaptability (Stebbins, 1971). With the advent of new molecular and genomic tools, recent recognition of genome-level rearrangement and diversified fates of duplicated genes such as subfunctionalization and neofunctionalization, provide new perspectives to explore the genomic and genetic attributes and functional consequences in polyploids (Lynch & Conery, 2000; Lynch & Force, 2000; Force et al., 2005; He & Zhang, 2005; Grover et al., 2012; Soltis & Soltis, 2012; Madlung & Wendel, 2013).

From genomics to proteomics

Today, the use of genomic sequence data and transcriptomic methods allows molecular description of the evolution of whole plant genomes, as well as the regulation of homoeologous genes in polyploidy species. Recent work in our lab described transcriptomic responses to polyploidy in cotton (Adams et al., 2003; Adams & Wendel, 2005a; Flagel et al., 2008; Hovav et al., 2008a; Hovav et al., 2008b; Hovav et al., 2008c; Hovav et al., 2008d; Rapp et al., 2009; Salmon et al., 2010; Yoo et al., 2013). However, as the prime targets of evolutionary selection, phenotypes and biological mechanisms are essentially influenced by proteins rather than transcripts. Moreover, there is no linear correlation between RNA transcription and protein abundance. Further,
no information on post-translational protein modification, which may play a key role in molecular interaction in cells, can be deduced from genomic studies. Therefore, it is critical that we extend the frontier of functional genomics of evolutionary study into the realm of the proteome, as a prelude to a fuller systems biology approach that integrates across the various “omics” that lie between genotype and phenotype (Olsen & Wendel, 2013a; Olsen & Wendel, 2013b). Early work on Brassica (Albertin et al., 2006; Albertin et al., 2007) using two-dimensional electrophoresis to assess the response of proteomes to allopolyploid formation, suggests that this will be a promising arena to enhance our knowledge about plant adaption and function.

Cotton as a model for evolutionary proteomics

The cotton genus (Gossypium L.) includes 45 diploid species and five allotetraploid species. Two diploid groups of species, known as Old World A-genome and New World D-genome, diverged from a common ancestor about 5-10 million years ago. These two genomes became reunited in a common nucleus through the allopolyploidization event approximately 1-2 million years ago, leading to the origination of the modern polyploid cotton species. Two A-genome diploid species G. arboreum L. and G. herbaceum L, and two allotetraploid species, G. hirsutum L. (‘Upland’ cotton) and G. barbadense L. (‘Pima’ or ‘Egyptian’ cotton) were independently domesticated at least 4000 years ago (Wendel et al., 1995; Dillehay et al., 2007) for cotton fiber. Cultivars derived from the latter two species now dominate the world cotton commerce. This complex evolutionary history of both natural and human-
mediated selection, as well as the marvelous variation among cotton species with respect to their seed fibers, makes cotton as an excellent model for evolutionary studies. These years, fast growing genomic resources of cotton including a recently completed genome sequence of \textit{G. raimondii} (Paterson \textit{et al.}, 2012), have greatly increased the research potentiality for proteomic study, of which the mass spectrometry data interpretation relies on searching genomic sequence database. In addition, proteome profiling of cotton species, along with our available transcriptomic data of cotton fiber development and evolution would provide us a unique opportunity to decipher the connection of cotton genotype and phenotypic traits in the manner of system biology.

\textbf{Figure 1} The evolutionary history of diploid and allotetraploid cotton species. Images of a single seed with attached trichomes (“cotton fiber”) are shown from A-genome \textit{G. arboreum}, D-genome \textit{G. raimondii}, and a wild form (accession TX2094; YUC) and a domesticated stock TM1 (TM1) of \textit{G. hirsutum}, a wild form (K101) and a modern cultivar (Pima S-7) of \textit{G. barbadense}.
Research objectives

My doctoral research aims to develop our understanding how evolutionary history, including natural polyploidization and human-mediated domestication, sculpted cotton proteomes. These in-depth functional genomic data are also expected to provide us new opportunities for cotton improvement. Our specific research objectives are:

1. To develop technology and tools for describing and studying the cotton fiber and seed proteomes.

2. To describe the cotton proteome from the standpoint of fiber development, which will allow us to assess the changes that accompany fiber evolution and domestication, and how this correlates with existing information on the transcriptome.

3. To understand how the proteome responds to genome doubling; that is, what is novel about polyploid cotton fiber and seed relative to that of its antecedent diploids?

4. To detail proteomic consequences of cotton fiber evolution and domestication; for example, to catalog the key proteins associated with and therefore possibly responsible for phenotype changes and important traits relevant to crop improvement.
References


CHAPTER 2
GENOMICALLY BIASED ACCUMULATION OF SEED STORAGE PROTEINS IN ALLOPOLYPLOID COTTON

A paper published in Genetics in 2011 (Genetics 189: 3)
Guanjing Hu, Norma L. Houston, Dharminder Pathak, Linnea Schmidt, Jay J. Thelen and Jonathan F. Wendel

Abstract

Allopolyploidy is an important process during plant evolution that results in the reunion of two divergent genomes into a common nucleus. Many of the immediate as well as longer-term genomic and epigenetic responses to polyploidy have become appreciated. To investigate the modifications of gene expression at the proteome level caused by allopolyploid formation, we conducted a comparative analysis of cotton seed proteomes from the allopolyploid G. hirsutum (AD-genome) and its model A-genome and D-genome diploid progenitors. An unexpectedly high level of divergence among the three proteomes was found, with about one-third of all protein forms being genome-specific. Comparative analysis showed that there is a higher degree of proteomic similarity between the allopolyploid and its D-genome donor than its A-genome donor, reflecting a biased accumulation of seed proteins in the allopolyploid. Protein identification and genetic characterization of high abundance proteins revealed that two classes of seed storage proteins, vicilins and legumins, comprise the major component of cotton seed proteomes. Analyses further indicate differential regulation or modification of homoeologous gene products, as well as novel patterns in the polyploid proteome that may result from the interaction between homoeologous
gene products. Our findings demonstrate that genomic merger and doubling have consequences that extend beyond the transcriptome into the realm of the proteome, and that unequal expression of proteins from diploid parental genomes may occur in allopolyploids.

**Introduction**

Genome doubling, or polyploidization, is a phenomenon prevalent in eukaryotes and particularly in higher plants. Genomic studies indicate that all angiosperm species have undergone at least two rounds of polyploidization during their evolutionary history, with most lineages having experienced additional whole genome duplications (Cui *et al.*, 2006; Jiao *et al.*, 2011). Allopolyploid species are particularly intriguing in that their formation entails the merger of diverged genomes, which often results in myriad dramatic and large-scale genomic and transcriptomic responses (Wendel, 2000; Comai, 2005), including structural and epigenetic modifications (Shaked *et al.*, 2001; Gaeta *et al.*, 2007; Buggs *et al.*, 2009; Ha *et al.*, 2009; Schnable *et al.*, 2011), as well as changes in gene expression (Wang *et al.*, 2006; Bottley & Koebner, 2008; Flagel *et al.*, 2008; Hovav *et al.*, 2008; Flagel *et al.*, 2009; Rapp *et al.*, 2009; Flagel & Wendel, 2010; Koh *et al.*, 2010). Compared to their progenitors, polyploids often display different physiological, morphological and ecological phenotypes (Pires *et al.*, 2004; Gaeta *et al.*, 2007; Anssour *et al.*, 2009; Ni *et al.*, 2009; Ramsey, 2011), which suggests functional and phenotypic evolution may be driven by these genomic changes.

Notwithstanding these and other recent insights into the genomic and transcriptomic consequences of genomic merger and doubling, the fate of translated gene products, i. e., the proteome, remains poorly studied in the context of
polyploidization. Because protein levels are influenced by post-translational processing and inherent variation in stability, it is difficult to infer the representation and regulation of proteins and participating metabolic pathways from transcriptomic data alone, and the correlation between protein and transcript expression levels has been shown to vary extensively depending on the system being analyzed and profiling approach used (Hajduch et al., 2010). As proteins represent the key players in cellular activities, characterizing the proteome using appropriately targeted approaches constitutes an important component of the evolutionary analysis of polyploidy and its consequences. A classical proteomic technique, two-dimensional gel electrophoresis (2-DE), has the potential to assess the expression patterns of proteins displayed by polyploid species relative to their diploid progenitors, as demonstrated in Brassica (Albertin et al., 2005; Albertin et al., 2006; Albertin et al., 2007). This technique allows the resolution of hundreds of protein spots within a single gel, which are accessible to identification through Mass Spectrometry (MS) analysis; moreover, some post-translational modifications corresponding to protein activities can be inferred via interpretation of the on-gel and MS properties. This comparative quantification of resolved spot profiles permits a proteome-scale comparison of the polyploid and its parental species.

Over the past decade, Gossypium has emerged as a model for studies of polyploidy, particularly with respect to the genomic and transcriptomic consequences of allopolyploidization (Adams et al., 2003; Senchina et al., 2003; Grover et al., 2004; Grover et al., 2007; Flagel et al., 2008; Hovav et al., 2008; Chaudhary et al., 2009; Flagel et al., 2009; Rapp et al., 2009; Flagel & Wendel, 2010; Salmon et al., 2010). As shown in Figure 1A, A- and D-genome Gossypium diverged for approximately 5-10 million years before becoming reunited in an allopolyploid
nucleus about 1-2 million years ago (Wendel & Cronn, 2003). Extensive research has identified the best models of the diploid progenitors involved in the creation of the allopolyploid lineage, the latter of which includes the most important of the cultivated species, *G. hirsutum*. This well-documented evolutionary framework, coupled with the substantial resources available, e. g., a comprehensive EST database (Udall *et al.*, 2006a; Udall *et al.*, 2006b), and the prior genomic/transcriptomic research into the consequences of polyploidy (Adams *et al.*, 2003; Senchina *et al.*, 2003; Grover *et al.*, 2004; Grover *et al.*, 2007; Flagel *et al.*, 2008; Hovav *et al.*, 2008; Chaudhary *et al.*, 2009; Flagel *et al.*, 2009; Rapp *et al.*, 2009; Flagel & Wendel, 2010; Salmon *et al.*, 2010), makes *Gossypium* an excellent system to extend research on genomic merger and doubling to the proteomic level. In this study, we profile and analyze the proteomes of cotton seeds in the polyploid (AD genome) *G. hirsutum*, and its two model diploid progenitors, *G. herbaceum* (A genome) and *G. raimondii* (D genome). Despite being best known for fiber production, the high-quality oil and proteins produced in the seeds of domesticated *G. hirsutum* have increased the agronomic and economic importance of cotton as a crop plant. The increased interest in cotton seeds (Sunilkumar *et al.*, 2006), and the relatively simplified protein composition of mature, dormant seeds, makes cotton a useful model for studying protein accumulation in the context of polyploidy.

**Materials and Methods**

**Plant materials**

Three *Gossypium* species were used in the present study: one polyploid species *G. hirsutum var. Acala Maxxa* (AD-genome), and two diploid species that represent the model diploid progenitors of allopolyploid cotton, namely *G. herbaceum*
(A1-73; A-genome) and G. raimondii (D-genome). For each species, seeds were collected and pooled from three to four plants that were grown in the Pohl Conservatory at Iowa State University, Ames, Iowa. After boll opening, mature seeds were hand-harvested and air-dried at room temperature for at least one month. Prior to protein extraction, the fiber-containing seeds were de-linted with concentrated sulphuric acid. The weight of the seeds collected was measured both before and after de-linting.

**Protein extraction**

Total protein was isolated from mature desiccated seeds based upon a phenol extraction procedure successfully applied to other oilseeds (Hajduch et al., 2005; Hajduch et al., 2006; Hajduch et al., 2007; Houston et al., 2009). Briefly, for each sample, 250 mg of de-linted seeds was ground to a fine powder in liquid nitrogen with a mortar and pestle, and homogenized with 10 mL of a 1:1 mixture of extraction buffer (100 mM Tris-HCl pH 8.8, 900 mM sucrose, 10 mM EDTA and 0.4% [v/v] 2-mercaptoethanol) and Tris-saturated phenol. The homogenate was agitated for 30 min and centrifuged at 4000 rpm for 30 min at 4°C. The upper phenol phase was extracted and combined with 5 volumes of 0.1 M ammonium acetate in methanol and placed at -20°C for overnight protein precipitation. The protein pellet was subsequently collected via centrifugation for 30 min at 4°C. The recovered pellet was thoroughly washed over 4 washing steps: once with 0.1 M ammonium acetate in methanol, twice with 80% acetone, and once with 70% ethanol. The final pellet was dried at room temperature and solubilized in Isoelectric Focusing (IEF) buffer (8 M urea, 2 M thiourea, 2% [w/v] CHAPS, 2% [v/v] Triton X-100, 50 mM DTT). Protein concentration was determined using the modified Bradford total protein assay.
(Bradford, 1976) from Bio-Rad (Hercules, CA) using bovine gamma globulin as standard.

2-DE

As described in (Hajduch et al., 2005), 1 mg of seed proteins was dissolved in 450 µL IEF buffer and separated by Isoelectric Focusing (IEF) for the first dimension and by SDS-PAGE in the second dimension. IEF was performed in a Bio-Rad PROTEAN IEF System (Hercules, CA) using 24 cm linear immobilized pH gradient (IPG) strips with pH ranges of 4-7 and 3-10 (GE Healthcare), using the following conditions: active rehydration at 50 V for 10 h, 100 V for 100 V·h, 500 V for 500 V·h, and 8000 V for 99 kV·h. After completion of IEF, the strips were prepared for SDS-PAGE as follows: the strips were reduced for 15 min with 2.0% [w/v] DTT in equilibration buffer (6 M urea, 50 mM Tris-HCl pH 8.8, 30% [v/v] glycerol, 2% [w/v] SDS), alkalized for 15 min with 2.5% [w/v] iodoacetamide in equilibration buffer, rinsed with SDS running buffer (1.5 M Tris-HCl, 6 M urea, 30% [v/v] glycerol, 5% [w/v] SDS), and then transferred onto 12% self-cast polyacrylamide gels. The second dimension SDS-PAGE was performed in an Ettan DALT 12 System (GE Healthcare) using 1 W/strip for 1 h and 2 W/strip for 15 h. Finished gels were washed twice in deionized water for 10 min and stained overnight with colloidal Coomassie (20% [v/v] ethanol, 1.6% [v/v] phosphoric acid, 8% [w/v] ammonium sulfate, 0.08% [w/v] Coomassie Brilliant Blue G-250). Stained gels were stored in 250 mL storage solution (10% v/v CBB solution, 0.02% w/v sodium azide) per gel at 4 °C.

Image and statistical analysis: Gels were imaged with a ScanMaker 9800XL (Microtek, Carson, CA, USA) using a resolution of 300 dpi and 16-bit grayscale pixel depth. Image analysis was conducted with ImageMaster 2D platinum software version
6.0 (Amersham Biosciences, Uppsala), which allows spot detection, quantification, and cross-image spot matching. Using the built-in normalization method implemented in the ImageMaster 2D platinum software, spot expression was represented by relative spot abundance, dividing each absolute spot volume by the total volume of all spots selected for analysis. For spots shared by all three genomes, differential protein expression was tested using a one-way ANOVA model with a fixed effect: \( Y_{ij} = \mu + G_i + e_{ij} \), where \( \mu \) represents the overall mean, \( G_i \) denotes a genome fixed effect, and \( e_{ij} \) is the random error term used for significance test. When spots were shared by all three genomes, the hypothesis of additive parental expression in the allopolyploid was tested; for this, a spot was considered additive with respect to expression if the spot abundance in the polyploid AD-genome was equal to the average abundance found in the parental A- and D-genomes. Any deviation from the average parental value was considered non-additive expression, which then was further categorized by comparing the AD-genome value to both diploid values. Possible deviations from additivity include: parental genomic dominance (where the expression found in the AD-genome is statistically equal to one parental value for spots differentially expressed between parents), and transgressive expression (where the expression in the polyploid AD-genome statistically falls either below or above that found in both parental genomes). The p-values of these analyses were adjusted for multiple testing (Benjamini & Hochberg, 1995), and the false discovery rate (FDR) was controlled at 5%.

**MS analyses**

Selected spots for protein identification were excised from gels and subjected to in-gel trypsin digestion followed by LC-MS/MS (liquid chromatography-tandem mass spectrometry) using an LTQ XL ProteomeX ion trap mass spectrometer.
Thermo-Fisher, San Jose, CA). Tandem mass spectral data were searched against the Arabidopsis protein database and an in-house Gossypium EST/contig translated database (provided by J Udall), using SEQUEST, which is part of the BioWorks 3.1SR1 software suite (ThermoFisher, San Jose, CA). The instrument and searching parameters were applied according to (Hajduch et al., 2006).

**Genetic analysis**

Primer sequences (Table S1) designed to amplify the suite of SSP genes were derived from the conserved regions of each SSP gene in Gossypium, identified from the alignment of publicly available G. hirsutum sequences [GenBank: M19378.1, M16891.1, M69188.1, M16936.1] and SSP ESTs (identified by blast hits against an in-house cotton EST database). Amplified PCR products were excised from 1.0% agarose gels, purified using Qiaquick gel purification kit (Qiagen Inc., Valencia, CA), and cloned with the pGEM®-T Easy Vector System (Promega, Wisconsin). Cloned products were sequenced using the Applied Biosystems 3730xl DNA Analyzer at ISU DNA facility. Because the PCR products are approximately 2000bp in length, internal primers were also designed for sequencing. The resulting sequences were aligned with those derived from GenBank and the cotton EST database using MUSCLE (Edgar, 2004) and inspected manually. The obtained Gossypium SSP gene sequences were deposited in Genbank under accession numbers JN602029-JN602047. Neighbor-joining analysis was conducted on the aligned sequences using PAUP* (Swofford, 2001). Uncorrected (“p”) DNA/RNA distances were set for distance analysis and missing data were ignored for affected pair-wise comparisons.
Results

The proteomic profiles of mature cotton seeds

Allopolyploid *G. hirsutum var. Acala Maxxa* (AD genome) and representatives of its diploid progenitors (A genome *G. herbaceum* and D genome *G. raimondii*) were used to profile the cotton seed proteome. Mature seed mass and size of intact and de-linted seeds were recorded before protein extraction (Figure 1B). The protein yields from phenol extraction ranged from 6.3% to 8.9% (dry weight) without significant variation among three genomes, which is in agreement with the seed protein contents previously reported for *Gossypium* (Frampton *et al.*, 1958; Pandey & Thejappa, 1975).

Mature seed proteins from each genome were isolated in biological quadruplicate for two-dimensional gel electrophoretic (2-DE) separation (Figure S1). In order to construct a high-quality proteomic map, two overlapping ranges of immobilized pH gradient (IPG) strips were used for the first dimensional Isoelectric Focusing (IEF): a broader range pH 3-10 strip and a narrower range pH 4-7 strip. As shown in Figure 2A, the use of pH range 4-7 largely enhanced spot resolution in the signal-dense area of pH 3-10; therefore, the proteomic profiles were constructed using spots detected in the pH 4-7 gels and the subsection of the pH 3-10 gels containing the pH 7-10 region. For each gel of each pH range, detected spots were matched within biological replicates and then between genomes. Spot detection was only considered for spots reproducibly represented by at least three biological replicates; spots meeting this criterion were selected for profiling and subjected to qualitative and quantitative analyses. According to this criterion, 646 spots were confidently detected from pH 4-7 gels, and 208 spots were resolved from pH 3-10 gels (pH 7 to 10 region only). Of the 854 total spots, 315 were present in the polyploid (AD-genome) seed
proteome, and fewer were recovered from the A- and D-genome seed proteomes (250 and 289 spots, respectively).

**Comparative proteomics of allopolyploid *G. hirsutum* and models of its diploid progenitors**

Proteomic profiles of the allopolyploid and two progenitor diploid species were first compared qualitatively through spot matching between genomes (Table 1; File S1). The pairwise comparison between genomes revealed that only 92 spots were observed in all cotton seed proteomes, corresponding to 29.0%, 36.8% and 32.2% of the total spots detected from the AD-, A- and D-genomes, respectively. Surprisingly, given the high degree of genetic similarity between A and D genome orthologs (Senchina et al., 2003), about one third of the spots from each genome were found to be genome-specific. This result shows that considerable proteomic variation exists not only between diploid species but also among polyploid and diploid genomes in *Gossypium*. Spots represented by only two genomes were also noted, and interestingly, 26.2% of AD-genome spots were found in the D-genome, whereas only 10.7% of the polyploid proteome was represented by the A-genome, suggesting a higher similarity between AD- and D-genome seed proteomes than between the AD- and A-genomes. This compositional bias towards the D-genome suggests an unequal contribution of the diploid genomes to global protein expression in cotton seeds, a potential response to hybridization and genome doubling similar to that experienced by the transcriptome (Flagel et al., 2008; Rapp et al., 2009).

In addition to the qualitative variation observed in the 2-DE spot patterns, differential quantitative expression of shared spots (illustrated in Figure 3) was analyzed. Quantitative changes attributed to polyploidy were inferred by testing the
additivity of parental contributions to the allopolyplloid proteome, and by classifying the expression patterns (see examples in Figure 4) using a two-step procedure. This procedure first tested for non-additivity, i.e. spots that deviated significantly in abundance from the average of the parental diploids. These non-additive spots were further categorized by comparison to their homologous expression levels in both parents (Table 2). Of the 92 common spots analyzed, 33 spots (35.9%) were detected as non-additive, among which 8 spots were expressed transgressively; that is, their expression was either greater or less than both parental diploids. Twenty-two non-additive spots displayed statistically equivalent expression as one of the two diploid parents, and 13 and 9 spots were sorted into the D-dominant and A-dominant patterns, respectively. The remaining 3 spots displayed an intermediate level of expression between the parental values and were considered co-dominant.

Identification of major components of cotton seed proteomes

To characterize and compare the major components of seed proteomes in the three *Gossypium* species, high-abundance spots (>1% mean relative volume of each genome at each pH range) were targeted for Mass Spectrometry (MS)-based protein identification. Because it is also of interest to determine the proteins or functional categories that contribute to the expression patterns observed in the polyploid, representative spots from the expression categories defined in Table 2 were also included for protein identification. According to these nonexclusive criteria, a total of 199 spots corresponding to approximately 80% of the total spot abundance for the three species were subjected to tandem MS analysis. Searching against a custom *Gossypium* and *Arabidopsis* protein database successfully identified 155 spots (62 from AD-genome, 55 from A-genome, and 53 from D-genome; see File S2 for spot
selection and identification). The majority of the identified spots (140 spots) belonged to the category of seed storage proteins (SSPs), including vicilin A (19 spots), vicilin B (5 spots), legumin A (83 spots) and legumin B (27 spots), and one vicilin-like gene (6 spots) not previously reported in *Gossypium*. The remaining spots (9.7%, 15/155) identified were classified to the functional categories of cellular organization (4 spots), molecular function (4 spots), and stress response (3 spots).

Due to the experimental design of using two IEF pH ranges, the relative expression of identified spots was independently profiled from pH range 4-7 (pH 4-7 gels) and 7-10 (subsection of the pH 3-10 gels). To estimate the overall protein composition of each cotton seed proteome, the percent spot abundances for pH 4-7 or pH 7-10 sub-proteomes were normalized by their composition relative to the full pH 4-10 range (0.4 to pH 4-7, 0.6 to pH 7-10; calculated using the ratio of total spot abundances of pH 4-7 and 7-10 subsections on the pH 3-10 gels), and summed for the identified proteins, as shown in Table 3. Two principal SSP families, vicilin and legumin, constituted a major fraction of the cottonseed proteomes, representing 71%, 68% and 72% of the total protein in the AD-, A- and D-genome species, respectively. Although the overall accumulation of SSPs appears constant among the three species, variation in relative expression of each individual SSP was observed among the diploid parents and the polyploid. For example, vicilin A was more highly expressed in the AD genome (19% of the total protein abundance) than in its diploid progenitors (13% and 10%, in the A- and D- parents, respectively); in contrast, the allopolyploid species utilized less legumin A (18%) as a nutrient reservoir than either of the two diploid species (26% and 33% in the A- and D- parents, respectively). Moreover, the composition of both legumin B and the vicilin-like protein in the AD-genome seeds displayed a similar expression pattern to the D-genome progenitor, which could
clearly be distinguished from the pattern of the A-genome progenitor (Table 3). This previously unreported vicilin-like protein was identified via BLAST homology to a *Pistacia vera* vicilin protein [GenBank: ABO36677.1], and was detected only in the AD- and D-genome species, while no corresponding spots or peptides were detected in the A-genome species (Figure 5); this pattern was also evident by surveying the *Gossypium* EST database (data not shown). This indicates that expression of this vicilin-like gene may be specific to the diploid D-genome species, and was subsequently recruited into AD-genome species during or post-allopolyploidization.

**Genetic analysis of *Gossypium* SSP genes**

The major *Gossypium* SSPs have previously been characterized (Dure & Chlan, 1981; Galau *et al.*, 1983; Chlan *et al.*, 1986; Chlan *et al.*, 1987; Dure, 1989; Galau *et al.*, 1991), although complete gene sequences were limited to *G. hirsutum*. In surveying the *Gossypium* EST databases, a considerable level of nucleotide diversity became evident, not only for the orthologous genes obtained from the diploid species, but also for copies found in the *G. hirsutum* EST database. These data indicated that some of the major SSPs in *Gossypium* are encoded by multigene families (data not shown). Thus, to understand the genetic basis of the proteomic profiles and their compositional diversity, gene family structures were characterized using conserved primers designed to amplify each SSP from four *Gossypium* species: AD-genome *G. hirsutum* var. Acala Maxxa, D-genome *G. raimondii* and two A-genome species *G. herbaceum* (which was used in the proteomic analysis, noted as A1 here) and *G. arboreum* (another putative A-genome progenitor of allopolyploid cotton, noted as A2, included for additional perspective on the genetic diversity of *Gossypium* SSP genes).
Vicilin A and vicilin B, which share 72% amino acid similarity, belong to the vicilin (7S globulin or alpha globulin) gene family (Chlan et al., 1987), and represent the first discovered cotton SSPs. To test the single-copy status for both vicilins, over ten sequences per gene were generated from each of the four Gossypium species mentioned above. After removing sequencing errors and redundancy, both vicilins were determined to exist as single-copy genes in the diploid genomes (A₁, A₂, D), corresponding to two homoeologous copies (Aₜ deriving from the A-genome progenitor, modeled by A₁ or A₂; Dₜ deriving from the D-genome progenitor) that were retained in the allopolyploid AD-genome. Gene trees for both vicilin genes were generated using the neighbor-joining method (Figure 6), and the same tree topologies were resolved using Maximum Likelihood and Maximum Parsimony methods (data not shown). These trees also were congruent with the phylogeny of Gossypium (Cronn et al. 2002; Wendel and Cronn 2003) shown in Figure 1A. Nucleotide variation, including indels and non-synonymous and synonymous substitutions, were identified among the orthologous and homoeologous genes copies for each gene (see File S3 for gene sequences and alignments). As noted on the branches of gene trees, the majority of substitutions were inferred to have occurred since the divergence of the A- and D-genome divergence from their common ancestor, with a few lineage-specific substitutions having arisen after polyploidization 1-2 million years ago. The Aₜ copy of both vicilin genes exhibited less lineage-specific nucleotide substitution than did the Dₜ copies in polyploid AD-genome, which likely is explained by the fact that the A-genome diploid species used are better models of the actual A-genome progenitor than the D-genome diploid is of the actual D-genome progenitor of allopolyploid cotton (Senchina et al., 2003).
The legumin (11-12S globulin or beta globulin) gene family is the other major SSP group found in *Gossypium*, and its members, legumin A and legumin B, are more diverged compared to the vicilin gene family members, sharing only 58.5% similarity in amino acid sequences. Multiple sequences of legumin A and legumin B were also generated and characterized from allopolyploid and diploid *Gossypium*. As with the vicilins, both legumins were also found as single-copy genes in the diploid genomes; however, only the D-genome derived copy was detected for legumin B in the AD-genome. This observation was further supported by the absence of an A-genome derived copy in the cotton EST database. Two possible explanations exist for the loss of the original A-genome derived copy in the polyploid: gene deletion and concerted evolution that resulted in the homogenization of the homoeologous pair towards the D-genome derived copy, a phenomenon previously demonstrated for ribosomal genes in allotetraploid *Gossypium* (Wendel et al., 1995), and more recently for numerous protein-coding genes (Salmon et al., 2010). Interestingly, despite the fact that both A- and D-genome derived copies were recovered for legumin A, a non-synonymous substitution in the A-genome derived copy of legumin A caused a premature stop codon. Additionally, an accelerated rate of nucleotide substitution in the legumin A A-genome derived copy was also observed (Figure 6), which suggests that this copy is non-functional. Together, these observations suggest that there exist different regulatory mechanisms and uneven selection pressures on vicilin and legumin genes, even though they both function as storage proteins in cotton seeds.

**Detailed proteomic characterization of *Gossypium* SSPs**

An observation common to 2-DE gels, and pertinent to the analysis of the SSP proteome in *Gossypium*, is that numerous spots often correspond to isoforms of the
same protein accession, as previously demonstrated with 2-DE analyses of SSPs in pea, soybean, rapeseed and Arabidopsis (Hajduch et al., 2005; Hajduch et al., 2006; Higashi et al., 2006; Bourgeois et al., 2009). By contrasting the isoform peptide sequences obtained through MS analysis to the full-length proteins, the on-gel spot location and the computationally predicted location can be examined to determine the formation and features of SSP isoforms, used here to characterize the isoforms of the vicilins (Figure 7). In the three Gossypium species studied, the most abundant vicilin A and vicilin B spots were identified to have molecular weights of 48 kDa (Figure 7; isoform “a”) and 52 kDa (Figure 7, isoform “e”), respectively, and both were composed of a horizontal isoform series spanning from pH 6 to 8.5. These spots were previously reported as the mature forms of vicilin that were processed through a series of post-translational modifications (Dure & Chlan, 1981; Dure & Galau, 1981). By mapping the peptides derived from MS analysis to the full-length protein sequences, these two isoform series were characterized and determined to derive from the approximately 70 kDa vicilin A and vicilin B prepropolypeptides through the cleavage of signal peptides together with the N-terminal fragments, respectively. Similarly, less abundant vicilin isoforms (Figure 7; isoforms “b”, “c”, “d”, and “f”) observed at a lower molecular weight were also evaluated and characterized as the products of proteolytic cleavage or peptide degradation. Protein modifications (e. g., glycosylation, phosphorylation, acetylation and methylation) also likely contributed to the formation of these vicilin isoforms, which can be inferred by slight shifts in spot pI and/or molecular weight and thus were evident when comparing the isoform on-gel and predicted locations.

It is worth noting that these isoform patterns varied not only among polyploid and diploid species, but also between the A_T and D_T homoeolog-derived isoforms
within the allopolyploid AD-genome. Using the analyzed SSP gene sequences as a reference, amino acid variations between the homoeologous peptides were identified, to enable diagnosis of whether the AD-genome polypeptides and their corresponding modifications were inherited from the A- or D- parental species, or if they exhibited a novel pattern after polyploidization. As shown in Figure 7, the 48 kDa (Figure 7; isoform “a”) and 52 kDa (Figure 7; isoform “e”) polypeptides common to all species analyzed were expressed by both homoeologous genes in AD-genome. Alternatively, only A_T polypeptides (Figure 7; isoform “c”) were recovered in the acidic 12-15 kDa region, consistent with the A-genome-specific pattern in diploids. A novel modification in the allopolyploid was observed for the 17 kDa vicilin B polypeptides (Figure 7; isoform “f”), which appears to be the result of retention of the N-terminal fragment of the vicilin B precursor in the allopolyploid only (whereas the parental diploids experience cleavage and degradation of this fragment). The polypeptides of this fragment were further determined to originate from expression of the D_T homoeolog of vicilin B; however, not enough peptide information was recovered to completely rule out the presence of A_T homoeolog products. Altogether, these findings suggest differential regulation or modification of homoeologous gene products, as well as novel patterns in the polyploid proteome that may result from the interaction between homoeologous gene products.

More than thirty spots were identified corresponding to legumin isoforms in each *Gossypium* species, commonly distributed at molecular weights of 30 kDa, 17-20 kDa and 11-12 kDa as legumin A, and at a molecular weight of 11-13 kDa as legumin B. As with vicilins, these legumin isoforms are also processed through a series of modifications, including proteolytic cleavage and peptide degradation. Isoform analysis through peptide mapping indicated that the 30 kDa polypeptides of
legumin A derived from the C-terminal fragment of the 58 kDa prepropolypeptide (see File S4 for peptide mapping analysis). Other isoform peptide sequences obtained through MS analysis failed to be clustered and mapped to continuous polypeptide regions, perhaps reflecting a lower peptide coverage recovered from MS analysis compared to that of vicilins. The contribution of homoeologous polypeptides within the allopolyploid was also evaluated, showing that all peptides detected for legumin A and legumin B were encoded by the \(\text{D}_\text{T}\) gene copy. This result is consistent with the gene family structure of legumins: in the allopolyploid AD-genome, only the D-genome derived copy of legumin B exists, and the A-genome derived copy of legumin A appears to be non-functional, due to a premature stop codon (as noted above). Considering this strict \(\text{D}_\text{T}\) homoeologous expression of legumin isoforms in the AD-genome, the legumin SSPs are possibly the key components that contribute to the biased accumulation of cotton seed proteins in allopolyploid cotton.

**Discussion**

**Vicilin and legumin are the major proteins in mature cotton seeds**

Seed storage proteins, which accumulate during seed filling and store nutrients for seed germination and seedling growth, comprise one of the most important protein categories in plant seeds. Due to their high abundance in nature and their economic importance as a major source of dietary protein, detailed studies of SSPs date to the early part of the 20th century (Osborne, 1924), when Osborne classified them according to their solubility in water (albumin), neutral saline (globulin), alcohol/water mixtures (prolamin), and acids or alkalis (glutelin). The most widely distributed and prevalent SSP group is globulin, which can be divided based on the sedimentation rate of their aggregated forms into the 7S vicilins and 11/12S legumins.
(Shewry et al., 1995). In our survey of the most abundant cotton seed proteins, nearly all of the proteins identified belong to the vicilin and legumin families, comprising 60-70% of the total seed proteins in abundance and suggesting that vicilins and legumins are the major component of mature cotton seeds, as well as the major cotton SSPs.

Quantification of the SSPs, made possible by 2-DE technology, permitted the precise estimation of each SSP category in cotton seeds. In agreement with prior research, which characterized the two principal forms of vicilin as occurring at 48 kDa and 52 kDa (Dure & Chlan, 1981; Dure & Galau, 1981; Dure et al., 1981; Chlan et al., 1986), these vicilin isoforms were also observed as the most abundant proteins on our proteomic maps. Their relative abundances (37% in AD-genome, 36% in A-genome and 28% in D-genome), however, were a little higher than the previous estimate of 27% by cylindrical SDS-PAGE (Dure & Chlan, 1981). In addition to these vicilin isoforms, the overall composition of vicilins and legumins was also estimated, together with a water-soluble fraction of SSPs termed as albumin in prior research, which suggested that each of these three SSP categories may account for up to 1/3 of the total protein amount in cotton seeds (Youle & Huang, 1979; Youle & Huang, 1981). Subsequent research, which characterized the albumin mRNA, noted that not only does albumin encode a low molecular weight protein of only 139 amino acids, but the albumin mRNA makes up a much smaller proportion of the total mRNA pool (2%) in developing seeds, when compared to vicilins (15%) and legumins (30%) (Hughes & Galau, 1989; Galau et al., 1992). It is not surprising, then, that this protein was not detected in our proteomic analyses, which encompass approximately 80% of total seed protein abundance. The discrepancy between previous protein quantifications and the current analysis is likely due to the more ambiguous
classification of globulin and albumin in early studies, which were based on protein solubility and sedimentation rates instead of actual sequences. Although intact globulins are mostly insoluble in water, their degraded or cleaved forms can gain higher water-solubility and display a molecular weight similar to that of albumin; therefore, these albumin-like globulin forms could contribute to overestimation of the amount of albumin in cotton seeds. Because mature albumins are typically cleaved into smaller polypeptides which fall outside of the effective separation range of SDS-PAGE, we cannot rule out the possibility that the poor representation of albumin in the present protein profiles may be due to a technical limitation of 2-DE profiling; however, other estimates of protein abundances, which rely on amino acid composition, concur with our assessment. That is, cotton seeds have been characterized as deficient in sulfur-containing amino acids, indicating that sulfur-rich proteins (such as albumin) constitute a low fraction of the total seed proteins (Bressani et al., 1966; Chlan et al., 1986; Galau et al., 1991; Galau et al., 1992).

**Biased accumulation of D-genome proteins in polyploid cotton seeds**

Allopolyploidization involves the merger of two different, and often divergent, genomes, whose reconciliation in a common nucleus often leads to myriad changes, including unequal integration and expression of the two merging genomes. Recent studies into the consequences of allopolyploidization have underscored this possibility of nonequivalence by demonstrating biased expression among homoeologs and a phenomenon termed transcriptional genome dominance (Rapp et al., 2009). In F\textsubscript{1} hybrids between the allotetraploids *Arabidopsis thaliana* and *A. arenosa*, an analysis of nonadditively expressed genes revealed that, for those genes more highly expressed in *A. thaliana*, the F\textsubscript{1} allotetraploid hybrid preferentially exhibited repressed
expression, much like that in *A. arenosa* (Wang *et al.*, 2006). In the recently formed natural allotetraploid *Tragopogon miscellus*, higher levels of expression have been reported for homoeologs originating in *T. dubius* versus those originating in *T. pratensis* (Buggs *et al.*, 2010). Similar studies in allopolyploid wheat have also demonstrated nonequivalent expression patterns among homoeologs (Bottley *et al.*, 2006; Bottley *et al.*, 2008; Bottley & Koebner, 2008; Pumphrey *et al.*, 2009). In cotton, biased expression of D-genome homoeologs has been reported for petal and fiber tissues in *G. hirsutum* (Flagel *et al.*, 2008; Hovav *et al.*, 2008). These data were later extended to a synthetic F₁ hybrid and the other four natural allotetraploid species that originated from the same genomic merger and doubling, an analysis that suggested that the D genome homoeolog bias was established during genome merger and was subsequently retained during the divergence of all five extant allopolyploid species (Flagel & Wendel, 2010). This observation was later augmented by the discovery of the phenomenon of transcriptional genomic dominance, whereby gene expression levels in a nascent (synthetic) AD-genome allopolyploid mimicked those in the parental D genome more often than those of the A genome (Rapp *et al.*, 2009). Because this was true both for genes that were up- and down-regulated in D relative to A, we termed this phenomenon genomic dominance (in this case biased toward D).

A natural extension of these transcriptional characterizations regarding the nonequivalence accompanying polyploidy is to ask whether similar patterns are exhibited at the protein level, and whether any observed nonequivalencies are linked to phenotypic and functional variations. Attempts to address these questions were first made in allopolyploid *Brassica* using a neosynthesized tetraploid (Albertin *et al.*, 2006; Albertin *et al.*, 2007), where little qualitative variation (less than 1% deviation in spot presence/absence) was observed between the neosynthesized *B. napus*
allotetraploid and its diploid progenitors, *B. oleracea* and *B. rapa*. For the 25%-38% of spots displaying quantitative difference (i.e., those expressed non-additively), expression patterns were slightly closer to that of *B. rapa* parent rather than *B. oleracea*, in accordance with a previous study that suggested bias toward the *B. rapa* genome in the transcriptional expression of rRNA genes (Chen & Pikaard, 1997). The lack of genomic and transcriptomic data, however, makes it difficult to infer the structural and functional significance of these observations in *Brassica*. In the present work, we profiled total mature seed proteins in a naturally formed allotetraploid whose genomic and transcriptomic reactions to genomic merger and doubling have been extensively studied (Adams et al., 2003; Senchina et al., 2003; Grover et al., 2004; Grover et al., 2007; Flagel et al., 2008; Hovav et al., 2008; Chaudhary et al., 2009; Flagel et al., 2009; Rapp et al., 2009; Flagel & Wendel, 2010; Salmon et al., 2010), and address the question whether the D-genome bias and dominance previously observed in the transcriptome of polyploid cotton is reflected at the protein level.

Consistent with transcriptomic data suggesting D-genome dominance (Rapp et al., 2009), the proteome of the allopolyploid was more similar to the D-genome parent, with 26.2% of the 2-DE spots detected in the allopolyploid being present only in the progenitor D-genome versus 10.7% which were present only in the A-genome diploid (Table 1). This observation was extended by quantitative profiling of shared spots, which displayed a higher level of non-additive expression equivalent to that of the progenitor D-genome than to that of the A-genome (14.1% “D-genome dominant” versus 9.8% “A-genome dominant”; Table 2). Hence, the cotton seed proteome displays an overall dominance reflecting its D-genome component; however, by parsing the qualitative and quantitative expressions for each individual SSP,
additional patterns of diversity in dominance become evident, including dominant 
expression by both progenitor genomes, particularly with respect to multiple isoforms 
corresponding to each SSP (Table 4). Similar patterns of differential and 
uncoordinated expression of protein isoforms was also demonstrated in the synthetic 
allopolyploid *B. napus* (Albertin *et al.*, 2007).

Biased expression of D-genome derived homoeologs, another remarkable 
transcriptional feature of allopolyploid *Gossypium* (Flagel *et al.*, 2008; Hovav *et al.*, 
2008; Flagel *et al.*, 2009; Flagel & Wendel, 2010), was also observed in the biased 
accumulation of cottonseed proteins expressed by D-genome derived homoeologs (e. 
g., the vicilin-like protein as shown in Figure 5). Biased expression is generally more 
difficult to uncover in complex protein data than in transcriptomic sequence data and 
is most easily recognized by the absence of A- or D- genome derived homoeologous 
copies (e. g., for legumins). Genetic changes at the protein level accumulate slowly 
relative to the scale of diversification among cotton lineages, such that exons differ by 
only about 1%, mostly at synonymous sites, among A- and D-genome orthologs 
(Senchina *et al.*, 2003), making it difficult to identify and distinguish homoeologous 
protein isoforms. Thus, the biased expression reported here is likely a significant 
underestimate of biased protein expression in the polyploid, as it relies on the limited 
cases of homoeolog loss that are most readily detected.

**Unexpectedly high variation among *Gossypium* seed proteomes**

Comparative proteomics, which permits the characterization of protein level 
variation among related species, is in its infancy as an evolutionary approach, 
particularly with respect to polyploid species. The few studies that have applied 
modern proteomic techniques in a comparative fashion have found little variation
among species (Albertin et al., 2005; Albertin et al., 2006; Albertin et al., 2007). In contrast to the expectations implied by this work as well as the high level of coding sequence conservation among the species studied, the cotton seed proteomes analyzed here display extraordinary variation. This variation not only occurs between diploid species, but also between the allopolyploid species and its model diploid progenitors; only one third of protein features were common to the three *Gossypium* species profiled. Although amino acid sequence variation of SSPs can in principle account for some of this variation, the fact that there is only ~1% synonymous nucleotide differences among orthologous *Gossypium* exons, both in previous studies (Senchina et al., 2003) and for the SSP genes analyzed here, indicate that amino acid substitutions account for only a very small part of the variation detected. This exaggerated interspecific expression variation observed in the *Gossypium* seed proteomes, when compared to similar research in *Brassica*, which revealed 15% divergence between diploid species and only 1% between the synthesized allotetraploid *B. napus* and its diploid progenitors (Albertin et al., 2006), may reflect a gradual accumulation pattern of differential protein expressions in allopolyploid cotton naturally formed 1-2 million years ago versus newly synthesized *B. napus*, as well as, at least in part, differences in the tissues examined. That is, the stem and root proteomes studied in *Brassica* are likely more complex with respect to their proteomes than are the seed proteomes studied here, which tend to be composed of fewer protein types that are extensively modified into many isoforms (Hajduch et al., 2005; Higashi et al., 2006; Bourgeois et al., 2009; Larre et al., 2010). Thus, relatively few underlying differences in post-translational modification programs among cotton species may propagate to affect multiple isoforms, in the process generating a relatively large impact on inferences of similarity, at least in comparisons of seed
versus stem or root proteomes. In addition to this speculation, it may be that the magnitude of proteomic variation is dependent not only on tissue type, but also on ploidy level (Feldman et al., 1986). Classic isozyme analyses, which is able to detect variable protein expression in the form of inferred gene losses or silencing, were previously applied to analyze homoeologous expression patterns in polyploids (Wendel, 2000), and suggested higher variability in expression of seed storage proteins versus other classes of genes in allopolyploid wheat (Galili & Feldman, 1983; Feldman et al., 1986). Furthermore, some of the differential expression patterns observed between diploid and polyploid wheat were inferred to result from proteomic interactions between the contributing genomes (Islam et al., 2003). As the very nature of allopolyploid species involves the co-existence of homoeologous genomes, which itself often involves conflict or competition between regulatory machineries that independently evolved in progenitor species, one can readily envision that the merger of diverged regulatory and post-translational machineries will lead to vastly enhanced combinatorial complexity, which in turn is detected in studies such as ours as “novel” spot presence/absences and transgressive expression levels.

Conclusions

This work presents the first high-quality proteomic map for mature seeds in cotton, a vital oil and meal seed crop. In total, 155 SSP spots and 5 non-storage protein spots were identified. In addition to this comprehensive characterization of protein composition, proteomic profiles were generated, revealing a pattern of interspecific complexity and non-additive protein accumulation in cotton allopolyploids. The biased accumulation of seed proteins toward the D-genome progenitor, combined with the genetic analyses presented here, provides a novel
perspective on the proteomic consequences of polyploidization. One caveat to our study is that we included only one accession of allopolyploid *Gossypium* and its diploid progenitors; therefore, some of the proteomic variation observed might reflect choice of accession rather than between species. The accessions studied were selected as the most widely used models for exploring the genomic and transcriptomic consequences of polyploidy, thereby providing additional among-study perspective. Further exploration into comparative proteomics, including the analysis of additional accessions, will be necessary to identify and characterize the regulatory mechanisms involved in generating the proteomic complexity and novelty observed in these and other species. It also will be interesting to explore the relationships among tissue choice, ploidy level, and multiple experimental variables in developing an enhanced understanding of the effects of hybridization and genome doubling on the proteome of higher plants.

**Acknowledgements**

We thank Kara Grupp and Anna Krush for their help in tissue collection and Corrinne Grover for helpful discussions and comments on the manuscript. We also thank Cotton Incorporated and the National Science Foundation Plant Genome Program for financial support. Dharminder Pathak was supported by a training grant from Punjab Agricultural University, Ludhiana, India.

**References**


Larre C, Penninck S, Bouchet B, Lollier V, Tranquet O, Denery-Papini S, 


Figures and tables

Fig. 1 Evolutionary history of *Gossypium* species.

A. Phylogenetic framework of diploid and allopolyploid *Gossypium*, illustrating the divergence of the A- and D-genome diploids and their polyploidization leading to the evolution of the AD-genome allopolyploid cottons.

B. Morphology of mature cottonseeds. Key to species: AD = *G. hirsutum* var. Acala Maxxa; A = *G. herbaceum* (A1-73); D = *G. raimondii*. Shown are seeds with (above) and without lint (below).
**Fig. 2** Proteomic profile of *G. hirsutum* seeds.

A. Experimental design for two-dimensional gel electrophoresis using two pH ranges, 3 to 10 and 4 to 7.

B. Synthetic proteome map of *G. hirsutum* constructed with images from pH 4 to 7 (left) and pH 3 to 10 (pH 7-10 region only; right). Identified protein spots are indicated as: vicilin A - circle; vicilin B - rectangle; vicilin-like - parallelogram; legumin A – diamond; legumin B – triangle; others - cross.
Fig.3  Differential expression patterns of cotton seed proteomes.

Partial 2-DE gels (pH 5 to 6, molecular weight 26 to 17 kD) of AD-genome, and diploid A- and D-genome seed proteomes. Spots shared by all species and having consistent expression levels are shown as black circles. Variation in either expression level or presence is indicated by color, where red denotes expression in the allopolyploid and blue and gold represent expression levels in the A- and D-genome diploids, respectively. Circle sizes corresponded to spot volumes.
Fig. 4 Representative 2-DE gels illustrating additive and non-additive quantitative expression patterns.

Proteins 621, 616, 627 and 1437 display additive patterns. Proteins deviating from statistical additivity (see text for details) were further categorized as follows: 626 & 636 – D-dominant; 1345 – A-dominant; 657 – transgressive expression above both diploids; 1321 - transgressive expression lower than both diploids. These proteins were identified by mass spectrometry as vicilin A (657, 1321), vicilin B (1437), and legumin A (621, 616, 627, 636, 1345).
Fig. 5 2-DE gels of Vicilin-like isoforms.

Vicilin-like isoforms are indicated by ovals.
**Fig. 6** Neighbor-joining trees of SSP genes in *Gossypium*

Numbers of total and non-synonymous (in parenthesis) nucleotide substitutions are indicated on branches. Those distinguishing the (A, A_T) from (D, D_T) clades are unpolarized and hence are shown at the root of each tree. A_T and D_T refer to homoeologous copies in the allopolyploid genome. The symbol * indicates a sequence with a stop codon.
**Fig. 7** Distribution and polypeptides of vicilin isoforms in *Gossypium*

The symbol * indicates a glycosylation site.

<table>
<thead>
<tr>
<th>Isoform types in Genome(s)</th>
<th>[VicilinA]</th>
<th>Predicted</th>
<th>On-gel</th>
</tr>
</thead>
<tbody>
<tr>
<td>a  (A_T, D_T, A, D)</td>
<td>162 - 584</td>
<td>48.7 - 8.4</td>
<td>6-8.5</td>
</tr>
<tr>
<td>b  (A, D)</td>
<td>162 - 464</td>
<td>35.0 - 8.7</td>
<td>28 - 8.1</td>
</tr>
<tr>
<td>c  (A_T, A)</td>
<td>10.6 - 584</td>
<td>10.6 - 10.0</td>
<td>5-6</td>
</tr>
<tr>
<td>d  (A_T, D_T, A)</td>
<td>250 - 396</td>
<td>16.6 - 10.1</td>
<td>AD: 9-10 A: 7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>[VicilinB]</th>
<th>Predicted</th>
<th>On-gel</th>
</tr>
</thead>
<tbody>
<tr>
<td>e  (A_T, D_T, A, D)</td>
<td>158 - 582</td>
<td>49.4 - 6.4</td>
</tr>
<tr>
<td>f  (D_T)</td>
<td>26 - 157</td>
<td>16.8 - 7.6</td>
</tr>
</tbody>
</table>
Table 1

Qualitative comparison of seed proteomes of allopolyploid *G. hirsutum* (AD-genome) and its parental diploid A- and D-genomes

<table>
<thead>
<tr>
<th>Pattern of qualitative expression</th>
<th>No.</th>
<th>A-genome</th>
<th>D-genome</th>
<th>AD-genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shared spots in all three genomes</td>
<td>92</td>
<td>36.8%</td>
<td>32.2%</td>
<td>29.0%</td>
</tr>
<tr>
<td>Genome-specific spots</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-specific</td>
<td>78</td>
<td>31.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D-specific</td>
<td>65</td>
<td></td>
<td>22.7%</td>
<td></td>
</tr>
<tr>
<td>AD-specific</td>
<td>108</td>
<td></td>
<td></td>
<td>34.1%</td>
</tr>
<tr>
<td>Spots found in two genomes A and D</td>
<td>46</td>
<td>18.4%</td>
<td>16.1%</td>
<td></td>
</tr>
<tr>
<td>A and AD</td>
<td>34</td>
<td>13.6%</td>
<td></td>
<td>10.7%</td>
</tr>
<tr>
<td>D and AD</td>
<td>83</td>
<td></td>
<td>29.0%</td>
<td>26.2%</td>
</tr>
<tr>
<td>Total no. of spots</td>
<td></td>
<td><strong>250</strong></td>
<td><strong>286</strong></td>
<td><strong>317</strong></td>
</tr>
</tbody>
</table>
Table 2
Quantitative analysis of protein additivity of shared spots in allopolyploid

*Gossypium* seed proteomes

<table>
<thead>
<tr>
<th>Pattern of shared spots</th>
<th>No.</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive</td>
<td>59</td>
<td>64.1%</td>
</tr>
<tr>
<td>Non-additive</td>
<td>33</td>
<td>35.9%</td>
</tr>
<tr>
<td>Higher than both diploids</td>
<td>6</td>
<td>6.5%</td>
</tr>
<tr>
<td>A-genome dominant</td>
<td>9</td>
<td>9.8%</td>
</tr>
<tr>
<td>Co-dominant</td>
<td>3</td>
<td>3.3%</td>
</tr>
<tr>
<td>D-genome dominant</td>
<td>13</td>
<td>14.1%</td>
</tr>
<tr>
<td>Lower than both diploids</td>
<td>2</td>
<td>2.2%</td>
</tr>
<tr>
<td>Total no. of spots</td>
<td><strong>92</strong></td>
<td></td>
</tr>
</tbody>
</table>
Table 3

Composition of cotton seed proteomes (percent abundance ± standard error)

<table>
<thead>
<tr>
<th>Type</th>
<th>AD-genome</th>
<th>A-genome</th>
<th>D-genome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>G. hirsutum</em></td>
<td><em>G. herbaceum</em></td>
<td><em>G. raimondii</em></td>
</tr>
<tr>
<td>Vicilin A</td>
<td>19.14 ± 1.53</td>
<td>13.06 ± 2.48</td>
<td>10.41 ± 1.95</td>
</tr>
<tr>
<td>Vicilin B</td>
<td>23.20 ± 3.89</td>
<td>24.66 ± 1.64</td>
<td>18.03 ± 2.04</td>
</tr>
<tr>
<td>Legumin A</td>
<td>18.20 ± 2.27</td>
<td>25.78 ± 4.90</td>
<td>32.94 ± 2.75</td>
</tr>
<tr>
<td>Legumin B</td>
<td>10.34 ± 1.14</td>
<td>5.95 ± 0.75</td>
<td>10.06 ± 0.87</td>
</tr>
<tr>
<td>Vicilin-like</td>
<td>0.20 ± 0.09</td>
<td>0</td>
<td>0.13 ± 0.04</td>
</tr>
<tr>
<td>Non-SSP</td>
<td>0.63 ± 0.19</td>
<td>0.36 ± 0.04</td>
<td>0.79 ± 0.21</td>
</tr>
<tr>
<td>Unknown n*</td>
<td>7.26 ± 1.54</td>
<td>4.73 ± 1.13</td>
<td>6.48 ± 0.86</td>
</tr>
</tbody>
</table>

Total 78.96 ± 2.00 74.54 ± 2.43 78.85 ± 1.51

*Spots were selected for protein identification, but no matched peptides were retrieved from databases.
Table 4

Multiplicity of expression patterns displayed by SSP isoforms

<table>
<thead>
<tr>
<th>Genome specific</th>
<th>Found in two genomes</th>
<th>Shared by all three genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome specific</td>
<td>AD  A  D</td>
<td>AD-A  AD-D  A-D</td>
</tr>
<tr>
<td>SSP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vicilin A</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Vicilin B</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Vicilin-like</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Legumin A</td>
<td>6  8  1  1  5</td>
<td>13  3  3  1</td>
</tr>
<tr>
<td>Legumin B</td>
<td>1  3</td>
<td>1  1  1</td>
</tr>
</tbody>
</table>

*expression dominance
Supporting information

**Figure S1** 2-DE gel images of mature cotton seed proteins

<table>
<thead>
<tr>
<th></th>
<th>pH range 3-10</th>
<th></th>
<th>pH range 4-7</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AD-genome G. hirsutum</td>
<td>![Image]</td>
<td>AD-genome G. hirsutum</td>
<td>![Image]</td>
<td></td>
</tr>
<tr>
<td>D-genome G. Raimondii</td>
<td>![Image]</td>
<td>D-genome G. Raimondii</td>
<td>![Image]</td>
<td></td>
</tr>
</tbody>
</table>
Table S1  Primer sequences used for genetic analysis

<table>
<thead>
<tr>
<th>Gene</th>
<th>Forward primer (5’-3’)</th>
<th>Reverse primer (5’-3’)</th>
<th>Internal primer (5’-3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VicilinA</td>
<td>CGG AGC AAG ACA</td>
<td>CCC TCC TCT CAT CAA</td>
<td>TTA GGG GCA TCA</td>
</tr>
<tr>
<td></td>
<td>AGT GTG AA</td>
<td>ATC CA</td>
<td>ACG AGT TC</td>
</tr>
<tr>
<td>VicilinB</td>
<td>CCG AAG AGG TAC</td>
<td>CTC TTG TGG GTT GCT</td>
<td>ATT TCA GGG TCC TCC</td>
</tr>
<tr>
<td></td>
<td>GAG GAC TG</td>
<td>GTT GA</td>
<td>AAA GG</td>
</tr>
<tr>
<td>LeguminA</td>
<td>CAG ACC TTT TCG TCA</td>
<td>CCG CCT TGT AAG ACC</td>
<td>CAG GAC CAG CAC</td>
</tr>
<tr>
<td></td>
<td>CAG CA</td>
<td>CTT TC</td>
<td>CAG AAA GT</td>
</tr>
<tr>
<td>LeguminB</td>
<td>ATG GCT GCT GTG CTC</td>
<td>CTT GGG CTA AAA ACC</td>
<td>CCA CAA AGA GAT</td>
</tr>
<tr>
<td></td>
<td>AGA TA</td>
<td>GAC AC</td>
<td>CAG CAC CA</td>
</tr>
</tbody>
</table>
Large supplemental files are available for download at

http://www.genetics.org/content/suppl/2011/09/07/genetics.111.132407.DC1

**File S1** 2-DE spot profiles including spot matching and relative volumes from twenty-four 2-DE gels

**File S2** Spot selection and protein identification

**File S3** Gene sequences and alignment of *Gossypium* SSPs

**File S4** Alignment of legumin isoform peptide sequences
CHAPTER 3

PROTEOMIC PROFILING OF DEVELOPING COTTON FIBERS FROM WILD AND DOMESTICATED GOSSYPIUM BARBADENSE

A paper published in New Phytologist in 2013 (New Phytologist 200: 2)

Guanjing Hu, Jin Koh, Mi-Jeong Yoo, Kara Grupp, Sixue Chen, and Jonathan F. Wendel

Abstract

Pima cotton (Gossypium barbadense L.) is widely cultivated because of its long, strong seed trichomes (“fibers”) used for premium textiles. These agronomically advanced fibers were derived following domestication and thousands of years of human-mediated crop improvement. To gain insight into fiber development and evolution, we conducted comparative proteomic and transcriptomic profiling of developing fiber from an elite cultivar and a wild accession. Analyses using iTRAQ LC-MS/MS technology identified 1317 proteins in fiber. Of these, 205 were differentially expressed across developmental stages, and 190 showed differential expression between wild and cultivated forms, 14.4% of the proteome sampled. Human selection may have shifted the timing of developmental modules, such that some occur earlier in domesticated than in wild cotton. A novel approach was used to detect possible biased expression of homoeologous copies of proteins. Results indicate a significant partitioning of duplicate gene expression at the protein level, but an approximately equal level of bias for each of the two constituent genomes of allopolyploid cotton. Our results demonstrate the power
of complementary transcriptomic and proteomic approaches for the study of the domestication process. They also provide a rich database for mining for functional analyses of cotton improvement or evolution.

**Introduction**

Primarily grown for its highly elongated, unicellular seed epidermal trichomes, cotton is the world’s largest source of renewable nature textile fiber. Two major forms of cultivated cotton, *Gossypium hirsutum* L. (Upland cotton) and *G. barbadense* L. (Pima or Egyptian cotton), account for approximately 99% of the world’s cotton production. *Gossypium barbadense*, which comprises a relatively small proportion of US plantings (4%), carries a price premium of 50-80% more than *G. hirsutum* fiber, due to its superior fiber properties (longer staple length and higher strength) that can generate stronger and softer threads, yarns, and fabrics ([http://www.cotton.org](http://www.cotton.org)).

Cotton plants belong to *Gossypium*, which includes ~45 diploids that represent approximately 10 million years of evolutionary divergence and collectively encompass extraordinary morphological variability, geographic distribution and life history variation (Wendel & Cronn, 2003; Wendel *et al.*, 2012). *Gossypium* is noteworthy for the polyploidy event that occurred 1-2 million years ago, giving birth to five allopolyploid species carrying two genomes, an A-genome from Africa or Asia, and a D-genome similar to that found in the New World, primarily Mexican diploids. Among the allopolyploids, *G. hirsutum* and *G. barbadense* were independently domesticated approximately 5,000 years ago in the Yucatan Peninsula and the intermontane Peruvian Andes areas, respectively (Wendel & Cronn, 2003; Wendel *et al.*, 2012). During
thousands of years of human-mediated selection and agronomic improvement, both species underwent many phenotypic modifications, including a shift to more compact plant architecture, establishment of annualized growth habit and photoperiod independence, and reduction in seed dormancy. The most notable changes, however, are in the seed trichomes, including enhanced fineness and length in fibers from the modern crop. Trichome growth curves for wild and cultivated forms indicate the latter have increased fiber growth rate during primary wall synthesis and a prolonged fiber elongation period (Applequist et al., 2001).

Notwithstanding progress in understanding the genetic basis of morphological change in crop plants (Doebley et al., 2006; Burke et al., 2007; Burger et al., 2008; Gross & Olsen, 2010; Gross & Strasburg, 2010; Olsen & Wendel, 2013), little is known about the alterations that mediate the dramatic transformations observed between wild and domesticated cotton. Comparative expression profiling in developing fiber cells from wild and domesticated cotton provided insights into a key metabolic transformation associated with prolonged fiber elongation in domesticated cotton, namely, the modulation of reactive oxygen species that control cellular redox levels (Hovav et al., 2008; Chaudhary et al., 2009). In Upland cotton, the fiber transcriptome has been dramatically rewired by domestication and crop improvement: nearly a quarter of the genes in the genome were differentially expressed, suggesting that the phenotypic changes between fibers from wild and domesticated cotton reflect a high level of underlying genetic complexity (Rapp et al., 2010).

In addition to studies of the genome and transcriptome, proteomic investigations may provide important perspectives on evolutionary transformations, although at present
there are few high-throughput or genome-scale proteomic evolutionary studies in plants. The promise of proteomics lies, at least partially, in the realization that proteins are the direct executers for most cellular activities, e.g., the physiological and biochemical reactions that link phenotypes to genotypes (Karr, 2008; Diz et al., 2012). With respect to the morphological and molecular changes observed in cotton domestication, it is natural to ask how the proteome has responded to human-mediated selection, thereby extending our understanding across distinct “omics” levels. Here, we demonstrate this approach through analyzing the fiber proteomes of wild and domesticated *G. barbadense* at four developmental time points with an advanced iTRAQ (isobaric tag for relative and absolute quantification) technology coupled with LC-MS/MS. Our results revealed a global shift of protein expression patterns corresponding to the domestication process, entailing expression changes of many candidate proteins and metabolic processes. We also conducted co-analysis using transcriptomic data for fiber elongation, demonstrating the different and hence complementary nature of gene and protein expression profiles for tackling complex evo-devo problems.

**Materials and methods**

**Plant materials, tissue collection and protein extraction**

The elite cultivar Pima S-7 and a wild accession K101 from Bolivia were chosen to represent domesticated and wild accessions of *G. barbadense*, respectively (Figure 1). Plants were grown in the Bessey Hall Greenhouse at Iowa State University. Flowers were tagged at anthesis and harvested at four key developmental stages, i.e., 5, 10, 20 and 25 days post anthesis (dpa), representing primary wall synthesis (5) and elongation (10), and
the transition to (20) and continuation (25) of secondary wall synthesis. Harvested cotton bolls were dissected and ovules were frozen in liquid nitrogen and stored at -80°C. For each developmental stage, we used three biological replicates. For each replicate, 2 g of ovules were pooled from five plants to account for variation among individuals, and were thereafter subjected to protein extraction. Cotton fiber proteins were isolated and purified as described (Yao et al., 2006) with the following modifications. Frozen ovules together with 10% [w/w] glass beads and 10% [w/w] PVPP were vortexed four to five times in liquid nitrogen (30 min.), and suspended in 5 ml Tris-saturated phenol and 5 mL extraction buffer (50 mM Tris-HCl, pH 8.8, 30% [w/v] sucrose, 2% [w/v] SDS, and 2% [v/v] 2-mercaptoethanol). The use of glass beads was adapted to separate fibers from ovules without tissue contamination from seeds (Taliercio & Boykin, 2007). Following phenol extraction, ammonium acetate precipitation and acetone washing (Koh et al., 2012), the protein pellet was dissolved in protein buffer (8M urea, 25mM TEAB, 2% [v/v] TX-100, 0.1% SDS [w/v], pH 8.5) at room temperature, and centrifuged at 20,000 × g for 20 min to remove insoluble materials. The supernatant was washed again with cold 80% acetone thrice and solubilized in the protein buffer. Protein assays were performed using an EZQ® Protein Quantitation Kit (Invitrogen, Carlsbad, CA, USA).

iTRAQ labeling, strong cation exchange and liquid chromatography-tandem mass spectrometry

For each sample, 100 µg of protein was reduced, alkylated, and trypsin-digested using the iTRAQ Reagents 8-plex Kit according to the manufacturer’s instructions (AB Sciex, Inc., Foster City, CA, USA). The developmental stages of 5, 10, 20 and 25 dpa of
Pima S-7 were labeled with iTRAQ tags 113, 114, 115, and 116, and those of K101 were labeled with 117, 118, 119, and 121, respectively. The combined peptide mixtures were lyophilized, dissolved in strong cation exchange (SCX) solvent A (25% [v/v] acetonitrile, 10 mM ammonium formate, and 0.1% [v/v] formic acid, pH 2.8), and fractionated using an Agilent HPLC system 1260 with a polysulfoethyl A column (2.1×100 mm, 5 µm, 300 Å; PolyLC, Columbia, MD, USA; Supporting Information Figure S1). Peptides were eluted with a linear gradient of 0–20% solvent B (25% [v/v] acetonitrile and 500 mM ammonium formate, pH 6.8) over 50 min followed by ramping up to 100% solvent B in 5 min and holding for 10 min. Twelve fractions were collected by monitoring the absorbance at 280 nm and area size of each fraction was calculated for % coefficient of variation among three biological replicates (Figure S1).

Tryptic peptides were loaded into a C18 capillary trap cartridge (Dionex, San Francisco, CA, USA) and separated with a LC Packing C18 Pep Map HPLC column (Dionex, San Francisco, CA, USA). A hybrid quadrupole time-of-flight QSTAR Elite MS/MS system (AB Sciex, Inc., Foster City, CA, USA) was used for data acquisition as described previously (Zhu et al., 2012).

Protein database search and analysis of differential protein expression

For comprehensive protein identification, we constructed a non-redundant Gossypium protein database (122,785 entries) using the recently sequenced genome (Paterson et al., 2012) of the diploid D-genome species G. raimondii and a cotton SNP index (Page et al., 2013) generated between G. raimondii and the A-genome diploid G. arboreum. These data were used to infer protein sequences of both diploid species for database construction. The MS/MS data were processed by a thorough database search
considering biological modifications and amino acid substitutions under the Paragon™ algorithm (Shilov et al., 2007) and the Pro Group™ algorithm, using ProteinPilot version 4.5 software (AB Sciex, Inc., Foster City, CA, USA). To examine homoeolog-specific expression, i.e., distinguishing the expression patterns of each homoeolog, the MS/MS data were subsequently searched against the separate D-genome G. raimondii (77,267 entries) and A-genome G. arboreum databases (65,170 entries). Search parameters included iTRAQ 8-plex quantification, cysteine modified with methyl methanethiosulfonate, trypsin digestion, thorough searching mode and variable modifications for known post-translational modifications (PTMs, http://www.abrf.org/index.cfm/dm.home). The confidence level of protein identifications was set to 95%, reflecting a less than 5% local false positive identification rate (FDR).

The global FDR of identified protein lists was determined by performing searches against the reversed protein databases, with estimates derived from both the conventional approach and a nonlinear fitting method (Tang et al., 2008) as shown in Supporting Information Table S1.

Bias correction (built-in function of ProteinPilot) was applied to normalize protein quantification across samples. Relative quantification of proteins was performed using the ratios from MS/MS spectra only when the peptide sequences were uniquely assigned to detected proteins. To be identified as being significantly differentially expressed, a protein must have been quantified with at least three spectra in at least two of the biological triplicates, along with a Fisher’s combined probability of <0.05 (Fisher, 1948).
Annotation, classification and expression clustering analysis

In addition to the released gene descriptions derived from the sequenced G. raimondii genome (Paterson et al., 2012), the functional annotation Blast2GO suite (Conesa et al., 2005) was used to annotate identified protein sequences (http://www.blast2go.com/b2ghome). Protein family and subfamily classification was performed using PANTHER (Mi et al., 2010), a database of protein functions inferred from phylogenetic trees of protein families from all kingdoms, thereby associating protein with a simplified but more accurate and complete ontology, compared to general GO annotation. Supported by the Blast2GO suite, InterProScan (Zdobnov & Apweiler, 2001) was launched to search against the PANTHER HMM library that maps protein sequences to PANTHER IDs. The mapping list was uploaded to the PANTHER system (http://www.pantherdb.org), and analyzed using the protein class option. The over- and under- representation of any particular protein class was tested using the binomial test (Cho & Campbell, 2000) with bonferroni correction for multiple comparisons. Proteins with annotations or assigned to known family groups were functionally classified based on the Arabidopsis functional catalogue (Bevan et al., 1998). Hierarchical clustering of protein expressions was performed and visualized on heatmaps in R using the gplots package (http://www.R-project.org), specifying average linkage and Pearson’s correlation distance metric.

RNA-seq analysis and comparison with proteomic data

Wild (K101) and domesticated (Pima S-6, an elite cultivar closely representing Pima S-7 used for proteomics) cotton bolls collected at two developmental stages (10 and
20 dpa) were used for fiber transcriptomic profiling. RNA extraction, purification, RNA-seq library construction and sequencing followed by data analyses including mapping, homoeolog-specific regulation and differential gene expressions were conducted as described (Paterson et al., 2012: supplemental information S5). To examine the degree of concordance between transcript and protein levels, Pearson correlation tests were conducted using expression ratios of 20 versus 10 dpa in both accessions and Pima S-6 versus K101 at both time points, respectively.

Results

Identification of *G. barbadense* fiber proteins

Using our *Gossypium* protein database, a total of 1317 proteins (File S1) were identified at a 95% confidence level and a 1% false discovery rate (Tables S1, S2). These fiber-expressed proteins represent most protein families (Mi et al., 2010) encoded in the *Gossypium* genome (Paterson et al., 2012), except categories that are less likely to be expressed in single-celled fibers (e.g., cell junction protein, cell adhesion molecule, defense/immunity protein). Chaperone (6.9%) and metabolic enzymes including oxidoreductase (13.7%), isomerase (7.0%), lyase (4.6%), ligase (4.4%) and kinase (4.0%) were found to be significantly over-represented, while nucleic acid binding (11.3%), transcription factor (2.0%), transporter (2.6%) and phosphatase (0.6%) proteins were under-represented in the fiber proteomes (Figure S2).
Quantitative proteomic changes during fiber development within wild and domesticated *G. barbadense*

Along with protein identification, our eight-plex iTRAQ experiments enabled simultaneous comparison of protein expression over a developmental course of fiber growth (5, 10, 20, 25 dpa) in Pima S-7 and K101 (File S2, S3). Proteomic changes were first examined between adjacent developmental stages - 5 to 10 dpa, 10 to 20 dpa, and 20 to 25 dpa, which revealed 205 proteins that were significantly differentially expressed within wild or domesticated cotton fibers (19.0% of 1317 proteins). Of these, Pima S-7 displayed a lower level of developmental expressional variation (151 proteins, 11.5%) than did K101 (198 proteins, 15.0%). As shown in Figure 1, the distribution of changes was biased toward the earliest developmental stage in both accessions ($p < 0.05$, Chi-square test): 90 (5 to 10 dpa), 64 (10 to 20 dpa), and 5 (20 to 25 dpa) differentially expressed proteins were identified in Pima S-7, and 126 (5 to 10 dpa), 76 (10 to 20 dpa) and 24 (20 to 25 dpa) proteins were differentially expressed in K101.

To discern the multivariate pattern of up- and down-regulation accompanying fiber development and the domestication process, we built a clustered heatmap of differentially expressed proteins (Figure 2). As demonstrated by the dendrogram at the top of the heatmap, the developmental course from 20 to 25 dpa in wild and domesticated accessions were clustered, both exhibiting little developmental or evolutionary differential expression in this phase of secondary cell wall synthesis. Interestingly, another cluster was formed between 10 to 20 dpa in K101 and the earlier course of 5 to 10 dpa in Pima S-7, instead of between the same time intervals, which suggests a general shift of protein regulation toward earlier fiber elongation in domesticated cotton.
Differentially expressed proteins were clustered according to their developmental changes (Figure 2, left dendrogram). Proteins involved in similar or relevant cellular activities in most functional classes displayed diverse expression patterns (Figure 2, grey columns; File S2). One apparent clustering was found in the class “protein synthesis”, where a collection of ribosomal protein subunits was concordantly down-regulated in Pima S-7 from 5 to 10 dpa.

**Quantitative proteomic changes between wild and domesticated *G. barbadense***

When comparing protein expression levels between Pima S-7 and K101 at each time point, 190 proteins experienced significant up- or down-regulation associated with domestication at one or more developmental stages (File S2). The highest number of differentially expressed proteins occurred during early fiber elongation (5 dpa, 122 proteins; Fig. 1), followed by fewer expression changes later during primary wall synthesis (10 dpa, 76 proteins) and the transition to secondary cell wall synthesis (20 dpa, 64 proteins), with the number increasing slightly during secondary wall synthesis (25 dpa, 104 proteins). The distribution of up-regulation in wild and domesticated accessions is statistically symmetric ($p > 0.05$, Chi-square test), although slightly more proteins were found up-regulated in Pima S-7 than in K101, except at 20 dpa. These proteins were classified into 12 functional categories (Bevan *et al.*, 1998); differential expression patterns within each were hierarchically clustered (Fig. 3; values in File S2). The largest functional class, “defense” (Fig. 3, middle part, central column), was associated with stress responses and detoxification. Four peroxidases were mostly up-regulated at 10 dpa while down-regulated at other stages, suggesting an involvement of reactive oxygen
relevant proteins. A large subclass of heat shock proteins (HSPs), including Hsp10/20, Hsp60, Hsp 70, Hsp83 and Hsp 90, exhibited diverse patterns. In the second largest class, “metabolism” (top, left column), proteins involved in nucleotide, amino acid, lipid, sugar and polysaccharide metabolism exhibited expression patterns that didn’t seem clustered by metabolic functions. The “protein synthesis” class (2nd cluster, left column) was composed largely of ribosomal proteins with increased expression at 5 dpa (13 of 19 ribosomal proteins; \( p < 0.05 \), Fisher’s exact test), whereas three out of five translational initiation and elongation factors were down-regulated at this stage. In the class “protein destination” (3rd cluster, left column), various subunits of T-complex protein 1 involved in protein folding and stabilization were concordantly regulated. For proteins involved in “energy” production (top cluster, central column), up-regulation by domestication was observed for most of this functional category along the developmental trajectory, except at 10 dpa. Among “cytoskeleton” related proteins (2nd cluster, middle column), actin and actin depolymerizing factor were oppositely regulated, and all three annexins decreased in abundance at 20 and 25 dpa. In the class “secondary metabolism” (top right), five enzymes of the flavonoid biosynthesis pathway were up-regulated by domestication throughout development, except at 10 dpa. Other protein classes without obvious patterns of functional clustering included “structure” proteins of cell wall and mitochondria (bottom left), “signal transduction” through G-protein binding, phosphorylation, and other signaling pathways (middle center), “transporter” comprising calcium-binding proteins, ATPases and ATP synthases (middle right), and “intracellular traffic” proteins (bottom center). The unclassified group (bottom right) comprised proteins with unknown or unclassified functions, such as 14-3-3 proteins. Only one transcription factor, a
homolog to *Arabidopsis* transcription factor Pur-alpha 1, was identified, exhibiting up-regulation in Pima S-7 at 25 dpa.

**Integrative Analysis of Proteome and Transcriptome during Fiber Elongation**

Transcriptomic changes in wild and domesticated *G. barbadense* were monitored using RNA-seq data from 10 and 20 dpa fibers (File S3), allowing a direct comparison of transcript and protein expression during cotton fiber elongation. Concordance tests revealed poor correlations ($r_{\text{Pearson}} = 0.06$–$0.24$) between mRNA and protein ratios (Figure 4). When considering significant changes only, 90 and 195 mRNA transcripts were differentially regulated from 10 to 20 dpa in wild and domesticated cotton, respectively, which are higher than the protein numbers (Figure 1). However, fewer genes (19 and 11) were differentially expressed at the mRNA level between the two accessions at the same time point, whereas 73 and 64 significant protein changes were observed at 10 and 20 dpa, respectively. Comparison of these changes revealed little overlap between transcript- and protein-level inferences. Only seventeen cases of concordant expression were evident, i.e., where protein accumulation was significantly correlated with transcript abundance (Figure 4, red dots).

**Homoeolog expression in allopolyploid *G. barbadense***

The analysis presented to this point concerns protein accumulation from both homoeologs ($A_T$ and $D_T$, where the subscript indicates the specific genome in the allopolyploid) from tetraploid cotton (which contains $A_T$ and $D_T$ genomes). To analyze the separate homoeologous contributions to the fiber proteome of allopolyploid *G.*
barbadense, we characterized the expression pattern of protein homoeologs for which we had evidence of genome-of-origin. Based on an average of 1.8% amino acid difference between protein orthologs in the diploid A- and D- genome protein databases, we were able to diagnose homoeolog-specific peptides for 729 proteins, or 55% of the 1317 fiber proteins identified (File S4). Of these, 296 proteins had genome-diagnostic peptides for both homoeologs, but without statistically significant differential expression of homoeologs \( p > 0.05 \). For the remaining 433 proteins, diagnostic peptides were only detected for one of the two homoeologs, which might suggest silencing (Hu et al., 2011), or a lower abundance of homoeologs from the “missing” genome due to the technical limitations of mass spectrometry (MS) to detect low-abundance proteins (Wang et al., 2006). Thus, these 433 proteins represent either biased expression of protein homoeologs or false positives; however, the symmetric distribution of proteins detected with either an A_T or a D_T homoeolog bias (212 vs 221; \( p > 0.05 \), Chi-square test) suggests equivalent detection power for both homoeologs in our analyses. As a more conservative approach, we limited our analysis to only proteins where the same directional homoeolog bias was observed in all replicates. This analysis yielded 57 proteins with robust estimates of homoeolog expression bias, distributed equally between the two constituent genomes, i.e. 27 A-specific and 30 D-specific proteins (Table 1). Gene members in the leucine aminopeptidase, glutathione peroxidase, aspartic proteinase and glutathione s-transferase families exhibited opposite homoeolog-specific expression patterns, while three malate dehydrogenases were all derived from the D_T homoeolog. Using RNA-seq data, homoeolog expression bias at the transcript level, as defined in Grover et al. (2012), was tested for the genes that encode these proteins with biased homoeolog expression. As
shown in Table 1, homoeolog ratios of transcripts were calculated for proteins exhibiting homoeolog bias (e.g., A_T/D_T for A_T-biased peptides, the reverse for D_T biased peptides); ratios over 1.0 suggest the same direction of homoeolog expression bias at both mRNA and protein levels (40 of 57 proteins, \( p < 0.05 \), Fisher’s exact test). Among the 33 proteins detected with significant transcriptional biases (Paterson et al., 2012: supplemental information S5), 27 exhibited the same directional bias at the protein level (Table 1, ratios marked with *), which also indicates a high concordance between gene transcription and translation levels (\( p < 0.05 \), Fisher’s exact test) for the analysis of homoeolog expression and bias.

**Discussion**

Ever since Darwin’s The Origin of Species (Darwin, 1859), it has been well-understood that the domestication of plants and animals offers windows into the evolutionary process and the genetic mechanisms by which traits arise. Over the past 20 years, there has been a concerted effort to identify the specific genes that control morphological transformations between wild crop ancestors and their modern descendants (Doebley et al., 2006; Burke et al., 2007; Burger et al., 2008; Gross & Olsen, 2010; Olsen & Wendel, 2013). As reviewed in Olsen and Wendel (2013), quantitative trait locus (QTL) mapping studies initially localized a handful of regulatory genes with large effects (Doebley et al., 1997; Frary et al., 2000; Doebley, 2004; Wang et al., 2005; Konishi et al., 2006; Li et al., 2006; Simons et al., 2006; Jin et al., 2008), while recent genome-wide analyses including association mapping and large-scale screens for signatures of selection provided evidence for many small-effect genes and a more
complex polygenic control of some domestication traits (Wright et al., 2005; Yamasaki et al., 2005; Chapman et al., 2008; McNally et al., 2009; Tian et al., 2009; Lam et al., 2010; Zhao et al., 2011; Huang et al., 2012; Xu et al., 2012). These data demonstrate that regulatory changes play an important and complementary role to causative mutations in structural genes during phenotypic evolution accompanying domestication. Accordingly, the comparison of transcriptomes of wild and domesticated derivatives, exemplified by work in maize and cotton, recently has provided insights into the astonishing complexity of molecular networks and metabolic pathways altered by the domestication process (Gross & Strasburg, 2010; Rapp et al., 2010; Hufford et al., 2012; Swanson-Wagner et al., 2012). The present study represents an initial effort at extending these types of analyses of the molecular basis of crop domestication to the proteomic level.

Application of proteomics to the study of crop domestication

In cotton, almost all the genes are actively transcribed during fiber development (Hovav et al., 2008; Rapp et al., 2010); Yoo et al., unpublished). Nearly a quarter of the transcriptome comprising 9465 genes, was differentially expressed between wild and domesticated fiber phenotypes in G. hirsutum (Rapp et al., 2010), and a less dramatic alteration of approximately 4200 genes was observed in domesticated G. barbadense compared to its wild form (Chaudhary et al., 2008). Here, comparison of the Pima S-7 (domesticated) and K101 (wild) fiber proteomes in G. barbadense revealed 190 proteins differentially expressed during fiber development (Figure 3), which account for 14.4% of the proteins we profiled. This proteomic change is of the same order of magnitude with that observed at the transcriptional level previously, and expression changes of some
protein groups, such as peroxidases and other stress response proteins, were also evident in the transcriptome data. However, a direct comparison of protein and their corresponding mRNA expression levels revealed poor correlations and few overlapped significant changes (Figure 4). Although two different but closely related modern cultivars of *G. barbadense* (Pima S-6 and Pima S-7) were used in the RNA-seq and iTRAQ analyses, respectively, which could contribute to the poor correlation between transcriptomic and proteomic data for the domesticated form, similar results were observed for wild *G. barbadense*, where the same accession was used in both analyses. This finding is in agreement with previous results from various organisms showing that transcript abundances only partially predict protein abundances, and that a series of regulatory processes involved in translation, localization, modification and protein degradation play a substantial role in controlling protein expression (Vogel & Marcotte, 2012).

When studying regulatory changes that contribute to crop domestication and adaptive evolution, it has been common practice to use gene transcription as proxies for the expression and activity of the corresponding proteins, thereby directly linking gene expression changes to phenotypic variations in response to selection. Transcriptomic studies using this approach have been reported particularly in the cotton model system (Chaudhary *et al.*, 2008; Hovav *et al.*, 2008; Rapp *et al.*, 2010), which led to in-depth investigations of parallel evolution under domestication for reactive oxygen species (ROS) scavenging (Chaudhary *et al.*, 2009) and profilin gene family up-regulation (Bao *et al.*, 2011) during cotton fiber development. It is worth noting that this up-regulation was observed at both the mRNA and protein levels, a result that is in accordance with the
presumptive foundation of the transcriptomic approach. However, the relationship between mRNA and protein abundances is complex, being subject to myriad transcriptional, post-transcriptional, translational, and post-translational determinants and regulations. As measured in mammalian cells, mRNAs are produced at a much slower rate than the rate of protein translation, and on average, protein products were five times more stable and 2800 times more abundant than mRNAs; more importantly and perhaps more surprisingly, protein abundance spanned a higher dynamic range (Schwanhausser et al., 2011). Therefore, expression changes detected at the mRNA level may or may not result in variable protein abundances as controlled by protein turnover, while at the same time expression changes at the protein level may or may not also be observed at the mRNA level.

A striking example of this discordance between transcript and protein regulation seen in our study concerns the continuous up-regulation of the cotton Flowering locus T (FT) protein in Pima S-7 relative to K101 from 10 to 25 dpa (Protein ID 467 in Figure 3, File S2); in contrast to the abundant proteins detected in fibers, FT mRNA transcripts were barely detected in our RNA-seq analyses and to our knowledge have never previously been reported in cotton fibers. Identified as a key regulator of floral transition in plants, FT is mainly transcribed and translated in leaves, and the protein travels as a long distance signal to induce flowering at the shoot apex (Wigge, 2011). The FT protein, also involved in several cell growth processes (Shalit et al., 2009; Kinoshita et al., 2011), interacts with transcriptional factors in various signaling pathways (Mimida et al., 2011). Further analysis is required to understand the function of FT and its increased expression accompanying domestication during cotton fiber development. Nevertheless, it is clear
that this type of protein-level specific alteration, either due to undetectable transcription or possible temporal/spatial separation of mRNA and protein presence, can only be studied by direct measurement of protein abundances.

Another type of expression change that can only be characterized at the protein level is variation arising from post-translational modification (PTM). For example, novel protein isoforms resulting from proteolytic cleavage, oxidation and deamidation were recently reported in *Tragopogon* in response to allopolyploidization (Koh *et al.*, 2012). Taken together, proteomics is not only complementary to transcriptomic screening for regulatory changes, but also has a distinct advantage, illuminating evolutionary processes relevant to protein function, and perhaps providing clues to adaptation and/or the origin of isoforms.

One novel application of proteomics, as applied in the present work, is the differentiation of protein homoeologs and their specific expression patterns in allopolyploid cotton. In contrast to the now widely appreciated transcriptomic concept of homoeolog expression bias (Grover *et al.*, 2012), the phenomenon of unequal homoeolog expression at the protein level has rarely been described (Hu *et al.*, 2011; Koh *et al.*, 2012). Two cases of homoeolog-specific expression were reported in allopolyploid *Tragopogon* proteomes (Koh *et al.*, 2012), where silencing of the maternal homoeolog at the transcript level led to exclusive expression of the paternal proteins in allopolyploid *T. mirus*. In cotton fibers, we were able to diagnose homoeolog-specific expression for 57 proteins, with directional bias equally distributed toward the two parental genomes. Additionally, for these proteins, our observation of high concordance between mRNA
and protein levels suggests that the genesis of homoeolog expression bias reflects regulatory processes that are mainly controlled at the gene transcription level.

**Functional interpretation of cotton fiber proteomes**

Our iTRAQ data obtained from wild and domesticated *G. barbadense* fiber proteins have provided a genome-scale proteomic analysis of fiber development in the evolutionary context of human selection. The resulting proteomic profiling of 1317 proteins demonstrates a clear technological advance of the MS-based approach for protein discovery over the traditional two-dimensional gel electrophoresis (2-DE) method, which previously has been used to identify up to 235 proteins in cotton fibers (Yao *et al.*, 2006; Yang *et al.*, 2008; Zhang *et al.*, 2013). Because of the nature of the MS technology, iTRAQ data acquisition is biased towards high-abundance proteins, especially in complex samples (Liu *et al.*, 2004; Wang *et al.*, 2006), which likely precludes our ability to discover additional proteins. Indeed, the fiber proteins we identified were over-represented by stable and abundant enzymatic proteins, such as oxidoreductases and isomerases.

With the objective of revealing proteomic changes resulting from cotton domestication, a key finding from our study is that the proteome of domesticated Pima cotton during early fiber elongation closely resembles a later developmental stage of wild *G. barbadense* (Figure 2). This systematic shift of protein regulation coincides with an increased fiber elongation rate in domesticated relative to wild cotton, as previously shown by fiber growth curves; that is, the most rapid period of fiber elongation began at about 10 dpa in domesticated *G. hirsutum*, whereas this phase was delayed until about 15
dpa in wild *G. hirsutum* and in another wild allopolyploid, *G. tomentosum* (Applequist *et al.*, 2001). By specifying proteins contributing to this pattern in *G. barbadense*, we identified concordant regulation of ribosomal protein subunits (File S2). With the peak expression at 5 dpa, ribosomal proteins were down-regulated from 5 to 10 dpa in Pima S-7, while their expression in K101 peaked at 10 dpa followed by down-regulation from 10 to 20 dpa (Figure 2). Often used as indicator of cell growth status, the altered expression of ribosomal proteins might imply unknowing human selection for earlier activation of fiber elongation networks. In *Arabidopsis*, ribosomal protein genes are co-regulated in growing axillary shoots and germinating seeds, and the common *cis* elements located in their promoter regions were shown to be promising target sequences to screen for upstream transcription factors regulating rapid developmental processes (Tatematsu *et al.*, 2008). This raises the possibility that similar regulatory regions of cotton ribosomal proteins could be discovered that are co-regulated and play a role in the gene networks governing fiber growth.

Also representing the temporal shift in expression pattern are many stress response proteins that regulate redox homeostasis, whose increased protein abundance during early fiber elongation was not seen until 20 dpa in wild *G. barbadense*, such as the expression curves of ascorbate peroxidase shown in Figure 2. Relevant to this finding is the earlier suggestion that regulation of hydrogen peroxide (H$_2$O$_2$) and other reactive oxygen species (ROS) is a key process in both cotton fiber development and evolution (Hovav *et al.*, 2008). H$_2$O$_2$ and other ROS at appropriate concentrations are required for cell elongation, being involved in the cleaving of polysaccharides during cell-wall relaxation (Fry, 1998; Foreman *et al.*, 2003; Liszkay *et al.*, 2004). They also appear to
serve as developmental signals for the onset of secondary wall differentiation (Potikha et al., 1999), but higher ROS levels may halt elongation through stimulation of cell wall stiffening, and can even promote programmed cell death or necrosis (Schopfer, 1996; Rodriguez et al., 2002). Many genes involved in modulating ROS levels were transcriptionally up-regulated in domesticated accessions of diploid and polyploid cotton species, suggesting parallel selection of this particular regulatory network in separate domestication events (Hovav et al., 2008; Chaudhary et al., 2009). In our data, ROS scavenging proteins were concordantly up-regulated from 5 to 10 dpa in Pima S-7 and from 10 to 20 dpa in K101 (Figure 2, File S2). In contrast to the up-regulation of mRNA levels by domestication throughout the developmental process, higher abundance of peroxidases was only observed at 10 dpa in domesticated cotton, accompanied by higher expression levels in wild cotton at early and later developmental stages (Figure 3). As implicated in a previous proteomic analysis, a G. hirsutum cytosolic ascorbate peroxidase (GhAPX1) functions to detoxify H$_2$O$_2$ produced during fast fiber elongation, as evidenced by the fact that transcript levels and enzymatic activity of GhAPX1, as well as fiber length, can be promoted by exogenous H$_2$O$_2$ (Li et al., 2007). Therefore, it is reasonable to speculate that accumulation of peroxidase was unknowingly targeted by humans so that lower levels of H$_2$O$_2$ were maintained, thereby facilitating fiber elongation. The ROS signaling network is highly dynamic and complex (Mittler et al., 2011; Suzuki et al., 2012), so only a glimpse of its significance with respect to cotton fiber development and evolution is provided here. Further experiments are warranted, focused on integrative comparative analyses of cellular ROS modulation and the genetic architecture of cotton fiber development.
Another possibly related expression pattern of redox homeostasis control revealed here is the up-regulation in domesticated relative to wild cotton at all stages except 10 dpa. This group included phospholipase D alpha (PLDα) (Figure 2), NADP-isocitrate dehydrogenase (NADP-ICDH) and a type III alcohol dehydrogenase (ADH). Activation of PLDα leads to hydrolysis of structural phospholipids into phosphatidic acid (PA) and choline. With both products serving as important signaling molecules, PLDα is involved in various cellular processes, among which PA plays a role in mediating superoxide production in Arabidopsis (Sang et al., 2001). It is possible that the accumulation of PLDα participates in signal transduction for the release of ROS in cotton fiber cells.

NADP-ICDH catalyzes the production of NADPH, which appears to be essential in the mechanism of plant defense against oxidative stress (Leterrier et al., 2012). Increased expression of NADP-ICDH in domesticated cotton except at 10 dpa may suggest elevated antioxidant activity at other developmental stages. ADH is an anaerobic protein that catalyzes the reduction of acetaldehyde to ethanol, resulting in continuous NAD+ regeneration. Its induction by anoxic or hypoxic stresses has been demonstrated in a variety of plants, including cotton (Millar et al., 1994; Millar & Dennis, 1996). The class III ADH we identified has not been characterized in the cotton genome, to our knowledge. Functionally diverged from the classic ethanol-active enzyme types, class III ADH has been implicated to play an essential role in formaldehyde detoxification (Achkor et al., 2003), which also is associated with the ROS detoxification function of the ascorbate-glutathione cycle (Reumann et al., 2007). Also identified in our data, two ethanol-active ADHs were characterized with different expression patterns during fiber
development and a lack of regulatory change between wild and domesticated cotton, which may suggest a unique detoxifying role of the class III ADH in cotton fibers. Overall, the data generated here will serve as an accessible source of clues for functional analyses, be they targeted at crop improvement or evolutionary understanding. For example, proteins more abundantly expressed in Pima S-7 at the later stages of 20 and 25 dpa were often found coupled with decreased expression during K101 fiber development after 10 dpa. The maintenance or up-regulation of these proteins may provide candidate biological processes to interpret the continuous elongation and delayed onset of secondary wall synthesis in domesticated cotton. Three enzymes involved in biosynthesis of polyphenol compounds were identified in this group, i.e., phenylalanine ammonia-lyase (PAL), catalyzing the first committed step in the phenylpropanoid pathway, chalcone isomerase (CHI) and dihydroflavonol 4-reductase (DFR) (Figure 2), the latter two functioning in the biosynthesis of flavonoids. The detection of flavonoid-related transcripts during cotton fiber development has been previously noted, where higher expression was observed during fiber elongation in comparison to secondary wall synthesis and in other ovular cells (Arpat et al., 2004; Gou et al., 2007; Hovav et al., 2008; Al-Ghazi et al., 2009; Rapp et al., 2010). Like many other secondary metabolites, flavonoids are thought to have numerous roles in the interactions of plants with their environments, including protection via the antioxidant activity of hydroxyl groups against diverse biotic and abiotic stresses (Lepiniec et al., 2006). A recent study indicated that silencing of a core flavonoid pathway enzyme flavanone 3-hydroxylase (F3H) as well as introduction of exogenous naringenin (NAR), a substrate of F3H, could significantly retard fiber development (Tan et al., 2013), perhaps linking fiber elongation under
domestication to the ROS signaling discussed above. It was also reported that the products of phenylpropanoid pathway could be deposited to fiber cell wall in the form of wall-linked phenolics (Fan et al., 2009), thereby facilitating secondary wall synthesis. Nevertheless, the regulation of phenylpropanoid and flavonoid pathways is further complicated by the dynamics of fiber developmental changes; that is, coordinate up-regulation of these enzymes also occurs during fiber initiation in domesticated cotton, as shown here and in previous transcriptomic studies (Hovav et al., 2008; Rapp et al., 2010).

Conclusion

The present study represents the first large-scale comparative proteomic analysis of development and the domestication process for cotton, and to our knowledge, for any crop plants. Compared to other analyses using 2-DE methods (Yao et al., 2006; Yang et al., 2008; Zhang et al., 2013), iTRAQ data resolved at least five-fold more fiber proteins, and provided simultaneous protein quantification from all sample conditions with low technical variation. In addition to demonstrating the altered protein expression patterns associated with fiber development and evolution, our study has highlighted the complementary roles of transcriptomic and proteomic views of crop domestication, leading us one step closer toward understanding the morphological and physiological transformations accompanying domestication and crop improvement. We identified a modular development shift in domesticated cotton, and concordant regulation of certain enzymes and biological processes such as redox homeostasis. Collectively these data provide clues into the fundamental regulatory network targeted by aboriginal
domesticators, and will lead to future functional analyses that may be valuable for both agronomic improvement and our understanding of the means by which new phenotypes may arise.

Finally, we note a promising application of plant proteomics described here relevant to our understanding of the evolutionary significance of polyploidy in plants. Specifically, we were able to document the level of homoeolog-specific protein expression and its directional bias, using protein databases constructed from genomic and transcriptomic data sets. Given the prevalence of whole genome duplications during crop evolution, we foresee the fruitful future application of these and related methods to our understanding of how gene and genome duplication generate new expression space for evaluation by human and natural selection.

Acknowledgements

We thank Anna Tuchin for help in tissue collection, Lei Gong and Chunmin Shan for helpful discussion on the manuscript. We acknowledge the Proteomics Division of the University of Florida’s Interdisciplinary Center for Biotechnology Research for assistance in LC-MS/MS analysis. The LC-MS/MS system was funded by National Institute of Health grant 1S10RR025418-01 to SC. This work was funded by Cotton Incorporated grant 09-558 and by the NSF Plant Genome Research Program, both to JFW.
References


*Cell Res* 17(5): 422-434.


Leterrier M, Barroso JB, Valderrama R, Palma JM, Corpas FJ. 2012. NADP-dependent isocitrate dehydrogenase from *Arabidopsis* roots contributes in the
mechanism of defence against the nitro-oxidative stress induced by salinity.


search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. Molecular & Cellular Proteomics 6: 1638-1655.


Tan J, Tu L, Deng F, Hu H, Nie Y, Zhang X. 2013. A genetic and metabolic analysis revealed that cotton fiber cell development was retarded by flavonoid naringenin. Plant Physiol.


Figures and tables

**Fig. 1.** Number of proteins differentially expressed during fiber development within and between wild and domesticated *G. barbadense*. A representative image of a single seed with attached trichomes, i.e. cotton fibers, is shown for each accession. Arrows represent comparisons conducted in the quantitative analyses. Numbers by the arrows denote the numbers of proteins differentially expressed for the specified comparison. Numbers in parentheses indicate the numbers of genes that were diagnosed as differentially expressed at the mRNA level, as measured by RNA-seq (data only generated for 10 and 20 dpa). For example, between stages 5 and 10 dpa within domesticated *G. barbadense*, 42 proteins were up-regulated at 5 dpa, whereas 48 were more highly expressed at 10 dpa. Similarly, between wild and domesticated accessions at 5 dpa, 63 proteins were more highly expressed in the domesticated form, while 59 were up-regulated in the wild cotton.
**Fig. 2.** Analysis of expression changes in the *G. barbadense* fiber proteome accompanying development and domestication. Expression ratios of adjacent time points were calculated using the earlier stage as denominator, and plotted in a heatmap on a log$_2$ scale. Differential expression patterns of 205 proteins are clustered on the vertical axis of the heatmap, and developmental courses from wild and domesticated accessions are clustered on the horizontal axis. Up- and down-regulation are shown in green and red colors, respectively; black color corresponds to no significant change. Based on the scheme of Bevan et al. (1998), the functional category was assigned to each protein, whose corresponding row is marked black in the central grey columns. Examples of protein expression profiles are shown on the right.
**Fig. 3.** A total of 190 proteins differentially expressed in domesticated *G. barbadense* relative to its wild counterpart at one or more developmental stages. Expression ratios were plotted in a heatmap on a log$_2$ scale. The green and red colors indicate up- and down-regulation in domesticated Pima S-7, respectively, relative to the wild form. Black color represents no significant expression change.
**Fig. 4.** Comparison of expression ratios from transcriptomic (y-axis) and proteomic (x-axis) profiling. Log$_2$ expression ratios were calculated from 20 dpa versus 10 dpa within wild (a), and domesticated cotton (b), wild versus domesticated cotton at 10 (c), and 20 dpa (d). Significant expression changes were labeled in colors: blue - proteins only, green - transcripts only, red points - both. Lines represent fitted straight trend lines from data points. Red colored proteins include: (a) - beta-alanine-pyruvate aminotransferase [Gorai.006G233600], a ferredoxin-like protein [Gorai.010G096500] and mlp-like protein 28 [Gorai.N010400]; (b) - alcohol dehydrogenase 1 (Gorai.002G222900), fasciclin-like arabinogalactan protein 9 [Gorai.008G155400] and a ferredoxin-like protein [Gorai.010G096500], endo-alpha-1,4-glucanase [Gorai.013G114800]; (d) - UDP-glucose:flavonoid 3-o-glucosyltransferase [Gorai.012G009300].
Table 1. Homoeolog-specific expression at the transcript level for genes showing biased expression at the proteomic level.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Description</th>
<th>K101</th>
<th>Pima</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gorai.001G016900</td>
<td>chaperonin 20</td>
<td>0.9</td>
<td>0.6</td>
</tr>
<tr>
<td>Gorai.001G032300</td>
<td>leucine aminopeptidase 1</td>
<td>14*</td>
<td>5.5*</td>
</tr>
<tr>
<td>Gorai.001G261100</td>
<td>aldehyde dehydrogenase family 2 member B4, mitochondrial</td>
<td>0.8</td>
<td>0.7</td>
</tr>
<tr>
<td>Gorai.001G261400</td>
<td>formate dehydrogenase</td>
<td>1.5</td>
<td>1.8</td>
</tr>
<tr>
<td>Gorai.002G062100</td>
<td>4-methyl-5(b-hydroxyethyl)-thiazole monophosphate biosynthesis</td>
<td>1.1</td>
<td>1.6</td>
</tr>
<tr>
<td>Gorai.002G119600</td>
<td>dehydrin</td>
<td>2.8</td>
<td>1.2</td>
</tr>
<tr>
<td>Gorai.002G233700</td>
<td>Unknown</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>Gorai.004G056700</td>
<td>nadh-ubiquinone oxidoreductase 24 kda subunit</td>
<td>1.7</td>
<td>1.3</td>
</tr>
<tr>
<td>Gorai.004G211400</td>
<td>glutathione peroxidase</td>
<td>1.2</td>
<td>0.8</td>
</tr>
<tr>
<td>Gorai.004G246300</td>
<td>auxin-induced protein PCNT115 isoform 1</td>
<td>7.0</td>
<td>10.1*</td>
</tr>
<tr>
<td>Gorai.004G260100</td>
<td>26S proteasome non-ATPase regulatory subunit RPN12A-like</td>
<td>1.0</td>
<td>0.9</td>
</tr>
<tr>
<td>Gorai.005G096500</td>
<td>aspartic proteinase nepenthesin-</td>
<td>6.9</td>
<td>8.4*</td>
</tr>
</tbody>
</table>

Detection of A-specific peptides only: ratios of A_T/D_T transcripts

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Description</th>
<th>K101</th>
<th>Pima</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gorai.001G016900</td>
<td>chaperonin 20</td>
<td>0.5</td>
<td>0.5*</td>
</tr>
<tr>
<td>Gorai.001G032300</td>
<td>leucine aminopeptidase 1</td>
<td>21*</td>
<td>33.6*</td>
</tr>
<tr>
<td>Gorai.001G261100</td>
<td>aldehyde dehydrogenase family 2 member B4, mitochondrial</td>
<td>0.4*</td>
<td>1.1</td>
</tr>
<tr>
<td>Gorai.001G261400</td>
<td>formate dehydrogenase</td>
<td>2.5</td>
<td>1.7</td>
</tr>
<tr>
<td>Gorai.002G062100</td>
<td>4-methyl-5(b-hydroxyethyl)-thiazole monophosphate biosynthesis</td>
<td>1.8</td>
<td>1.7</td>
</tr>
<tr>
<td>Gorai.002G119600</td>
<td>dehydrin</td>
<td>3.2*</td>
<td>2.6*</td>
</tr>
<tr>
<td>Gorai.002G233700</td>
<td>Unknown</td>
<td>1.6</td>
<td>1.3</td>
</tr>
<tr>
<td>Gorai.004G056700</td>
<td>nadh-ubiquinone oxidoreductase 24 kda subunit</td>
<td>1.3</td>
<td>1.3</td>
</tr>
<tr>
<td>Gorai.004G211400</td>
<td>glutathione peroxidase</td>
<td>1.4</td>
<td>0.8</td>
</tr>
<tr>
<td>Gorai.004G246300</td>
<td>auxin-induced protein PCNT115 isoform 1</td>
<td>11.8*</td>
<td>8.7*</td>
</tr>
<tr>
<td>Gorai.004G260100</td>
<td>26S proteasome non-ATPase regulatory subunit RPN12A-like</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>Gorai.005G096500</td>
<td>aspartic proteinase nepenthesin-</td>
<td>10.4*</td>
<td>22.6*</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Description</td>
<td>D/T Ratio 1</td>
<td>D/T Ratio 2</td>
</tr>
<tr>
<td>------------</td>
<td>--------------------------------------------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Gorai.005G187700</td>
<td>pyruvate dehydrogenase alpha subunit</td>
<td>1.3</td>
<td>3.4*</td>
</tr>
<tr>
<td>Gorai.006G172100</td>
<td>Oxysterol-binding family protein</td>
<td>0.8</td>
<td>0.6</td>
</tr>
<tr>
<td>Gorai.007G090400</td>
<td>chloroplast transketolase</td>
<td>1.6</td>
<td>2.3</td>
</tr>
<tr>
<td>Gorai.007G168600</td>
<td>Unknown</td>
<td>0.5</td>
<td>0.1*</td>
</tr>
<tr>
<td>Gorai.007G175100</td>
<td>glutathione s-transferase</td>
<td>0.9</td>
<td>2.6</td>
</tr>
<tr>
<td>Gorai.007G257500</td>
<td>methionine synthase</td>
<td>7.3</td>
<td>12.4*</td>
</tr>
<tr>
<td>Gorai.007G307000</td>
<td>20S proteasome beta subunit PBG1</td>
<td>1.6</td>
<td>0.8</td>
</tr>
<tr>
<td>Gorai.008G099900</td>
<td>Unknown</td>
<td>21.6*</td>
<td>4.6</td>
</tr>
<tr>
<td>Gorai.009G035800</td>
<td>fiber protein GLP1</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>Gorai.009G203100</td>
<td>deoxyuridine 5'-triphosphate nucleotidohydrolase-like</td>
<td>1.4</td>
<td>1.7</td>
</tr>
<tr>
<td>Gorai.009G234600</td>
<td>14-3-3 protein</td>
<td>1.6</td>
<td>0.9</td>
</tr>
<tr>
<td>Gorai.009G418800</td>
<td>syntaxin-71</td>
<td>6.1</td>
<td>2.9</td>
</tr>
<tr>
<td>Gorai.010G159200</td>
<td>ketol-acid reductoisomerase</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>Gorai.010G185300</td>
<td>Unknown</td>
<td>16.9</td>
<td>13.3*</td>
</tr>
<tr>
<td>Gorai.012G062300</td>
<td>Hop-interacting protein THI141</td>
<td>5.1</td>
<td>3.4</td>
</tr>
</tbody>
</table>

**Detection of D-specific peptides only: ratios of D<sub>T</sub>/T<sub>T</sub> transcripts**

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Description</th>
<th>D/T Ratio 1</th>
<th>D/T Ratio 2</th>
<th>D/T Ratio 3</th>
<th>D/T Ratio 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gorai.001G038600</td>
<td>glutathione peroxidase</td>
<td>0.6</td>
<td>1.0</td>
<td>0.8</td>
<td>1.1</td>
</tr>
<tr>
<td>Gorai.001G103700</td>
<td>quinone-oxidoreductase-like protein</td>
<td>0.6</td>
<td>1.5</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Gorai.002G036600</td>
<td>elongation factor ef-2</td>
<td>1.9</td>
<td>1.9</td>
<td>1.9</td>
<td>2.1*</td>
</tr>
<tr>
<td>Gorai.002G063300</td>
<td>mitochondrial nad-dependent malate dehydrogenase</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>1.6</td>
</tr>
<tr>
<td>Gorai.002G252300</td>
<td>glutathione S-transferase</td>
<td>71.2”</td>
<td>91.3”</td>
<td>86.1”</td>
<td>91.7”</td>
</tr>
<tr>
<td>Gorai.003G097400</td>
<td>nad-malate dehydrogenase</td>
<td>1.8</td>
<td>1.8</td>
<td>2.2</td>
<td>2.4*</td>
</tr>
<tr>
<td>Gorai.004G009000</td>
<td>60s acidic ribosomal protein</td>
<td>10.1</td>
<td>14.8”</td>
<td>6.7”</td>
<td>8.5”</td>
</tr>
<tr>
<td>Gorai.004G048300</td>
<td>utp:alpha-d-glucose-1-phosphate uridylyltransferase</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>1.0</td>
</tr>
<tr>
<td>Gorai.004G063900</td>
<td>6-phosphogluconate dehydrogenase</td>
<td>2.7</td>
<td>2.4</td>
<td>2.3</td>
<td>1.9</td>
</tr>
<tr>
<td>Gorai.004G096800</td>
<td>glutaredoxin-C4</td>
<td>1.5</td>
<td>1.5</td>
<td>1.7</td>
<td>1.6</td>
</tr>
<tr>
<td>Gorai.004G166200</td>
<td>Unknown</td>
<td>0.4</td>
<td>0.3</td>
<td>0.5</td>
<td>0.5*</td>
</tr>
<tr>
<td>Gorai.004G226600</td>
<td>RPM1-interacting protein 4-like</td>
<td>2.4</td>
<td>0.7</td>
<td>2.7”</td>
<td>0.9</td>
</tr>
<tr>
<td>Gorai.005G028100</td>
<td>rna-binding glycine-rich protein</td>
<td>1.3</td>
<td>1.3</td>
<td>1.3</td>
<td>1.1</td>
</tr>
<tr>
<td>Gorai.005G050200</td>
<td>malate dehydrogenase</td>
<td>1.4</td>
<td>1.4</td>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td>Gorai.006G070400</td>
<td>N-carbamoyl-L-amino acid hydrolase</td>
<td>14.1”</td>
<td>13.7”</td>
<td>23.6”</td>
<td>108.5 *</td>
</tr>
<tr>
<td>Gorai.006G156700</td>
<td>60s acidic ribosomal protein</td>
<td>0.6</td>
<td>0.4</td>
<td>0.6</td>
<td>0.4*</td>
</tr>
<tr>
<td>Gorai.007G022600</td>
<td>elicitor-responsive protein 3-like</td>
<td>3.7</td>
<td>1.4</td>
<td>2.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Gorai.008G048100</td>
<td>glycine decarboxylase complex h-protein</td>
<td>1.1</td>
<td>0.9</td>
<td>1.0</td>
<td>0.8</td>
</tr>
<tr>
<td>Gorai.008G155400</td>
<td>fasciclin-like arabinogalactan protein 9</td>
<td>9.2</td>
<td>5.0”</td>
<td>8.8”</td>
<td>5.3*</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Protein Name</td>
<td>10 dpa</td>
<td>20 dpa</td>
<td>20-10 dpa</td>
<td>10-20 dpa</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------------------------------------------</td>
<td>---------</td>
<td>---------</td>
<td>------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Gorai.008G169300</td>
<td>glutamate decarboxylase</td>
<td>2.8</td>
<td>2.9</td>
<td>4.1*</td>
<td>1.2</td>
</tr>
<tr>
<td>Gorai.008G285900</td>
<td>aspartic proteinase nepenthesin-1</td>
<td>68*</td>
<td>31.0*</td>
<td>150.</td>
<td>8.6*</td>
</tr>
<tr>
<td>Gorai.009G060500</td>
<td>aspartate aminotransferase</td>
<td>6.2</td>
<td>5.4*</td>
<td>1.2</td>
<td>1.6</td>
</tr>
<tr>
<td>Gorai.009G078400</td>
<td>beta-galactosidase 8</td>
<td>2.6</td>
<td>1.2</td>
<td>4.5*</td>
<td>1.6</td>
</tr>
<tr>
<td>Gorai.009G432600</td>
<td>Uridine 5'-monophosphate synthase</td>
<td>1.0</td>
<td>0.7</td>
<td>0.3*</td>
<td>0.7</td>
</tr>
<tr>
<td>Gorai.010G221900</td>
<td>probable leucine-rich repeat receptor-like protein kinase at5g49770-like</td>
<td>0.7</td>
<td>1.0</td>
<td>0.7</td>
<td>1.2</td>
</tr>
<tr>
<td>Gorai.011G026000</td>
<td>clathrin light chain protein</td>
<td>1.2</td>
<td>1.2</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Gorai.011G035400</td>
<td>GDSL esterase/lipase</td>
<td>1.7</td>
<td>1.9</td>
<td>3.8*</td>
<td>3.5*</td>
</tr>
<tr>
<td>Gorai.012G142600</td>
<td>2-nitropropane dioxygenase</td>
<td>34.8*</td>
<td>41.8*</td>
<td>2.8*</td>
<td>4.4*</td>
</tr>
<tr>
<td>Gorai.012G146500</td>
<td>dolichyl-diphosphooligosaccharide--protein glycosyltransferase 48 kda subunit-like</td>
<td>1.3</td>
<td>1.3</td>
<td>1.5</td>
<td>1.6</td>
</tr>
<tr>
<td>Gorai.013G216000</td>
<td>leucine aminopeptidase 3</td>
<td>1.1</td>
<td>3.8*</td>
<td>1.3</td>
<td>1.0</td>
</tr>
</tbody>
</table>

1RNA-seq datasets of 10 and 20 dpa were separately analyzed.

* Significant transcriptional bias toward the same direction of protein level bias

& Significant transcriptional bias toward the opposite direction of protein level bias
Supporting Information

**Table S1.** Analyses of false discovery rates (FDRs)

Number of proteins identified from iTRAQ run 1

<table>
<thead>
<tr>
<th>Critical FDR</th>
<th>Local FDR</th>
<th>Global FDR</th>
<th>Global FDR from Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0%</td>
<td>1007</td>
<td>1143</td>
<td><strong>1158</strong></td>
</tr>
<tr>
<td>5.0%</td>
<td><strong>1080</strong></td>
<td>1267</td>
<td>1278</td>
</tr>
<tr>
<td>10.0%</td>
<td><strong>1116</strong></td>
<td>1377</td>
<td>1382</td>
</tr>
</tbody>
</table>

Number of proteins identified from iTRAQ run 2

<table>
<thead>
<tr>
<th>Critical FDR</th>
<th>Local FDR</th>
<th>Global FDR</th>
<th>Global FDR from Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0%</td>
<td>933</td>
<td>1092</td>
<td><strong>1073</strong></td>
</tr>
<tr>
<td>5.0%</td>
<td><strong>1000</strong></td>
<td>1204</td>
<td>1198</td>
</tr>
<tr>
<td>10.0%</td>
<td><strong>1040</strong></td>
<td>1291</td>
<td>1302</td>
</tr>
</tbody>
</table>

Number of proteins identified from iTRAQ run 3

<table>
<thead>
<tr>
<th>Critical FDR</th>
<th>Local FDR</th>
<th>Global FDR</th>
<th>Global FDR from Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0%</td>
<td>863</td>
<td>959</td>
<td><strong>960</strong></td>
</tr>
<tr>
<td>5.0%</td>
<td><strong>904</strong></td>
<td>1053</td>
<td>1049</td>
</tr>
<tr>
<td>10.0%</td>
<td><strong>922</strong></td>
<td>1149</td>
<td>1152</td>
</tr>
</tbody>
</table>

* Numbers in bold are close to that from protein identification with above 95% confidence level.
Table S2. Protein identifications at 95% confidence level

<table>
<thead>
<tr>
<th></th>
<th>run 1</th>
<th>run 2</th>
<th>run 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectra</td>
<td>25638</td>
<td>27140</td>
<td>23863</td>
<td></td>
</tr>
<tr>
<td>Distinct peptides</td>
<td>13900</td>
<td>13529</td>
<td>11275</td>
<td></td>
</tr>
<tr>
<td>Proteins</td>
<td>1093</td>
<td>993</td>
<td>878</td>
<td>1317</td>
</tr>
</tbody>
</table>
Large supplemental files are available to be downloaded at


**Table S3** Protein identification and quantification by ProteinPilot iTRAQ analyses

**Table S4** Significant protein expression changes during fiber development and between wild and domesticated *G. barbadense*

**Table S5** RNA-seq analysis of gene differential expressions

**Table S6** Homoeolog-specific peptides identified in *G. barbadense* proteomes
**Fig. S1.** Representative example of the cation exchange chromatograph of iTRAQ-labeled peptides derived from wild and domesticated *G. barbadense*. An off-line 2D LC-MS/MS method with SCX chromatography as a first step was used to fractionate the fiber proteome. The collected fractions were combined into 12 final fractions as indicated by pink lines, and submitted into a quadrupole TOF MS/MS system. The area coverage with standard deviation (SD) and coefficient of variation (CV) in each fraction among three sets of iTRAQ is shown.

<table>
<thead>
<tr>
<th>Fraction no.</th>
<th>Retention Time (min)</th>
<th>area coverage (% with SD)</th>
<th>CV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.00</td>
<td>37.15 ± 1.37</td>
<td>3.69</td>
</tr>
<tr>
<td>2</td>
<td>23.60</td>
<td>5.13 ± 0.37</td>
<td>7.22</td>
</tr>
<tr>
<td>3</td>
<td>26.40</td>
<td>1.34 ± 0.09</td>
<td>6.70</td>
</tr>
<tr>
<td>4</td>
<td>29.20</td>
<td>1.61 ± 0.13</td>
<td>8.23</td>
</tr>
<tr>
<td>5</td>
<td>32.00</td>
<td>2.07 ± 0.1</td>
<td>4.85</td>
</tr>
<tr>
<td>6</td>
<td>34.80</td>
<td>2.52 ± 0.13</td>
<td>5.28</td>
</tr>
<tr>
<td>7</td>
<td>40.40</td>
<td>5.53 ± 0.28</td>
<td>5.00</td>
</tr>
<tr>
<td>8</td>
<td>46.00</td>
<td>5.08 ± 0.26</td>
<td>5.08</td>
</tr>
<tr>
<td>9</td>
<td>51.60</td>
<td>4.30 ± 0.21</td>
<td>4.88</td>
</tr>
<tr>
<td>10</td>
<td>63.65</td>
<td>7.34 ± 0.33</td>
<td>4.50</td>
</tr>
<tr>
<td>11</td>
<td>70.91</td>
<td>3.00 ± 0.27</td>
<td>8.87</td>
</tr>
<tr>
<td>12</td>
<td>130.00</td>
<td>24.93 ± 0.29</td>
<td>1.16</td>
</tr>
</tbody>
</table>
**Fig. S2.** Panther protein family classification. (a) Pie chart of fiber protein families. (b) Over- and under-representation of fiber protein families relative to the entire proteome. Fractional differences were calculated by each category. *Significant difference, based on binomial tests with bonferroni correction.
CHAPTER 4

PROTEOMIC PROFILING OF FIBER DEVELOPMENT AND DOMESTICATION IN UPLAND COTTON (*GOSSYPIUM HIRSUTUM* L.)

A paper in preparation for submission to *Molecular Biology and Evolution*

Guanjing Hu, Jin Koh, Dharminder Pathak, Sixue Chen, and Jonathan F. Wendel

Abstract

A comparative proteomic analysis was performed to explore the evolutionary genetics underlying the enhanced fiber traits in modern upland cotton (*G. hirsutum* L.). The dynamics of fiber proteomes were examined for four representative developmental stages using two complementary proteomic approaches, 2-DE and iTRAQ LC-MS/MS. Approximately 1000 protein features were characterized using each strategy, collectively resulting in identification and functional categorization for 1223 proteins. Notably, homoeolog-specific peptides were diagnosed for 558 proteins in the allopolyploid cotton proteome. About 30% of the proteomes were differentially expressed during fiber development within wild and domesticated cotton, between which the largest developmental divergence was found for the period between of 10 to 20 days following pollination. Furthermore, expression of 240 iTRAQ proteins and 293 2-DE spots were altered by domestication, collectively representing multiple cellular and metabolic processes with functional tendencies toward metabolism, energy, protein synthesis and destination, defense and stress response. Analyses of homoeolog-specific expression indicate that duplicated gene products in cotton fibers can be differently regulated in
response to evolutionary change. Altogether, these results demonstrate the power of proteomics for the analysis of crop domestication and phenotypic evolution.

**Introduction**

Lint fibers of upland cotton (*Gossypium hirsutum* L.) are the most prevalent natural fibers used in the textile industry, which support the manufacture of diverse consumer and industrial products throughout the world. In 2012, the planting of upland cotton covered over 12 million acres, and the business revenue stimulated by the production is estimated to be approximately $100 billion in the USA alone (http://www.cotton.org). In addition to their agronomic and economic importance, cotton fibers provide an excellent single-celled model for the study of basic biological processes in plants (Haigler *et al.*, 2012). Each cotton “fiber” is a single remarkably elongated cell that differentiated from individual epidermal cells of the ovule integument. Following cell differentiation, which typically occurs 2-3 days before anthesis, fiber cells become dramatically elongated during the next 20 days post-anthesis (dpa), followed by cell wall thickening via secondary wall synthesis. At full maturity, fibers of commercial upland cotton lines typically reach up to 40 mm in length, being one to three thousand times greater in length than in diameter. Thus, it is of great importance to understand the genetic basis of this striking example of plant cell development.

As a tetraploid species, *G. hirsutum* originated 1-2 million years ago through hybridization and genome doubling between a diploid, A-genome *Gossypium* species from Africa or Asia, and a D-genome diploid similar to that found in Central and South America. This polyploidy event, occurring after 5-10 million years of evolutionary
divergence between the diploid progenitors, also gave birth to four other allopolyploid species including *G. barbadense*, whose cultivated form is often known as Egyptian or Pima cotton. Both grown for their seed hairs, and secondarily for their oil seeds, *G. hirsutum* and *G. barbadense* were independently domesticated at least 5000 years ago in the Yucatan Peninsula and the intermontane Peruvian Andes areas, respectively, by ancient human cultures (Wendel *et al.*, 2010). Within a large indigenous range encompassing most of Mesoamerica and the Caribbean, *Gossypium hirsutum* exists as a continuum of morphological forms from wild to domesticated plants. Its most wild form, race *yucatanense* (Tx2094), is a sprawling perennial shrub found as dominant component of native beach strand vegetation in the Yucatan Peninsula (Brubaker & Wendel, 1994). In comparison with the modern crop with enhanced fiber fineness and length, the shorter and un-spinable fibers of the wild plant provide us the baseline condition to understand the evolutionary steps involved in the morphological transformation, and ultimately the underlying genetic mechanisms that led to modern, agronomically improved cotton.

Growth curve analysis of wild and domesticated cotton fibers indicated that long fibers are associated with increased growth rate during primary wall synthesis and a prolonged fiber elongation period (Applequist *et al.*, 2001). The agronomically advanced cultivar TM1 of *G. hirsutum* reaches a maximum growth rate between 10 and 15 dpa, whereas the wild accession Tx2094 displayed little or no growth between 10 and 15 dpa, followed by a significantly increased rate between 15 and 20 dpa. Although the final fiber length of Tx2094 is about 60% of that in TM1, microscopic observation indicated that 20 dpa is near the beginning of secondary wall synthesis for both accessions, and the domesticated cotton is only slightly ahead of the wild form (Rapp *et al.*, 2010).
Comparative expression profiling at the transcriptome level demonstrated that the transcriptomes of these examples of wild and domesticated *G. hirsutum* were massively rewired by domestication and crop improvement, with nearly a quarter of genes in the genome were differentially expressed. These data suggest that the phenotypic changes that exist between fibers from wild and domesticated cotton reflect a complex alteration of the transcriptional developmental network (Rapp *et al*., 2010). A number of genes and metabolic pathways with altered expression were identified in domesticated cotton, such as profilins and the modulation of reactive oxygen species (ROS), confirmed to be associated with increased fiber length and quality seen in the crop (Chaudhary *et al*., 2008; Hovav *et al*., 2008; Chaudhary *et al*., 2009; Bao *et al*., 2011).

In addition to genetic and transcriptomic analysis, proteomic investigation of fiber development has proven to be a powerful and complementary approach to study cotton domestication (Hu *et al*., 2013). Between wild and domesticated forms of *G. barbadense*, 190 out of 1317 fiber proteins sampled were differentially expressed at one or more stages of fiber development, whereas only one of these proteins was differentially expressed at the transcriptional level. It is clear that mRNA expression levels only partially predict protein abundances, as well as the cellular and physiological activities mainly executed by proteins. Although a two-dimensional gel electrophoresis (2-DE)-based approach has been established and used to identify more than 200 proteins that change during fiber development in upland cotton (Yao *et al*., 2006; Li *et al*., 2007; Yang *et al*., 2008; Pang *et al*., 2010; Zhang *et al*., 2013), there are no reports of proteomic transformations accompanying human-mediated selection in *G. hirsutum*. 
Here, we analyzed the fiber proteomes of wild and domesticated *G. hirsutum* at four developmental time points using two independent proteomic approaches, 2-DE and isobaric tag for relative and absolute quantification (iTRAQ) technology coupled with LC-MS/MS. Using each method, we characterized the global expression patterns in both wild and domesticated *G. hirsutum*, and documented those proteins differentially expressed during fiber development. We further identified a number of differentially expressed proteins as candidates for functional analyses that may yield insight into domestication and future cotton improvement. Finally, we compared the results produced by each proteomic strategy and the reproducibility of protein profiling by these methods, a subject of current interest in proteomics research (Aggarwal et al., 2006; Chong et al., 2006; Wu et al., 2006; Fenselau, 2007; Thelen, 2007; Vercauteren et al., 2007).

**Materials and methods**

**Plant materials, tissue collection and protein extraction**

For elite cultivar of modern upland cotton, we used the genetic and cytogenetic standard, Texas Marker Stock 1 (TM-1). For wild *G. hirsutum*, we chose var. *yucatanense* accession Tx2094 (US Department of Agriculture GRIN accession PI 501501, collected by J McD Stewart), an unambiguously wild form, based on previous morphological and molecular evidence (Wendel et al., 2010). Plants from each accession were grown in the Bessey Hall Greenhouse at Iowa State University. Flowers were tagged at anthesis and developing bolls were harvested at 5, 10, 20 and 25 days post-anthesis (dpa), representing key developmental stages of primary wall synthesis and elongation (5 and 10 dpa), and the transition to (20 dpa) and continuation (25 dpa) of
secondary wall synthesis. Bolls were dissected immediately after harvest and ovules were frozen in liquid nitrogen and stored at -80°C. Fiber proteins were extracted and purified as previously reported (Hu et al., 2013). Three samples per developmental stage and per cotton accession were prepared accounting for three biological replicates.

2-DE and image analysis

Extracted proteins were dissolved in isoelectric focusing (IEF) buffer (8 M urea, 2 M thiourea, 4% [w/v] CHAPS, 2% [v/v] Triton X-100, and 50 mM DTT) at room temperature, and quantified using the 2-D Quant Kit (GE Healthcare Biosciences, Pittsburgh, PA, USA) with bovine serum albumin as standard. For each 1mg of fiber protein, the final volume was adjusted to 450 µl with IEF buffer and then 2.25 µl of IPG buffer 3–10 NL (GE Healthcare Biosciences) was added. After centrifugation at 10,000 x g for 10 min to remove insoluble materials, the supernatants were loaded on 24-cm 3-10 NL immobilized pH gradient (IPG) strips (GE Healthcare Biosciences) followed by rehydration for 90 min at room temperature. Using an IPGphor unit (GE Healthcare Biosciences), IEF was carried out with the following protocol: active rehydration at 50 V for 10 hr, 100 V for 100 V hr, 500 V for 500 V hr, and 8000 V for 99 kV hr at 20 °C, with a maximum current setting of 20 μA per strip. After completion of IEF, the strips were reduced for 20 min with 2.0% [w/v] DTT in equilibration buffer (6 M urea, 50 mM Tris-HCl, pH 8.8, 30% [v/v] glycerol, 2% [w/v] sodium dodecyl sulfate (SDS)), cysteine alkalized for 20 min with 2.5% [w/v] iodoacetamide in equilibration buffer, rinsed with SDS running buffer (1.5 M Tris-HCl, 6 M urea, 30% [v/v] glycerol, 5% [w/v] SDS), and then placed on 12% self-cast polyacrylamide gels. Second-dimensional electrophoresis
was performed in an Ettan DALT six System (GE Healthcare Biosciences), using 80 V for 2 hr and 100 V for 16 hr. Finished gels were rinsed in deionized water and stained overnight with colloidal Coomassie staining solution (20% [v/v] ethanol, 1.6% [v/v] phosphoric acid, 8% [w/v] ammonium sulfate, 0.08% [w/v] Coomassie Brilliant Blue G-250).

Stained gels were scanned using ImageScanner (GE Healthcare Biosciences) at a resolution of 600 dpi and 16-bit grayscale pixel depth. Image and statistical analysis was conducted with the Progenesis SameSpots software version 4.0 (Non Linear Dynamics, Durham NC, USA), using standard procedures and default parameters. Briefly, spot detection, background subtraction, and normalization were performed following automatic alignment of all images with manual inspection. Detected spots were 100% matched across images, so that all gels contain the same number of spots without missing values. Differential expression of spot volumes that represent protein abundances was assessed using the built-in statistical tool in the SameSpots software for contrasts of interest. Our criteria for significant changes were based on an ANOVA p-value < 0.05 as calculated with a fold change cutoff of > 1.2 or < 0.8.

iTRAQ LC-MS/MS, protein identification and quantification

Extracted protein were dissolved in protein buffer (8M urea, 25mM triethylammonium bicarbonate (TEAB), 2% [v/v] TX-100, 0.1% SDS [w/v], pH 8.5), and prepared for iTRAQ labeling as described in Hu et al. (2013). For each sample, 100 µg of protein was reduced, alkylated, and trypsin-digested using the iTRAQ Reagents 8-plex Kit according to the manufacturer’s instructions (AB Sciex, Inc., Foster City, CA, USA).
The TM1 proteins were labeled with iTRAQ tags 113 (5 dpa), 114 (10 dpa), 115 (20 dpa) and 116 (25 dpa), and Tx2094 proteins were labeled with tags 117 (5 dpa), 118 (10 dpa), 119 (20 dpa) and 121 (25 dpa). After combining the labeled samples, the peptide mixture was fractionated with strong cation exchange chromatography, and analyzed using an off-line 2D LC-MS/MS method as previously described (Hu et al., 2013).

The MS/MS data were processed by a thorough database search considering biological modifications and amino acid substitutions under the Paragon™ algorithm (Shilov et al., 2007) and the Pro Group™ algorithm, using ProteinPilot version 4.5 software (AB Sciex, Inc.). As described in Hu et al. (2013), three protein databases were used for iTRAQ protein and homoeolog identification, including a non-redundant Gossypium protein database and two separate A- and D- genome diploid databases. Methylthio-cysteine and amine groups at the N-terminus and lysine were considered for the fixed modifications and variable modifications were included for post-translational modifications (PTMs). The cutoff of protein identification was set to a confidence level of 95%. The global false discovery rate (FDR) of identified protein lists was determined by performing searches against the reversed protein databases, with estimates derived from both the conventional approach and a nonlinear fitting method (Tang et al., 2008) as shown in Supporting Information Table S1. The identified proteins were annotated with gene ontology categories using Blast2GO (Conesa et al., 2005), assigned to protein families using PANTHER (Mi et al., 2010) and functional classes based on the Arabidopsis functional catalog (Bevan et al., 1998).

For protein quantification, only MS/MS spectra that were uniquely identified for a particular protein were used to extract peak intensities of iTRAQ labeling tags, which
were subsequently normalized across samples using the built-in bias correction function of ProteinPilot (AB Sciex, Inc.). Relative protein quantification for comparisons of interest was generated with p values using the software’s standard procedures. To be identified as being significantly differentially expressed, a protein must have been quantified with a fold change of > 1.2 or < 0.8 and p < 0.05 in at least two of the biological triplicates, along with a Fisher’s combined probability of <0.05 (Fisher, 1948).

**Other statistical analysis**

Hierarchical clustering with regular bootstrap (BP) and approximately unbiased bootstrap (AU) p-values was performed using the R software package pvclust (Suzuki and Shimodaira, 2006), specifying average linkage and Pearson’s correlation distance metric with 10,000 iterations. The bootstrap p-values indicated how strongly the cluster is supported by the data. For example, for a cluster with p >95%, the hypothesis that “the cluster does not exist” is rejected with a significance level of 5%. The branch length represents the degree of dissimilarity in protein expression among conditions.

**Results**

**Parallel proteomic strategies applied to developing cotton fibers**

Proteins prepared from cotton fibers collected at four key stages, i.e, 5 (primary wall synthesis), 10 (fast fiber elongation), 20 and 25 dpa (the onset and continuation of secondary wall synthesis) were analyzed with two independent proteomic approaches, as shown in Figure 1. Electrophoresis of twenty-four 2-DE gels (Supporting information Figure S1) and three eight-plex iTRAQ experiments were performed in parallel to profile
fiber proteomes in wild (Tx2094) and domesticated (TM1) cotton accessions. A high-resolution 2-DE map was established on an isoelectric focusing range of pH 3-10, containing 907 spots detected and reproducibly matched across all analyzed gels (Table S2). By matching our spot map to a newly constructed 2-DE reference dataset of upland cotton (Zhang et al., 2013), protein identification was retrieved for 184 spots representing electrophoretic isoforms of 147 proteins. Using iTRAQ labeling coupled with 2D LC-MS/MS experiments, a total of 1189 fiber proteins were characterized (identification confidence > 95%, FDR < 1%; Table S1 and S3), among which 558 proteins were diagnosed with homoeolog-specific peptides (Table S4).

Collectively, 2-DE and iTRAQ analyses resulted in the identification of 1223 non-redundant proteins (i.e. unique Gossypium gene loci), with 113 proteins identified by both methods (Table S2), accounting for 77% and 10% of the 2-DE and iTRAQ datasets, respectively. Although both sets of protein identifications covered almost all functional families encoded in the Gossypium genome, protein families of transferase, isomerase, chaperone and transfer/carrier protein were over-represented by 2-DE proteins, while nucleic acid binding protein, ligase and membrane traffic protein were more highly represented by proteins identified by iTRAQ (Figure 2a). 2-DE and iTRAQ together identified proteins with a wide range of isoelectric points (pIs) and molecular masses without observable gaps (Figure 2b and c). Proteins with the pI range 4 to 7 dominated the pI spectrum, accounting for 86% (2-DE) and 74% (iTRAQ) of the total identified proteins from cotton fibers. High pI proteins (above 7) were better represented by iTRAQ, and notably, 52 proteins, with pI above 10, were identified outside the separation range of our 2-DE analysis (Figure 2b). Neither approach identified proteins with pI
values below 4. Proteins within the mass range 10 to 70 kD represented 89% (2-DE) and 86% (iTRAQ) of the identified proteins (Figure 2c). High mass proteins above 100 kD were only resolved by iTRAQ, and the protein with highest mass was identified as a 566 kD auxin transport protein [Gorai.001G206500]. These results suggested that iTRAQ is more inclusive with respect to proteins having high molecular masses and pIs, relative to the 2-DE method.

Quantitative proteomic changes during fiber development within wild and domesticated *G. hirsutum*

To study protein expression during fiber development, proteomic changes were examined between adjacent developmental stages (5-10, 10-20, and 20-25 dpa), which revealed 292 iTRAQ proteins (24.6% of 1189 proteins) and 331 2-DE spots (36.4% of 907 spots; Table S4) that were significantly differentially expressed within wild or domesticated cotton fibers. Of these, wild (Tx2094) and domesticated (TM1) cotton displayed similar amount of developmental expressional variation (iTRAQ, 202 vs 203 proteins; 2-DE, 193 vs 181).

As shown in Fig. 3, iTRAQ analysis (blue numbers) revealed that in Tx2094 the highest amount of protein expression change occurred early during fiber development (5-10 dpa, 156 proteins), which was two-fold higher than those differentially expressed during later time courses (10-20 dpa, 45; 20-25 dpa, 55; \(p < 0.05\), fisher’s exact test), while higher amounts of differential expression were found in TM1 until 20 dpa (10-20 dpa, 121; 10-20 dpa, 132); fewer expression changes were observed in TM1 from 20 to 25 dpa (29 proteins; \(p < 0.05\), fisher’s exact test). By examining the multivariate
expression pattern of these significantly changed proteins, the same time courses of 5-10 and 20-25 dpa were clustered between TM1 and Tx2094 with about 90% bootstrapping support, which suggested that the developmental change in domesticated cotton is more similar to its wild progenitor during early fiber elongation and secondary wall synthesis, than that during their intermediate period represented by the 10-20 dpa period of rapid fiber elongation during primary wall synthesis (Figure 4). As shown by functional categorization of these developmental changes (Table 1), more proteins in the functional classes of “metabolism”, “energy”, “protein synthesis”, “protein destination and storage” and “signal transduction” were differentially regulated from 10 to 20 dpa in TM1 versus Tx2094 (p < 0.05, fisher’s exact test). Compared to most functional classes, where proteins displayed diverse expression changes during fiber development, ribosomal protein subunits, translation factors and tRNA synthetases in the class “protein synthesis” were predominantly down-regulated from 5 to 20 dpa (p < 0.05, fisher’s exact test).

In contrast to the results obtained from the iTRAQ analysis, numbers of differentially expressed 2-DE spots did not decrease during fiber development (Figure 3, brown numbers); instead, statistically equivalent amounts of proteomic variations were found during early and late developmental courses within each accession (TM1, 69 vs 65 spots; Tx2094, 83 vs 75 spots). However, more proteins were differentially expressed from 10 to 20 dpa in TM1 than in Tx2094 (94 vs 44 spots; p < 0.05, fisher’s exact test), consistent with the pattern seen in iTRAQ data. In addition, multivariate expression patterns of 2-DE spots also indicated clustering of identical time courses between wild and domesticated cotton for 5-10 and 20-25 dpa (~80% and 90% bootstrapping, respectively; Figure 4), thereby suggesting that the most dramatic rewiring of the fiber
proteome caused by domestication and crop improvement occurred between 10 and 20 dpa. A total of 82 differentially expressed spots were identified, among which the identified proteins of 69 spots were also profiled by iTRAQ analysis. In comparison to the protein functional classes represented by iTRAQ data, no protein identified by 2-DE analysis were categorized into the classes of “transcription”, “intracellular traffic” or “cell structure” (Table 1).

Differential protein expression between wild and domesticated G. hirsutum

When comparing protein expression levels between TM1 and Tx2094 at each time point, 240 iTRAQ proteins and 293 2-DE spots were differentially expressed at one or more developmental stages (Table S4). As shown in Figure 3, the numbers of 2-DE spots displaying expression change were distributed equally along four developmental time points, while for the proteins profiled by iTRAQ, the highest number of expression changes occurred early in fiber elongation (5 dpa, 137 proteins), followed by fewer changes later during primary wall synthesis (10 dpa, 78 proteins) and the transition to secondary cell wall synthesis (20 dpa, 100 proteins), with the fewest changes during secondary wall synthesis (25 dpa, 48 proteins). The distribution of up-regulation between wild and domesticated accessions is statistically symmetric (p > 0.05, chi-squared test), except for 2-DE results where more proteins were up-regulated in TM1 than Tx2094 at 5 dpa, with the direction switched at 10 dpa. Functional analysis of these proteins, including all from iTRAQ and 64 identified 2-DE spots, indicated that all functional categories are involved in proteomic divergence between TM1 and Tx2094, with more than half of proteins belonging to the categories of “metabolism”, “energy”, “protein
synthesis”, “protein destination”, and “defense and stress response” (Figure 5).

Interestingly, while the class of “protein synthesis” was largely up-regulated at 5 dpa by domestication, 5 of the 7 “transcription” related proteins were down-regulated, including a transcriptional factor [Gorai.004G264900] (expression profile shown in Figure 5, right panel “288. TF3”), a transcription coactivator [Gorai.010G214300] and RNA binding proteins [D_Gorai.010G102100, Gorai.004G004300, Gorai.005G052700].

Homoeolog expression in allopolyploid G. hirsutum

Our analyses presented to this point studied total protein accumulation from both homoeologs (\(A_T\) and \(D_T\), where the subscript indicates the specific genome in the allopolyploid) from tetraploid cotton (which contains \(A_T\) and \(D_T\) genomes), utilizing iTRAQ and 2-DE approaches in parallel. To analyze the separate homoeologous contributions to the fiber proteome of allopolyploid G. hirsutum, protein identification with evidence of genome-of-origin was required to characterize the expression pattern for protein homoeologs separately. This is one notable advantage of MS-based iTRAQ analysis, which provides the capability of diagnosing homoeolog-specific peptides, as shown in Hu et al. (2013). Compared to iTRAQ proteins, no homoeolog-specific 2-DE spots were detected, an unsurprising result given that far fewer proteins (147 for 2 DE vs. 1189 for iTRAQ) were identified to screen for homoeolog-specific peptides.

Among the 558 proteins identified by iTRAQ with genome-diagnostic peptides, 137 proteins had peptides diagnostic for both homoeologs (Table S5). Different from the expression profile of total protein accumulation, homoeolog-specific expression was quantified using only the corresponding \(A_T\) or \(D_T\) peptides (as opposed to all peptides).
For example, the total expression of a dehydrin protein associated with stress responses [Gorai.002G119600] was measured using spectra of five distinct peptides, where significant expression changes were identified between wild and domesticated cotton at 25 dpa, as well as from 20 to 25 dpa within both accessions (Figure 6a, green dots). In the homoeolog-only analysis, two of these peptides were diagnosed as specific to D_T, one as specific to A_T, and two peptides that were shared by both homoeologs (Figure 6b). As shown in Figure 6a (blue and red dots), measurements of A_T and D_T peptides were separately clustered in some comparisons. At 25 dpa, for example, D_T peptides were down-regulated, as shown by total expression, but all A_T data (multiple spectra of one peptide) indicated up-regulation in domesticated cotton, which suggested differential regulation of protein homoeologs in allopolyploid G. hirsutum. Another example shown is Beta-hydroxyisobutyryl-CoA hydrolase 1 isoform 2 [Gorai.004G108600]: although no significant change was inferred based on total accumulation (across all diagnostic and non-diagnostic peptides) (Figure 5c), the D_T homoeolog appeared to be differentially expressed in most comparisons. Using this comparative framework, homoeolog expression in the developing fibers of wild and domesticated G. hirsutum was inspected for the 137 proteins; however, differential regulation was not clearly shown by other proteins, where the A_T and D_T homoeologs displayed similar expression patterns, such as aspartic proteinase A1 [Gorai.004G031800] shown in Figure 5d.

Among the remaining proteins with diagnostic peptides detected for one of the two homoeologs, 92 and 113 proteins displayed A_T and D_T peptides in all three biological replicates, respectively. As suggested in Hu et al. (2013), this biased expression of homoeolog-specific peptides reflects the non-random nature of mass spectrometry in
detecting high abundance peptides, which therefore can be used as an indicator of biased homoeolog expression favoring the sub-genome detected. Accordingly, the numbers of \( A_T \) and \( D_T \) biased homoeolog expression are statistically equal (\( p > 0.05 \), fisher’s exact test), suggesting that homoeolog expression bias of fiber proteins is pervasive, but also is balanced with respect to genome-of-origin in \( G. hirsutum \).

**Discussion**

**2-DE and iTRAQ: complementary proteomic approaches**

Here we report an original study of comparative proteomics that investigates how the evolutionary history of crop domestication has altered protein expression of developing fibers in modern upland cotton. We characterized the developmental dynamics of fiber proteomes and conducted intraspecific comparative analyses using 2-DE and iTRAQ methods, both known as powerful techniques to study comprehensive protein changes in the proteomics field (Rose *et al.*, 2004; Thelen, 2007). In 2-DE, protein mixtures are separated by native charge (pI) followed by molecular mass, i.e., the two properties referred to as two-dimensional gel electrophoresis, to be resolved into complex maps of spots. Due to the diverse properties of proteins, the reproducibility of 2-DE gels can be problematic, thereby imposing a great challenge on comparative analysis, which requires matching over a thousand spots across multiple gels. Even though all our 2-DE gels were checked for reproducibility and aligned for complete spot matching using an advanced 2-DE analysis software (Progenesis SameSpots), inter-gel variation can be reduced but remain an inherent problem with standard single-stain 2-DE assays. In contrast to that, one obvious advantage of iTRAQ technology is the ability to pre-label
protein samples with up to eight isobaric tags, hence allowing their comparison within the same experiment. Following protein separation, both methods relies on mass spectrometry for peptide recognition and protein identification, and comparison studies of their proteomic coverage have revealed these two techniques are complementary due to their methodological and technological differences (Thelen, 2007; Diz et al., 2012).

In our study, a small set of 2-DE spots were identified by searching against a reference dataset generated by Zhang et al. (2013), given that the same experimental method was used in both studies (Yao et al., 2006), and our gel images are compatible with the reference map. On the other hand, a much larger number of proteins were identified using iTRAQ coupled with LC-MS/MS analysis (Figure 1). By comparing the biochemical properties of proteins resulted from two methods, iTRAQ appears to provide a more comprehensive catalog, by including proteins with a wider range of pI and molecular mass. Nevertheless, the uniqueness of 2-DE for separating protein isoforms, allows it to provide additive hence complementary information on post-translational modifications and other protein-specific mechanisms. As shown in Figure 5 (top right panels), two spots were identified with opposite expression patterns for a dehydroascorbate reductase [Gorai.012G068600], suggesting that its isoforms were differentially affected by evolutionary force. Also identified by iTRAQ, this protein was uniquely profiled and displayed an intermediate expression pattern between those of 2-DE isoforms. This discrepancy in protein identification is because of the peptide-centric nature of iTRAQ identification: for different protein isoforms, only a few peptides bear distinct point modifications or proteolytic sites, which are often under-represented due to their relatively low abundance, while their common peptides at higher abundance are
more likely to be identified and grouped together as a single identification; as a result, various protein isoforms are much less likely to be studied by iTRAQ compared to 2-DE. Overall, different strengths of 2-DE and iTRAQ underscore the importance of technical diversity in revealing the complexity of biological system and evolutionary studies at protein level.

Proteomic dynamics during fiber development

Since the 2-DE-based approach was established for cotton fiber proteomics, a number of studies have been performed to explore the mechanism of fiber development at the protein level (Yao et al., 2006; Li et al., 2007; Yang et al., 2008; Pang et al., 2010; Zhao et al., 2010; Zhang et al., 2013). Proteomic analyses of 10 dpa cotton ovules in contrast to the fuzzless-lintless mutant ovules identified that the biological processes involving hydrogen peroxide homeostasis (Li et al., 2007) and pectin synthesis (Pang et al., 2010) are essential for cotton fiber elongation. Using a different lintless mutant as control, Zhao et al. (2010) identified 81 fiber proteins at 12 dpa involved in a series of cellular activities, and abundances of cytoskeleton-related proteins were found remarkably decreased in the short-lint phenotype. In a comparative analysis of fiber proteomes at five representative developmental stages (5-25 dpa), Yang et al. (2008) found 235 spots differentially expressed during fiber development; until recently, the protein identification of all these spots was completed, and serves as a first large-scale proteomic dataset for developing fibers (Zhang et al., 2013). By matching our 2-DE spot maps to this reference dataset, a total 184 spots were identified and also found differentially expressed during fiber development, which suggests that the majority of
fiber developmental dynamics is profiled by 2-DE approach with substantial consistency, even though 2-DE experiments were conducted in different labs using different cultivars of *G. hirsutum*.

In comparison to these 2-DE based studies, a much higher number of fiber proteins were identified and quantitatively profiled in our iTRAQ analysis. In addition to representing 70% of the characterized 2-DE proteomes, over a thousand proteins were uniquely identified by the iTRAQ approach, which provides a more comprehensive and valuable resource for studying the molecular mechanism of fiber elongation and cell wall synthesis. For example, differential expression of several transcription factors and nucleotide binding proteins were detected during fiber development; given that these proteins often function as upstream regulators, molecular analysis of their genomic target sites may provide new insights into the gene regulatory network of fiber elongation.

**Fiber proteomes altered by crop domestication**

Comparisons between wild and domesticated fibers revealed that the quantitative expression of 20-32% of the fiber proteome has been altered by domestication and crop improvement. This level of change is in agreement with that previously reported at the transcriptome level (23% of 40,430 genes) using the same model accessions of wild and domesticated *G. hirsutum* (Rapp *et al.*, 2010). In that previous study, the greatest differential gene expression was early in fiber development (2 dpa), so it was suggested that the gene regulatory network was rewired by early developmental events, with these changes propagating through the transcriptional network during subsequent developmental periods. Also, these key changes arose prior to obvious phenotypic
variance that can be observed in fiber length (Rapp et al., 2010). At the protein level, the highest amount of expression change was also observed at the earliest developmental stage sampled in our study (5 dpa). Interestingly, however, the amount of proteomic difference was maintained at a fairly constant level across later developmental stages according to the 2-DE data. This is in contrast to the general quantitative lowering of expression change and increased degree of synchronization between developmental trajectories of wild and domesticated cotton suggested by transcriptomic profiling and the proteomic profiling by iTRAQ. This discrepancy between 2-DE and iTRAQ results likely reflects the fact that they reveal different proteins and hence offer different windows into the evolutionary process, as discussed earlier. For those 2-DE proteins that remained stable during the latest fiber developmental stages studied, and which were differentially expressed in wild vs. domesticated cotton, their corresponding transcriptional signals may have been degraded or spatially separated from protein products due to numerous and variable regulatory processes of gene expression (Vogel & Marcotte, 2012).

With the objective of revealing how human-mediated selection reshapes the dynamic course of fiber development, a key finding from our study is that over twice as many expression changes were found in domesticated cotton than in the wild form in the period from 10 to 20 dpa (iTRAQ, 132 vs 45; 2-DE, 94 vs 44), in contrast to the other between-stage transitions studied. A previous observation associated with this phenomenon is that at the transcript level more genes were differentially expressed during fiber development in domesticated cotton than those in the wild form, implicating an increase in overall regulatory dynamics caused by human domestication (Rapp et al., 2010). Our results based on protein expression data offer an additional perspective on this
process and specifies a temporal component. Specifically, our results suggest that the most dramatic proteomic change coincide with period of most rapid fiber elongation in domesticated *G. hirsutum* (Applequist *et al.*, 2001). As a result, the differential expression we observed at the protein level is likely to more sensitively represent the biological processes that underlie phenotypic changes in domesticated cotton. Examples of how these data are useful for making functional interpretations are provided below.

**Functional interpretation of differentially expressed fiber proteins**

The fiber proteins that are differentially expressed are involved in various cellular activities, and about 20% of these function as metabolic enzymes in biosynthetic pathways of nucleotide, amino acid, fatty acid, lipid, and carbohydrate (Table 1, Figure 5), offering some clues into the developmental basis of fiber growth. For example, carbohydrate metabolism is known to be critical to cell wall growth for both energy production and to provide intermediates for cell wall synthesis, which is highly active during fast fiber elongation. One important enzyme in this category is sucrose synthase (SUS), which catalyzes a reversible reaction but preferentially converts sucrose into fructose and UDP-glucose. Two differentially expressed SUS proteins [114, Gorai.010G091800; 918, Gorai.010G092300] were identified by iTRAQ, displaying distinct expression profiles. As shown in Figure 5 (right panels), protein 114 was up-regulated from 10 dpa and displayed a higher expression level in the wild accession at 20 and 25 dpa, while protein 98 was highly expressed at 5 dpa and decreased later in TM1, in contrast to having static expression in Tx2094. Another Sus protein was represented by two 2-DE spot isoforms [Gorai.009G038000; spot ID - 563, 604], with expression at 10
dpa up-regulated in TM1 compared to Tx2094 (Figure 5). Although it has been shown that SUS plays a key role in cell expansion for both primary (Ruan et al., 2005) and secondary cell wall synthesis (Amor et al., 1995; Salnikov et al., 2003), multiple SUS gene family members are differentially expressed in a wide range of tissues and developmental processes in cotton (Chen et al., 2012) and it is not clear how a variety of SUS enzymes are temporally differentiated while function collectively in developing fibers. We can speculate that future exploration of these SUS homologs could provide clues to important regulatory machinery governing fiber growth.

Also functioning in hydrolysis of sucrose, a vacuolar invertase (VIN) [80, Gorai.008G216800] and a sucrase-like invertase [855, Gorai.012G064500] were identified with increased expression in domesticated cotton (Figure 5). For the same reason as SUS, VIN has long been considered as a major player in cell expansion. Wang et al. (Wang et al., 2010) showed that VIN activity is required for fiber elongation, which becomes evident in initiating fibers and remained high during their fast elongation and dropped when elongation slowed. Furthermore, a genotype with faster fiber elongation had significantly higher VIN activity than a slow-elongating genotype, which is well resembled by its protein expression pattern here, suggesting that up-regulation of VIN is an important mechanism involved in cotton domestication.

As an extension to our previous analysis of crop domestication in Pima cotton G. barbadense (Hu et al., 2013), the established proteomic database for G. hirsutum provides a rich source of clues for key regulatory changes that were commonly targeted by independent human-mediated selection in the two species under domestication, as well as differential gene regulation associated with fiber qualities unique to each species.
Taking the regulation of ribosomal proteins as an example, both domesticated examples of the two species exhibit peak expression of some ribosomal proteins during earlier stages, followed by gradual down-regulation, in contrast to lower levels of protein expression in their wild progenitors at 5 dpa. What makes this case interesting is the distinct expression patterns in the two wild accessions, observed as decreasing levels during fiber elongation in \textit{G. hirsutum} but with a peak expression at 10 dpa in \textit{G. barbadense}. It appears that the regulatory genetics of ribosomal proteins, which has been diversified between \textit{G. hirsutum} and \textit{G. barbadense} since speciation, was later re-shaped into uniformity by convergent evolution under strong fiber selection. Therefore, as evident by two independent domestication events, the elevation of ribosomal protein expression and their associated protein synthesis machinery is more likely to be linked to the domesticated fiber phenotypes, making its regulatory mechanism a good candidate for future functional analyses of crop improvement. On the other hand, several proteins are identified with altered expression by domestication but differentially regulated in two domesticated accessions, which are primarily involved in flavonoid biosynthesis and oxidation reduction. In recent study using interspecific backcrossed population from \textit{G. barbadense} and \textit{G. hirsutum}, regulation change of these functions were also suggested by QTL and eQTL analyses corresponding to different fiber traits (Chen, 2007).

\section*{Proteomics and crop domestication}

Since the time of Darwin (Darwin, 1859), biologists have understood the promise of crop plants and their wild relatives for providing insight into the mechanisms of phenotypic evolution. Over the past 20 years, there has been a concerted effort to identify
the causal mutations and underlying genetic architecture that control morphological transformations between wild crop ancestors and their modern descendants (Doebley et al., 2006; Burke et al., 2007; Burger et al., 2008; Gross & Olsen, 2010; Olsen & Wendel, 2013a; Olsen & Wendel, 2013b). As reviewed in Olsen and Wendel (2013a; 2013b), the increasing application of genome-scale systems biology approaches (e.g., genomics, transcriptomics, proteomics, metabolomics) promises to shed qualitatively new light on crop plant evolution. Our present study of the domestication process for upland cotton $G. hirsutum$, together with our previous work on Pima cotton $G. barbadense$, demonstrate the use of advanced proteomic profiling tools in an elite cotton cultivar and a wild accession to gain insight into cotton fiber development and evolution. The cataloged regulation of certain proteins and biological pathways provide clues for future functional analyses that may be valuable for both agronomic improvement and phenotypic evolution.

In the context of crop domestication, a wealth of indirect evidence suggests a role for polyploidization in generating adaptive plasticity and novel phenotypic variation for domestication-related traits (reviewed by (Paterson, 2005; Udall & Wendel, 2006). Given that all crop plants are polyploids, one promising application of evolutionary proteomics described here is to document the modification pattern of homoeolog-specific protein expression, which may lead to new evidence that link the genetics of gene duplication to functional and phenotypic significance.
Acknowledgements

We thank Kara Grupp and Anna Tuchin for help in tissue collection. We acknowledge the Protein and Proteomics Facility of the Iowa State University for technical assistance in 2-DE analysis. We acknowledge the Proteomics Division of the University of Florida’s Interdisciplinary Center for Biotechnology Research for assistance in LC-MS/MS analysis. The LC-MS/MS system was funded by National Institute of Health grant 1S10RR025418-01 to SC. This work was funded by Cotton Incorporated grant 09-558 and by the NSF Plant Genome Research Program, both to JFW.

References


Figures and tables

Fig. 1 Schematic illustration of the proteomic workflow.
**Fig. 2** Classification and biochemical properties of proteins expressed during fiber development. Compositions of Panther protein classes (a), theoretical pIs (b) and molecular masses (c) were compared between proteins identified from 2-DE (dark columns) and iTRAQ (light grey columns) analyses. *Significant difference, based on binomial tests.
**Fig. 3** Number of proteins differentially expressed during fiber development within and between wild and domesticated *Gossypium hirsutum*. Differential protein expression was independently analyzed for 1189 iTRAQ (blue) and 907 2-DE spots (brown). A representative image of a single seed with attached cotton fibers is shown for each accession. Numbers by the end of arrows denote the numbers of proteins up-regulated for the specified comparison, and numbers by the beginning of arrows denote numbers of down-regulated proteins.
**Fig. 4** Hierarchical clustering of developmental expression changes in wild and domesticated *G. hirsutum*. Log2 expression ratios between adjacent time points were analyzed for iTRAQ (a) and 2-DE (b) datasets. A bootstrapping of 10,000 iterations was performed to calculate bootstrap probabilities (BP) and approximately unbiased bootstrap probabilities (AU).

(a) 
(b)
**Fig. 5** Differentially expressed proteins in domesticated *G. hirsutum* relative to its wild progenitor. Differential expression patterns of 240 iTRAQ proteins and 64 2-DE spots with identification are separately clustered on the vertical axis of the heatmap. Expression ratios of domesticated versus wild were calculated and plotted on a log$_2$ scale for each time point. Up- and down-regulation are shown in green and red colors, respectively; black corresponds to no significant change. Based on the scheme of Bevan et al. (1998), functional category was assigned to each protein, whose corresponding row is marked black in the central grey columns. Examples of protein expression profiles are shown on the right.
**Fig. 6** Homoeolog-specific expression at the protein level. Boxplot of log2 expression ratios of TM1 vs. Tx2094 at each developmental time point, and between adjacent time points within each accession, are shown for: (a) Gorai.002G119600, (c) Gorai.004G108600, and (d) Gorai.004G031800. (b) shows protein sequences and homoeolog-specific peptides of Gorai.002G119600.
(a)

(b)

Amino acid sequences

>D_Gorai.002G119600.1
MAEEHTKVGGEEAVASKERGMFDGLKKEEKPQPQEEVVATQFEKVKI
EEEKEDEKKHSLDDLKLRNSNTSSSSSDDEEEDGGGEKKKKKKEKKKGK
KEEDSAVPVEKCDEAAATVHHSSETPEKKGMEKIKDKLPGQHKKDEEVTT
PPPPAAPPTENDHHEGETKEKDFLEKIEKIPGYHSKTEDEKEKETTAPH*

>A_Gorai.002G119600.1
MAEEHTKVGGEEAVASKERGMFDGLKKEDEKPQPQEEVVATQFEKV
IEEKEDEKKHSLDDLKHLNSSSDDDEEEEGGEKKKKEKKGK
KEQDSAQPVEKCDEAAATVHHSSETPEKKGMDKIKDKLPGQHKKDEEDT
TPPPAAPPTENDHHEGETKEKDFLEKIEKIPGYHSKTEDEKEKETTAPHP*
Table 1 Functional categorization of proteins differentially expressed during fiber development

<table>
<thead>
<tr>
<th>Functional category</th>
<th>Protein number (up)</th>
<th>iTRAQ 2-DE</th>
<th>2-DE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5-10 dpa</td>
<td>10-20 dpa</td>
<td></td>
</tr>
<tr>
<td></td>
<td>T Y T Y T Y T Y T Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolism</td>
<td>15 14 5 5</td>
<td>12 3 13 6</td>
<td>53(18.2%)</td>
</tr>
<tr>
<td>Energy</td>
<td>3 6 6 6</td>
<td>5 1 6 1 2</td>
<td>25(8.6%)</td>
</tr>
<tr>
<td>Cell growth and cytoskeleton</td>
<td>3 6 3 1 1</td>
<td>1 2</td>
<td>10(3.4%)</td>
</tr>
<tr>
<td>Transcription</td>
<td>1 6 2 2</td>
<td>1</td>
<td>7(2.4%)</td>
</tr>
<tr>
<td>Protein synthesis</td>
<td>3 44 25</td>
<td>1 21 4</td>
<td>53(18.2%)</td>
</tr>
<tr>
<td>Protein destination</td>
<td>4 5 3 9</td>
<td>5 1 7 1 4</td>
<td>30(10.3%)</td>
</tr>
<tr>
<td>Transporter</td>
<td>3 4 7 2 1</td>
<td>1 2</td>
<td>9(3.1%)</td>
</tr>
<tr>
<td>Intracellular traffic</td>
<td>1 1 1</td>
<td>1</td>
<td>2(0.7%)</td>
</tr>
<tr>
<td>Cell structure</td>
<td>2 3 4 3 2</td>
<td>2 2 1</td>
<td>9(3.1%)</td>
</tr>
<tr>
<td>Signal transduction</td>
<td>7 13 1 8 1 1 1</td>
<td>22(7.5%)</td>
<td>7(8.5%)</td>
</tr>
<tr>
<td>Defense and stress response</td>
<td>5 8 5 9</td>
<td>7 3 7 4 3</td>
<td>30(10.3%)</td>
</tr>
<tr>
<td>Secondary metabolism</td>
<td>2 6 3 2</td>
<td>2 2 4 3 2</td>
<td>11(3.8%)</td>
</tr>
<tr>
<td>Unclassified</td>
<td>7 11 7 10</td>
<td>7 6 5 1 2</td>
<td>31(10.6%)</td>
</tr>
</tbody>
</table>

SUM 292 82
Supporting Information

**Fig. S1** Images of 2-DE gels

<table>
<thead>
<tr>
<th>Time</th>
<th>TM1 (cultivated)</th>
<th>Tx2094 (wild)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 dpa</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
</tr>
<tr>
<td>10 dpa</td>
<td><img src="image6" alt="Image" /></td>
<td><img src="image7" alt="Image" /></td>
</tr>
<tr>
<td>20 dpa</td>
<td><img src="image11" alt="Image" /></td>
<td><img src="image12" alt="Image" /></td>
</tr>
<tr>
<td>25 dpa</td>
<td><img src="image16" alt="Image" /></td>
<td><img src="image17" alt="Image" /></td>
</tr>
</tbody>
</table>
Table S1 Protein identification and false discovery rate (FDR) analysis

a. Number of proteins identified at 95% confidence level

<table>
<thead>
<tr>
<th></th>
<th>run 1</th>
<th>run 2</th>
<th>run 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Spectra</td>
<td>68551</td>
<td>96346</td>
<td>94751</td>
<td></td>
</tr>
<tr>
<td>Spectra Identified</td>
<td>28925</td>
<td>37261</td>
<td>38990</td>
<td></td>
</tr>
<tr>
<td>(42.2%)</td>
<td>(38.7%)</td>
<td>(41.1%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distinct Peptides</td>
<td>12016</td>
<td>11923</td>
<td>12708</td>
<td></td>
</tr>
<tr>
<td>Proteins</td>
<td>1050</td>
<td>1085</td>
<td>1117</td>
<td>1199</td>
</tr>
</tbody>
</table>

b. Number of proteins estimated according to FDR analysis

<table>
<thead>
<tr>
<th>Critical FDR</th>
<th>Local FDR</th>
<th>Global FDR</th>
<th>Global FDR from Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0%</td>
<td>973</td>
<td>859</td>
<td>886</td>
</tr>
<tr>
<td>5.0%</td>
<td>**1012</td>
<td>1014</td>
<td>1046**</td>
</tr>
<tr>
<td>10.0%</td>
<td>**1050</td>
<td>1059</td>
<td>1075**</td>
</tr>
</tbody>
</table>

* Protein numbers from triplicated experiments were listed; numbers in bold are close to that from protein identification with above 95% confidence level.

Table S2 Normalized spot volumes of 2-DE gels and protein identification

Table S3 Protein identification and quantification by ProteinPilot iTRAQ analyses

Table S4 Significant expression changes and functional categorization

Table S5 homoeolog-specific expression
CHAPTER 5

GENE-EXPRESSION NOVELTY IN ALLOPOLYPLOID COTTON: A PROTEOMIC PERSPECTIVE

A paper in preparation for *Heredity*

Guanjing Hu, Jin Koh, Dharminder Pathak, Sixue Chen, and Jonathan F. Wendel

Abstract

Allopolyploidization is accompanied by changes in gene expression that are thought to contribute to phenotypic diversification. Here we describe global patterns of protein expression change in the single-celled, cotton fiber proteome of two natural allopolyploid species and living models of their diploid parents, using 2-DE and iTRAQ proteomic approaches in parallel. In total, approximately 3000 proteins were quantitatively profiled during fiber elongation. The amount of differential expression ranged from 4.4% to 12.8% between cotton species, with asymmetric differences between each allopolyploid and their diploid A- and D-genome progenitors. Over 80% of the fiber proteome in allopolyploid cotton were additively expressed. Proteomic expression level dominance was unbalanced toward the parental A genome in allopolyploid *G. hirsutum*, whereas the direction in *G. barbadense* switched from D genome in 10 dpa fibers to A genome in 20 dpa fibers. An unexpectedly high level of difference was found for the two allopolyploid species studied, where only about 1% of the significant protein changes were shared, in spite of their shared ancestry, relatively recent evolutionary divergence, and similar fiber phenotypes. In addition to profiling
total protein expression of homoeolog pairs, we diagnosed homoeolog-specific expression for 1001 proteins, and demonstrated a novel approach to assess homoeolog expression bias relative to the expression of parental proteins. Our study revealed that the global protein expression pattern has been altered and largely diversified in cotton allopolyploid species, thereby providing a proteomic perspective on understanding the evolutionary consequences of allopolyploidization and phenotypic diversification.

Introduction

Polyploidy is now recognized as a fundamental process in plant evolution, and all flowering plant genomes are known to have experienced several or more rounds of genome doubling in their evolutionary history (Jiao et al., 2011). Compared to the intraspecific genome duplication of autopolyploidy, the formation of allopolyploids entails the merger and doubling of diverged genomes, which has been proposed as an important mechanism of functional and phenotypic evolution driven by structural and regulatory divergence between parental genomes and the attendant duplication of genetic material (Wendel, 2000; Wendel & Doyle, 2004; Comai, 2005; Doyle et al., 2008; Leitch & Leitch, 2008; Soltis & Soltis, 2009; Finigan et al., 2012).

A large body of literature demonstrates that allopolyploidy is accompanied by a series of non-Mendelian interactions and processes, including chromosomal rearrangement and variation (Ramsey & Schemske, 2002; Szadkowski et al., 2010; Xiong et al., 2011; Chester et al., 2012), DNA sequence elimination (Shaked et al., 2001; Ozkan et al., 2003; Blanc & Wolfe, 2004; Han et al., 2005; Skalicka et al., 2005; Anssour et al., 2009; Buggs et al., 2009; Tate et al., 2009; Schnable et al., 2011),
epigenetic modification (Madlung et al., 2002; Salmon et al., 2005; Bottley et al., 2006; Chen, 2007; Gaeta et al., 2007; Kovarik et al., 2008; Ni et al., 2009) and differences in small RNA regulation (Kovarik et al., 2008; Ha et al., 2009). With respect to the transcriptome, a common observation is that the increase of genetic information in allopolyploids leads to various deviations from additivity, as well as a variety of transcriptomic re-patternings relative to parental species (Grover et al., 2012a). One well recognized phenomenon is “homoeolog expression bias”, characterized by unequal expression of homoeologs, particularly when the homoeolog expression ratio is unexpected given its state in the diploid progenitors (Adams et al., 2003; Bottley et al., 2006; Gaeta et al., 2007; Flagel et al., 2008; Hovav et al., 2008; Chaudhary et al., 2009; Buggs et al., 2010a; Buggs et al., 2010b; Koh et al., 2010; Yoo et al., 2013). A second prominent phenomenon observed for gene expression in allopolyploids derives from a consideration of the total expression level of a homoeolog pair in allopolyploids. When the total expression level in allopolyploids is compared to the mid-parent value of diploid progenitors, deviations are inferred as nonadditive gene expression (Hegarty et al., 2005; Wang et al., 2006; Gaeta et al., 2007; Flagel & Wendel, 2010; Chelaifa et al., 2013). A more nuanced view of duplicate gene expression, termed expression level dominance, was first described in cotton (Rapp et al., 2009) and then later reported in Spartina (Chelaifa et al., 2010), wheat (Chague et al., 2010) and coffee (Bardil et al., 2011), as well as in further studies in cotton (Yoo et al., 2012). This phenomenon is characterized by a form of duplicate gene co-regulation where the total expression of a homoeolog pair is similar to that exhibited by only one of the two diploid parents, irrespective of the direction of differential regulation with respect to the other parent (Grover et al., 2012a).
In general, the advent and subsequent widespread utilization of microarray and next-generation sequencing technologies has led to a rapid increase in exploration of gene expression at the transcriptional level. However, the transcriptome only is insufficient for understanding the end products of gene expression and phenotypic outcomes because of the relatively weak correlation between the transcriptome and the proteome (Rose et al., 2004; Thelen, 2007; Karr, 2008). As proteins are the major executor of cellular activities, the phenotype of an organism may arguably be more directly related to protein abundance and function than to transcriptional abundance (Karr, 2008; Diz et al., 2012). Thus, comparative proteomics offers an important perspective on evolutionary processes.

Only a few studies have examined the outcome of gene expression at the protein level in allopolyploids with respect to diploid progenitors. Using the classical two-dimensional gel electrophoresis approach (2-DE), interspecific comparison of protein presence and abundance was applied in wheat (Bahrman & Thiellement, 1987; Islam et al., 2003), *Brassica napus* (Albertin et al., 2006; Albertin et al., 2007) and cotton (Hu et al., 2011), which demonstrated examples of nonadditive expression at the protein level. In allopolyploid (AD genome) cotton seeds, a high percentage of proteomic variations were found among diploid and allopolyploid proteomes, with a biased accumulation of seed storage proteins towards one of the two (the D genome) progenitors (Hu et al., 2011). As an alternative and potentially more powerful proteomic approach (relative to 2-DE), a gel-free method employing isobaric tags for relative and absolute quantitation (iTRAQ) was used to investigate proteomic variations with respect to polyploidy in *Arabidopsis* (Ng et al., 2012) and *Tragopogon* (Koh et al., 2012). Both proteomic methods have several technical limitations, and the resulting proteomic information
derived from each is known to be complementary (Rose et al., 2004; Thelen, 2007). For example, only proteins within a narrow range of hydrophobicity, mass and isoelectric point (pI), can be resolved by 2-DE gels, and the protein identification in a spot-by-spot manner is time-consuming and often not available to associate the detected expression pattern to functional interpretation; in iTRAQ analysis, protein quantification is determined by comparing peptide pairs labeled with variable iTRAQ tags, which leaves the relative protein abundance within the same sample immeasurable. It is not clear, for interspecific comparisons in evolutionary studies, how these technical differences between the two approaches would affect resulted proteomic inferences, or whether one method is better suited for assessment of gene expression patterns in a manner comparable to those described for transcriptomic studies.

Our goal in the present study was to utilize both 2-DE and iTRAQ proteomic methods to investigate the effects of genome doubling on allopolyplid cotton proteomes. One key advantage of this system is that we could select a relatively simple structure for evolutionary analysis, the remarkable single-celled epidermal trichome colloquially known as cotton “fiber”. A second advantage is that we were able to simultaneously study two natural allopolyplid species, G. hirsutum (AD1) and G. barbadense (AD2), which diversified from the same polyploidy event approximately 1-2 million years ago (Grover et al., 2012b; Wendel, JF et al., 2012). These allopolyplids were studied in parallel relative to models of their diploid progenitors, the A-genome and D-genome species G. arboreum and G. raimondii, respectively. The resulting data offered us the opportunity to determine the proteomic consequences of polyploidization, and provide a rich database for offering insights into phenotypic diversification.
Materials and Methods

Plant materials

Four Gossypium accessions were used in the present study, including two natural alloployploids, G. hirsutum var. Acala Maxxa (AD1) and G. barbadense var. Pima S-6 (AD2), and models of their diploid progenitors, the A-genome species G. arboreum and the D-genome species G. raimondii (Wendel, J et al., 2012). Three to four plants from each accession were grown in the Bessey Hall Greenhouse at Iowa State University (Ames, IA, USA). Flowers were tagged at anthesis and developing bolls were harvested at 10 and 20 day post-anthesis (dpa), representing the key developmental stages (Kim & Triplett, 2001; Wilkins & Arpat, 2005; Haigler et al., 2012) of primary wall synthesis and elongation (10 dpa), and the transition to secondary wall synthesis (20 dpa). Harvested bolls were dissected immediately and ovules were frozen in liquid nitrogen and stored at -80°C. For each developmental stage, we used three biological replicates. Fiber proteins were extracted and purified as described in Hu et al. (2013).

2-DE and image analysis

The extracted proteins were dissolved in isoelectric focusing (IEF) buffer (8 M urea, 2 M thiourea, 4% [w/v] CHAPS, 2% [v/v] Triton X-100, and 50 mM DTT) at room temperature, and quantified using the 2-D Quant Kit (GE Healthcare Biosciences, Pittsburgh, PA, USA) with bovine serum albumin (BSA) as standard. For each 1mg of fiber protein, the final volume was adjusted to 450 µl with IEF buffer and then 2.25 µl of IPG buffer 3–10 NL (GE Healthcare Biosciences) was added. After centrifugation at 10,
000 x g for 10 min to remove insoluble materials, the supernatants were loaded on 24-cm 3-10 NL immobilized pH gradient (IPG) strips (GE Healthcare Biosciences) followed by rehydration for 90 min at room temperature. Using an IPGphor unit (GE Healthcare Biosciences), IEF was carried out with the following protocol: active rehydration at 50 V for 10 hr, 100 V for 100 V hr, 500 V for 500 V hr, and 8000 V for 99 kV hr at 20 °C, with a maximum current setting of 20 µA per strip. After completion of IEF, the strips were reduced for 20 min with 2.0% [w/v] DTT in equilibration buffer (6 M urea, 50 mM Tris-HCl, pH 8.8, 30% [v/v] glycerol, 2% [w/v] sodium dodecyl sulfate (SDS)), cysteine alkalized for 20 min with 2.5% [w/v] iodoacetamide in equilibration buffer, rinsed with SDS running buffer (1.5 M Tris-HCl, 6 M urea, 30% [v/v] glycerol, 5% [w/v] SDS), and then placed on 12% self-cast polyacrylamide gels. Second-dimensional electrophoresis was performed in an Ettan DALT six System (GE Healthcare Biosciences), using 80 V for 2 hr and 100 V for 16 hr. Finished gels were rinsed in deionized water and stained overnight with colloidal Coomassie staining solution (20% [v/v] ethanol, 1.6% [v/v] phosphoric acid, 8% [w/v] ammonium sulfate, 0.08% [w/v] Coomassie Brilliant Blue G-250).

Stained gels were scanned using ImageScanner (GE Healthcare Biosciences) at a resolution of 600 dpi and 16-bit grayscale pixel depth. Image and statistical analysis was conducted with the Progenesis SameSpots software version 4.0 (Non Linear Dynamics, Durham NC, USA), using standard procedures and default parameters. Briefly, spot detection, background subtraction, and normalization were performed following automatic alignment of all images with manual inspection. Detected spots were 100% matched across images, so that all gels contain the same number of spots without missing
values. Differential expression of spot volumes that represent protein abundances was assessed using the built-in statistical tool in the SameSpots software for contrasts of interest. Our criteria for significant changes were based on an ANOVA p-value < 0.05 as calculated with a fold change cutoff of > 1.2 or < 0.8.

iTRAQ LC-MS/MS, protein identification and quantification

The extracted proteins were dissolved in protein buffer (8M urea, 25mM triethylammonium bicarbonate (TEAB), 2% [v/v] TX-100, 0.1% SDS [w/v], pH 8.5), and prepared for iTRAQ labeling as described in Hu et al. 2013 (in press). For each sample, 100 µg of protein was reduced, alkylated, and trypsin-digested using the iTRAQ Reagents 8-plex Kit according to the manufacturer’s instructions (AB Sciex, Inc., Foster City, CA, USA). The diploid A-genome proteins were labeled with iTRAQ tags 113 (10 dpa) and 114 (20 dpa), D-genome proteins were labeled with tags 115 (10 dpa) and 116 (20 dpa), AD2 proteins were labeled with tags 117 (10 dpa) and 118 (20 dpa), and AD1 proteins were labeled with tags 119 (10 dpa) and 121 (10 dpa), respectively. After combining the labeled samples, the peptide mixture was fractionated with strong cation exchange chromatography, and analyzed using an off-line 2D LC-MS/MS method as previously described (Hu et al., 2013).

The MS/MS data were processed by a thorough database search considering biological modifications and amino acid substitutions under the Paragon™ algorithm (Shilov et al., 2007) and the Pro Group™ algorithm, using ProteinPilot version 4.5 software (AB Sciex, Inc.). Methylthio-cysteine and amine groups at the N-terminus and lysine were considered for the fixed modifications and variable modifications were
included for post-translational modifications (PTMs). The cutoff of protein identification was set to a confidence level of 95%. The global false discovery rate (FDR) of identified protein lists was determined by performing searches against the reversed protein databases, with estimates derived from both the conventional approach and a nonlinear fitting method (Tang et al., 2008) as shown in Supporting Information Table S1. The identified proteins were annotated with gene ontology categories using Blast2GO (Conesa et al., 2005), assigned to protein families using PANTHER (Mi et al., 2010) and functional classes based on the Arabidopsis functional catalog (Bevan et al., 1998).

For protein quantification, only MS/MS spectra that were uniquely identified for a particular protein were used to exact peak intensities of iTRAQ labeling tags, which were subsequently normalized across samples using the built-in bias correction function of ProteinPilot (AB Sciex, Inc.). Relative protein quantification between any two sample conditions and \( p \) values were determined using the software’s standard procedures. To be identified as being significantly differentially expressed, a protein must have been quantified with a fold change of > 1.2 or < 0.8 and \( p < 0.05 \) in at least two of the biological triplicates, along with a Fisher’s combined probability of <0.05 (Fisher, 1948).

iTRAQ data analysis and homoeolog-specific expression

Three protein databases were used for iTRAQ protein identification, including a non-redundant *Gossypium* protein database and two separate A- and D- genome diploid databases, as previously described (Hu et al., 2013). To comprehensively identify fiber proteins from multiple cotton species, the MS/MS data were first searched against the non-redundant *Gossypium* protein database. This step provided identification of protein
orthologs in A- and D- genome diploids, and their counterparts in allopolyploids representing the combined expression of homoeolog pairs \((A_T \text{ and } D_T, \text{ where the subscript indicates the specific genome in the allopolyploid})\). The derived data were used to analyze the total expression level of homoeolog pairs for a given protein in allopolyploids relative to expression levels in the parental diploids, in parallel to 2-DE analyses.

To distinguish the expression patterns between homoeologs in allopolyploids, or homoeolog-specific expression, the MS/MS data were subsequently searched against separate diploid A- and D- genome databases using the same ProteinPilot parameters described above. Based on the amino acid difference between protein orthologs in the diploid protein databases, A- or D- specific peptides were identified by comparisons using the diploid databases. For a A- or D- specific protein, expression differences between sample conditions were considered statistically significant if at least three peptides specific to the corresponding diploid genome were identified from at least two of the replicates and quantified with \(p < 0.05\) via Student’s t-test.

**Additivity test and categorization of differential expression patterns**

The hypothesis of additive parental expression in the allopolyploid was tested using 2-DE data only, where a spot was considered additive if its expression in the allopolyploid genome was statistically equivalent to the average values expressed in the parental A and D genomes. Co-electrophoresis of 1:1 A- and D-genome protein extracts was conducted to obtain average values of the diploid parents. To test the reliability of the experiment, co-electrophoretic values were compared to the mean of parental values
by student $t$-test ($p < 0.05$). According to the contrast between each allopolyploid and co-electrophoresis, any allopolyploid expression significantly deviated from the co-electrophoretic value was considered nonadditive, in the form of up- or down-regulation relative to the average parental values.

Proteins in allopolyploids were categorized into six possible patterns based on their expression levels relative to the diploid parents (see Figure 4). First, a conserved expression was inferred (category VI) if the expression level in the allopolyploid was statistically equivalent to that in both diploid parents, whereas if it was equivalent to the value of only one parent, expression level dominance was inferred in the direction of this parent (I and II), irrespective of whether their expressions were up- or down-regulated relative to the other parent, according to Grover et al. (2012). For protein expression levels that were significantly different from both parental values, median (III) or transgressive expression (IV and V) was inferred depending on whether the expression in the allopolyploid statistically falls between or outside the range of the two diploid parents.

**Other statistical analysis**

Hierarchical clustering with regular bootstrap (BP) and approximately unbiased bootstrap (AU) $p$-values was performed using the R software package pvclust (Suzuki and Shimodaira, 2006), specifying average linkage and Pearson’s correlation distance metric with 10,000 iterations. The bootstrap $p$-values indicated how strongly the cluster is supported by the data. For example, for a cluster with $p > 95\%$, the hypothesis that `the
cluster does not exist' is rejected with a significance level of 5%. The branch length represents the degree of dissimilarity in protein expression among species.

Results

Parallel proteomic strategies applied to profile cotton fiber proteins

The proteomic workflow is schematically depicted in Figure 1, illustrating the fiber developmental stages studied, application of 2-DE and iTRAQ techniques, and MS and associated bioinformatics analyses for establishing a comparative fiber proteome database. Total proteins extracted from 10 and 20 dpa fibers were examined in two natural allopolyploids, *G. hirsutum* var. Acala Maxxa (AD1) and *G. barbadense* var. Pima S-6 (AD2), and representatives of their diploid progenitors *G. arboreum* (A) and *G. raimondii* (D), resulting in a total of eight sample conditions subsequently subjected to 2-DE and iTRAQ experiments in parallel. For 2-DE analyses, we also conducted co-electrophoresis of A- and D-genome protein extracts (1:1 mix) to obtain the average expression pattern of parental diploids. With an isoelectric focusing range of pH 3-10, a total of 1323 2-DE spots were detected and reproducibly cross-matched in all analyzed gels, where the profiled spot volumes represented fiber proteins being expressed over a range of at least 3 orders of magnitude (Supporting information Figure S1, Table S2).

Using iTRAQ labeling coupled with 2D LC-MS/MS experiments, a total of 1652 fiber proteins were identified with their relative expression changes measured simultaneously across sample conditions (*identification confidence > 95%, FDR < 1%*; Table S1 and S3), among which homoeolog-specific peptides were diagnosed for 1001 proteins in allopolyploid species, accounting for 60% of the fiber proteome characterized.
The proteins identified by iTRAQ covered almost all functional families encoded in the *Gossypium* genome, with over a third of the proteins classified into oxidoreductase (12.4%), nucleic acid binding protein (11.5%) and hydrolase (11.0%) (Figure S2a).

Relative to the whole *Gossypium* proteome, protein families over-represented in developing fibers included oxidoreductase, lyase, isomerase, chaperone, cytoskeleton protein, protease, kinase, transfer/carrier protein, and calcium-binding protein, while under-represented families were ligase, hydrolase, phosphatase, transferase, transporter, transcription factor and nucleic acid binding protein (Figure S2b).

**Differential protein expression during fiber development**

Because two developmental time points (10 and 20 dpa) were used to profile fiber protein expression, we first assessed the proteomic change associated with fiber elongation within each genotype, which revealed that 4.1-9.9% of the resolved proteins were significantly differentially expressed from 10 to 20 dpa (Figure 2). Of these, 2-DE and iTRAQ methods both identified about 7% significant developmental change in the D genome fibers, whereas different estimates (p < 0.05, Fisher’s exact test) were obtained for the two proteomic techniques for the other three species studied: fiber proteins in A and AD1 displayed a higher level of developmental variation profiled by 2-DE (8.5% and 7.6%) than by iTRAQ (4.1% and 4.1%), whereas iTRAQ identified more differentially expressed proteins in AD2 than did the 2-DE approach (9.9% vs 4.5%).

Notably, only approximately 1% of the developmental changes during fiber elongation in the two allopolyploid species were shared. This observation suggests that
the proteomic changes during fiber development may vary substantially among different cotton species that trace to the same allopolyploid event (Table S4, S5).

To compare the proteomic patterns accompanying fiber development between diploid and polyploid cottons, expression ratios of 20 to 10 dpa of all proteins were considered for pairwise concordance tests (Figure S3) and clustering analysis (Figure 2). For expression ratios profiled by the 2-DE approach, the only significant correlation ($r_{\text{Pearson}} = 0.11$, $p < 0.01$) was detected between the diploid A and allopolyploid AD1 genomes. For iTRAQ data, all pairwise correlations were statistically significant ($p < 0.01$), except that between the A genome and the allopolyploid AD2 ($r_{\text{Pearson}} = 0.06$, $p = 0.04$). While the highest correlation coefficient ($r_{\text{Pearson}} = 0.50$) was found between A and AD1, the diploid D genome was negatively correlated with AD2 ($r_{\text{Pearson}} = -0.29$). As shown in Figure 2, cluster analyses using both 2-DE and iTRAQ data sets revealed a definite cluster of A and AD1 genomes (over 95% bootstrap values). These results suggest that the protein-level developmental change in developing fibers from the allopolyploid AD1 species *G. hirsutum* is more similar to that of the diploid A-genome progenitor *G. arboreum*, compared to fibers from the other allopolyploid AD2 (*G. barbadense*).

Proteomic variations in diploid and allopolyploid cottons

To explore the variability of protein expression patterns in diploid and allopolyploid cotton fibers, we directly compared proteomes of different species at the same developmental time point. As shown in Figure 3 a and b, 2-DE profiling revealed 183 (13.8%) and 194 (10.4%) proteins differentially expressed between the diploid
parents at 10 and 20 dpa, respectively, exhibiting a higher level of proteomic variation compared to that between the two allopolyploid species, where 108 (8.2%) and 97 (7.3%) proteins were differentially expressed at 10 and 20 dpa, respectively (p < 0.05, Fisher’s exact test). In comparisons between allopolyploids and their two parents, the percentages of proteins showing differential expression at 10 dpa were statistically equivalent for both allopolyploids (11.3-12.8%; Figure 3a); at 20 dpa, AD1 exhibited a higher level of differential expression compared to the D progenitor than to the A progenitor (12.8% vs 9.4%; p < 0.05, Fisher’s exact test), whereas the differential expression between AD2 and the two progenitors was symmetric (8.5%; Figure 3b).

Proteomic variation resulting from iTRAQ analyses are shown in Figure 3c and d, with some apparent expression differences relative to those observed for the 2-DE results (panels a, b). For example, more differentially expressed proteins were found between parental diploids and between the two allopolyploids at 20 dpa (11.8% vs 6.8%; p < 0.05, Fisher’s exact test), while the percentages of differential expression were statistically equivalent at 10 dpa (8.6% vs 9.2%). For both allopolyploids, the amount of differential expression relative to their diploid parents was asymmetric (p < 0.05; Fisher’s exact test). That is, the global expression pattern of allopolyploids was biased towards the diploid parent with fewer differentially expressed proteins. The proteomic expression of allopolyploid AD1 was closer to the A progenitor than to the D progenitor at both developmental time points. Interestingly, the directional bias switched for AD2; that is, the expression pattern more closely mirrored the D progenitor at 10 dpa, but later switched toward the A genome progenitor at 20 dpa.
Additivity test and expression level dominance in allopolyploid cotton

As a prelude to test the additivity of parental contributions to the allopolyploid proteome, co-electrophoreses of 1:1 A- and D-genome proteins were compared to the mean of their independent 2-DE values: only 3 spots were significantly variable at 10 dpa while no difference was found at 20 dpa, much lower than expected by chance at a 1% false discovery rate (13 out of 1323 spots). These results allowed us to use co-electrophoretic patterns as control to detect spots in allopolyploids that deviated significantly in abundance from the average of the parental diploids, i.e. non-additive protein expression. As shown in Figure 3 e and f, approximately 13% of the spots were detected as nonadditive in AD1 at both time points, while the percentage of nonadditive spots in AD2 decreased from 10.6% at 10 dpa to 7.6% at 20 dpa; these nonadditive expressions were equally distributed in the direction of up- and down-regulation relative to the parental average (p > 0.05, Fisher’s exact test).

Regardless of protein expression relative to the average of parental diploids, categorization of expression patterns in allopolyploids was applied to 2-DE and iTRAQ datasets in parallel, according to differential expression relative to each of the diploid parents. Thus, we binned proteins into six possible categories as shown in Figure 4. The majority of fiber proteins, accounting for more than 70% of the allopolyploid proteomes, displayed conserved expression in comparison with the expression levels in both parents (category VI), while less than 1% of proteins displayed expression values that were intermediate to but statistically different from those of both parents when the two parents exhibited differential expression (category III). Expression level dominance of either parental genome (categories I and II) was evident for 4.8-7.8% of the proteins expressed
in allopolyploid fibers, when statistical equivalence of expression levels is diagnosed between the allopolyploid and only one of its diploid progenitors. According to both 2-DE and iTRAQ datasets, AD1 exhibited higher amounts of expression level dominance in the direction of the A genome than of the D genome (p < 0.05, Fisher’s exact test), or, following Grover et al. (2012), we can say that the expression level dominance displayed by AD1 was unbalanced toward the A-genome parent. Likewise, unbalanced expression level dominance of AD2 was suggested by iTRAQ data only, the direction of which, however, was toward the D genome at 10 dpa but switched to the A-genome at 20 dpa.

Novel expression patterns of transgressive up- (category V) or down-regulation (category IV) were identified for 0.8-3.8% of the proteins in allopolyploid fibers. The remaining proteins that were statistically excluded from these categories were grouped together (category VII), accounting for a higher percentage of 2-DE spots than of iTRAQ proteins, which may reflect the variable sensitivities of differential expression analysis of these two methods due to their technical differences.

**Homoeolog-specific expression in allopolyploid cotton**

Among 1001 proteins identified with genome-diagnostic peptides with respect to the diploid parental genomes, 423 proteins in allopolyploids had diagnostic peptides detected for both homoeologs (A\textsubscript{T} and D\textsubscript{T}) in the fiber proteomes. For each identified peptide, iTRAQ enabled relative quantification of expression levels in fiber proteins from diploid and polyploid cotton species (Table S6). For example, a D-specific protein can be detected in allopolyploid and the diploid D-genome fibers, with their quantitative comparison represented by expression level of a D\textsubscript{T} homoeolog in allopolyploid relative
to that of the parental protein in the diploid D-genome (that is, D<sub>T</sub>/D); likewise, the A<sub>T</sub>/A ratios can be calculated for a A-specific protein. Because an A- or D- genome specific peptide, in principle, doesn’t exist in the other diploid genome, any false-positive expression signals detected (measured as A<sub>false</sub>/D or D<sub>false</sub>/A) are possibly caused by background effect of mass spectrometry and/or inaccurate A-D amino acid differences present in our diploid protein database. Thus, any proteins displaying problematic signals were excluded from further analysis by applying criteria A<sub>false</sub>/D<1 and D<sub>false</sub>/A < 1 (p < 0.05, Student’s t-test).

As shown in Figure 5 (heatmap), a total of 34 proteins were identified with significant changes of homoeolog-specific expression relative to that of the parental gene in diploid species in at least one of the allopolyploid species, as expressed by A<sub>T</sub>/A and D<sub>T</sub>/D. Notably, down-regulation (red on heatmap) was mostly observed for homoeologs relative to their parental proteins, where less than 10 cases of up-regulation (green on heatmap) were shown. Co-analysis with expression pattern categories revealed that for over 70% of these proteins surveyed, their total expression of both homoeologs fell into the category of “Conserved” expression as characterized in Figure 4, suggesting that the total expression in allopolyploids (sum of A<sub>T</sub> and D<sub>T</sub>) is statistically equivalent to their parental proteins in A and D genomes (Figure 5, light grey color in right columns). Accordingly, the total abundance of protein expression in allopolyploids is maintained mainly through down-regulation of both homoeologs irrespective of any regulatory change on the relative contribution of homoeologs (i.e. A<sub>T</sub>/D<sub>T</sub> versus A/D); however, up-regulation of one homoeolog with respect to its parental protein while down-regulating the other (e.g. A<sub>T</sub>/A <1, D<sub>T</sub>/D >1), clearly suggests biased expression of the up-regulated
homoeolog, given conserved parental expression of origin (i.e., if A/D=1, A_T/D_T <1). One example of such biased regulation of homoeolog expression was evident for a class III peroxidase [Gorai.001G259400], which exhibited down-regulation in A_T/A coupled with mostly up-regulation in D_T/D (indicated by * beside the heatmap).

Discussion

Here we have presented a multi-level comparative approach to investigate modifications of gene expression at the proteome level, in a single celled structure, caused by allopolyploidization. With the enhanced analytical power provided by using both 2-DE and the gel-free iTRAQ methods for proteomic profiling, two natural allopolyploid cotton species were studied in parallel. By demonstrating the protein level consequences of gene expression evolution, such as expression level dominance and homoeolog expression bias, our results provide a broad view of the proteomic variations with respect to developmental dynamics, evolutionary divergence and genomic constitution in allopolyploid cotton relative to its parental diploid species.

Complementary application of 2-DE and iTRAQ approaches

Our 2-DE and iTRAQ analyses each present a quantitative and interspecific profile of fiber proteomes that include over one thousand protein features. Comparison of the parental G. arboreum and G. raimondii proteomes revealed that approximately 10% of the proteins were differentially expressed between the two diploid species, which have diverged for ~5-10 million years during which time they have accumulated ~3.6% synonymous and 0.9% non-synonymous nucleotide (Flagel et al., 2012) and ~1.8% of
amino acid differences between orthologous genes (Hu et al., 2013). The proteomic divergence revealed by the profiling experiments reported here is much lower than that previously observed for cotton seeds, where only 50% of the protein spots were qualitatively shared by the two diploids (Hu et al., 2011). In Tragopogon, 3.2% of differential protein expression was reported for leaf proteomes between the diploid T. dubius and T. porrifolius, which are represented with a 3.5% synonymous nucleotide divergence (Koh et al., 2012). These results suggest that protein expression differences vary largely across tissue types and are hard to predict from genetic divergence alone, especially when comparisons are made across tissue or organ types that vary in their cellular complexity. Compared to differences between the two diploids, fewer differential expression changes were identified between two allopolyploid cotton species – G. hirsutum and G. barbadense according to 2-DE data, which is in agreement with their closer evolutionary relationship (1-2 million years of divergence, compared to 5-10 myrs for the diploids), whereas this pattern was not shown by iTRAQ. Although mostly consistent results were derived from 2-DE and iTRAQ data in our analysis, our data demonstrate that the two platforms offer slightly different and hence complementary perspectives on proteomic evolution.

In 2-DE analysis, different proteomes are resolved into two-dimensional spot maps for quantification with the assumption that each spot represents one protein. However, due to protein microheterogeneity caused by post-translational modifications, proteolytic degradation and other causes, protein products from one gene may exhibit multiple different spots, thereby affecting quantitative sensitivity and accuracy (Wu et al., 2006). Besides protein isoforms, protein co-migration and in silico co-migration due to
incorrect spot matching across gels imposes an even greater challenge in quantification, especially for evolutionary studies when proteomes from different species are included for comparative analyses. As previously reported for cotton seeds, although compositions of major seed storage proteins were relatively conserved across species, their numerous isoforms contributed up to 50% of the interspecific variation (Hu et al., 2011). Thus, the protein expression change resulted from 2-DE analyses in fact reflects confounding effects of both protein abundance and post-translational modifications, which on the other hand, provides us a unique opportunity to investigate evolutionary processes that modify protein-level-specific properties. In contrary to that, protein microheterogeneity is less problematic in iTRAQ analysis, where quantification is assessed at the peptide level.

In the iTRAQ platform, labeling and pooling samples into one analysis has one key advantage over single-stain 2-DE gels, by reducing experimental variation and raising confidence in quantitative assessments of protein changes. A second and critical advantage, with respect to the study of allopolyploids, is that with appropriate data from reference genomes, it is possible to derive information on homoeolog-specific protein expression. The first application of iTRAQ for this purpose was described recently for allopolyploid *Tragopogon* (Koh et al., 2012): by specifying the genome-specific protein sequences between diploid progenitors as well as sub-genomes in allopolyploids, genome-specific peptides for a given pair of homoeolog proteins were examined for differential expression, which identified two cases of biased expression of a parental homoeolog in the natural allopolyploid *T. mirus* but not in the F1 hybrid or synthetic polyploid. Since that initial report, homoeolog expression bias has been discovered in developing fiber proteomes of two allopolyploid cottons, *G. barbadense* (Hu et al., 2013)
and *G. hirsutum* (Chapter 4). In the present study, we demonstrate a quantitative method to examine homoeolog-specific expression patterns, which led to a comparative profile of homoeolog expression for 34 proteins with respect to their parental proteins in diploids as well as to the corresponding total expression pattern in allopolyploid proteomes.

A potential pitfall of iTRAQ analysis results from the same factors that make iTRAQ approach so efficient in the first place, i.e. relative protein quantification dependent on the ratio of iTRAQ reagents. For a given protein, pairwise sample ratios instead of original peptide intensities are extracted and subjected to differential expression and other statistical analysis, which is less flexible compared to other methods utilizing direct measurements for high throughput quantification, e.g. spot volumes in 2-DE analysis and read numbers in RNA-seq analysis, and can affect cross-platform analysis in conjunction with these methods due to incompatible data structures. For example, the iTRAQ data in our analysis were not used for additivity test, whereas the 2-DE spot volumes of diploid parents can be used to estimate the average of parental expression. Although additivity test based on iTRAQ ratios was previously reported in *Arabidopsis* (Ng *et al.*, 2012), we argue that the use of in silico parental averages is difficult to justify without validation using values derived from mixed sample of parental proteins.

Taken together, 2-DE and iTRAQ methods have been not only provided complementary information for comprehensive proteomic profiling (Rose *et al.*, 2004; Thelen, 2007; Diz *et al.*, 2012), but also each exhibits different strengths, underscoring the importance of technical diversity in revealing the complexity of biological systems and evolutionary modifications.
Nonadditive expression and unbalanced expression level dominance in fiber proteomes

The profiling data demonstrate that 7-13% of the fiber proteome in allopolyploid cotton species is differentially expressed relative to the average of its two parental diploids. This offers one perspective on the pace and scope of proteomic evolution accompanying diploid divergence and subsequent genomic merger and doubling. A corollary is that the majority of fiber proteins are additively expressed. The degree of proteomic divergence for cotton fibers is lower than that previously observed in Brassica root and leaf proteomes (25–38%) (Albertin et al., 2006) and for the seed proteome in cotton (34%) (Hu et al., 2011).

The term expression level dominance, modified from genomic expression dominance, was suggested by Grover et al. (2012) to describe the phenomenon where the total expression level in an allopolyploid resembles that of one of the two parents. As first described in cotton for the leaf transcriptome by Rapp et al. (2009), microarray profiled gene expression levels in a synthetic allopolyploid genome mimicked those in the parental D genome more often than those in the A genome. Using a more sensitive transcriptomic approach (RNA-seq), a similar pattern of unbalanced expression level dominance was demonstrated in synthetic allopolyploid cotton, while the direction was reversed in an F1 diploid hybrid and two natural allopolyploids, which favored the parental A genome (Yoo et al., 2013). At the protein level, the seed proteome in cotton was characterized for one allopolyploid species using 2-DE method, where 16% more 2-DE spots in G. hirsutum proteome were shared with those in the D parent than in the A parent, and more proteins exhibited expression level dominance favoring the D than the
A genome. Here we complement these earlier studies with an expanded experimental design using two different proteomic approaches as well as two different allopolyploid species. A consensus observation of the present work is unbalanced expression level dominance for 10 dpa fibers in the direction of the A genome, which leads to the suggestion that there is rapid change during fiber development in a genomically biased fashion. The functional significance and underlying regulatory bases of this observation may be fruitful avenues for future exploration.

**Variable proteomic alteration following allopolyplolidization between G. hirsutum and G. barbadense**

The two natural allopolyploid species *G. hirsutum* and *G. barbadense* represent divergent branches from a monophyletic (Grover et al., 2012b) allopolyploidization event that occurred 1-2 mya (Wendel & Cronn, 2003; Wendel, JF et al., 2012). With both allopolyploids exhibiting the notable morphological trait of long spinable cotton fibers, having been derived from independent domestication events, it has long been hypothesized that allopolyploid cotton preferentially displays the physiological profile and molecular machinery from its A-genome progenitor, because it is in this genomic group that long fiber first evolved. Fiber elongation rate is dramatically higher in *G. hirsutum* and the A-genome diploids compared to fibers from the D-genome diploid *G. raimondii*, especially during primary cell wall synthesis from 10 to 20 dpa (Applequist et al., 2001). Although the most advanced modern cultivars of *G. barbadense* produce longer fibers than *G. hirsutum* cultivars at maturity, their growth curves generally overlap (Chen et al., 2012). The developmental divergence observed here between these two
species at the proteome level, represented by 5-10% of fiber proteins, is consistent with our previous work describing developing fiber proteomes in *G hirsutum* (see Chapter 4) and *G. barbadense* (Hu et al., 2013). A notable observation, and one that bears future study, is that there is little overlap in protein change during development in the two species (17 iTRAQ proteins and 4 2-DE spots). A second intriguing observation is that the fiber proteome of *G. hirsutum* resembles the parental A genome more closely than does the fiber proteome of *G. barbadense* (Figure 2, Figure S3). These results suggest that following allopolyploidization and subsequent independent domestication, the two allopolyploid species have achieved comparable fiber phenotypes through fairly different evolutionary roads at the proteome level.

**Conclusions**

Gene expression evolution accompanying allopolyploidization, including expression level dominance and homoeolog expression bias, has been acknowledged and widely studied at the transcriptomic level, but rarely evaluated at the protein level. Only a handful of species (*Brassica, Tragopogon, Arabidopsis* and cotton) have been examined to date, and each using only a single proteomic technique. Here we demonstrate the protein level consequences of gene regulation using two complementary proteomic strategies, which we apply to a readily harvested single-celled structure. This enabled a more comprehensive documentation of fiber proteomes in response to evolutionary change. Two natural cotton allopolyploids of the same origin were used to study the general consequences of polyploidization, which led to a consensus discovery of A-biased expression level dominance at the same fiber elongation stage. Interestingly, highly variable paths of proteomic modification were exhibited by the two allopolyploids
given their morphological similarity. One intriguing dimension of our study is that both allopolyplloid species were studied using modern cultivated lines; future work will reveal whether the proteomic distinctions between the two species reflect the effects of strong human-mediated directional selection that unknowingly targeted different components of the proteomic network, or if instead this happened during evolutionary divergence prior to domestication. Future exploration that includes wild accessions and various landraces will provide a powerful framework to identify and characterize the key protein and metabolic pathways corresponding to species evolution and phenotypic diversity in cotton.

Acknowledgements

We thank Kara Grupp and Anna Tuchin for help in tissue collection. We acknowledge the Protein and Proteomics Facility of Iowa State University for technical assistance in 2-DE analysis. We acknowledge the Proteomics Division of the University of Florida’s Interdisciplinary Center for Biotechnology Research for assistance in LC-MS/MS analysis. The LC-MS/MS system was funded by National Institute of Health grant 1S10RR025418-01 to SC. This work was funded by Cotton Incorporated grant 09-558 and by the NSF Plant Genome Research Program, both to JFW.

Reference


Advances in Agronomy: Academic Press, 139-186.


Figures and tables

**Fig. 1** Schematic illustration of the proteomic workflow. Parallel strategies of 2-DE and iTRAQ analyses were used for comparative quantification of fiber proteomes at 10 and 20 dpa from four cotton species. Proteins extracted from each sample condition were separated with a non-linear IEF range of pH 3-10 in triplicated 2-DE experiments, and the resulting gel images were aligned for spot detection and protein quantification with normalized spot volumes. In iTRAQ analyses, proteins extracted from eight sample conditions were separately digested and labeled with iTRAQ tags, and the combined peptide mixture was subjected to liquid chromatography (LC) coupled with tandem MS analyses. Three independent iTRAQ analyses were conducted for three replicated sets of protein samples. The iTRAQ-LC-MS/MS-acquired data were searched against three cotton databases for protein and homoeolog identification.
Developing Cotton fibers

Total Protein by Phenol Extraction Method

2-DE

iTRAQ

Quantitation by Spot Volumes

Quantitation by Tag Intensities

Mass Spectrometry Analysis
Hybrid quadrupole Time-Of-Flight ABI QSTAR elite MS system

Database Search & Bioinformatic Analysis
Protein ID by Gossypium protein database
Homoeolog ID by separate diploid databases

Comparative Cotton Proteome Database
Fig. 2 Hierarchical clustering of diploid and allopolyploid cotton according to developmental expression changes. Log2 expression ratios of 20 dpa versus 10 dpa were analyzed for 2-DE (a) and iTRAQ (b) datasets. Each branch represents a cotton species (A = G. arboreum; D = G. raimondii; AD1 = G. hirsutum; AD2 = G. barbadense) labeled with the number and percentage of significant protein changes. A bootstrapping of 10,000 iterations was performed to calculate bootstrap probabilities (BP) and approximately unbiased bootstrap probabilities (AU).
**Fig. 3** Number of proteins differentially expressed among diploid and allopolyploid species. Bold text indicates the total number and percentage of proteins differentially expressed in each comparison. Beside bold text, the total number of proteins is partitioned into both directions; e.g., of the 153 proteins differentially expressed between *G. hirsutum* and *G. arboreum* in panel (a), 69 and 84 were up-regulated in the former and latter species, respectively. For 2-DE comparisons at 10 (a) and 20 dpa (b), percentages were calculated based on a total of 1323 spots detected on gels. For iTRAQ comparisons at 10 (c) and 20 dpa (d), percentages were calculated based on 1652 non-redundant proteins. Using co-electrophoreses as control, additivity of protein expression in allopolyploids was tested at 10 (e) and 20 dpa (f) for 2-DE data.
**Fig. 4** Categorization of protein expression in allopolyploid proteomes at each developmental stage. Possible patterns of allopolyploid (AD) expression relative to their diploid parents (A and D) were tabulated for each category. Results based on 2-DE and iTRAQ datasets are presented in pairs. For example, for the developmental stage 20 dpa within *G. barbadense* (AD2), 33 and 82 proteins estimated by 2-DE and iTRAQ analyses, respectively, fall into the expression category I, which account for 2.5% of 1323 2-DE spots and 5.0% of 1652 iTRAQ proteins profiled. Significant differences between iTRAQ and 2-DE results are indicated using grey shadowing (p < 0.05, fisher’s exact test).

<table>
<thead>
<tr>
<th>Categories</th>
<th>I. A-expression level dominance</th>
<th>II. D-expression level dominance</th>
<th>III. Median expression</th>
<th>IV. Transgressive up-regulation</th>
<th>V. Transgressive down-regulation</th>
<th>VI. Conserved expression</th>
<th>VII. Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD1: 10 dpa</td>
<td>55 / 55 4.2% / 3.3%</td>
<td>32 / 27 2.4% / 1.6%</td>
<td>2 / 15 0.2% / 0.9%</td>
<td>24 / 5 1.8% / 0.3%</td>
<td>23 / 9 1.7% / 0.5%</td>
<td>968 / 1438 73.2% / 87.0%</td>
<td>219 / 103 16.6% / 6.2%</td>
</tr>
<tr>
<td>AD1: 20 dpa</td>
<td>39 / 100 2.9% / 6.1%</td>
<td>25 / 27 1.9% / 1.6%</td>
<td>7 / 5 0.5% / 0.3%</td>
<td>23 / 21 1.7% / 1.3%</td>
<td>14 / 11 1.1% / 0.7%</td>
<td>1009 / 1366 76.3% / 82.7%</td>
<td>206 / 122 15.6% / 7.4%</td>
</tr>
<tr>
<td>AD2: 10 dpa</td>
<td>33 / 25 2.5% / 1.5%</td>
<td>44 / 63 3.3% / 3.8%</td>
<td>7 / 11 0.5% / 0.7%</td>
<td>17 / 36 1.3% / 2.2%</td>
<td>33 / 8 2.3% / 0.3%</td>
<td>969 / 1360 73.2% / 82.3%</td>
<td>220 / 149 16.6% / 9.0%</td>
</tr>
<tr>
<td>AD2: 20 dpa</td>
<td>33 / 82 2.5% / 5.0%</td>
<td>33 / 46 2.5% / 2.8%</td>
<td>6 / 14 0.5% / 0.8%</td>
<td>10 / 16 0.8% / 1.0%</td>
<td>15 / 15 1.1% / 0.9%</td>
<td>1062 / 1341 80.3% / 81.2%</td>
<td>164 / 138 12.4% / 8.4%</td>
</tr>
</tbody>
</table>
Fig. 5 Co-analysis of homoeolog expression and total protein expression categories in allopolyploid cotton. (a) For 34 proteins with genome-diagnostic peptides for both diploid genomes, expression ratios of $A_T/A$ and $D_T/D$ were plotted in a heatmap on a log2 scale. Up- and down-regulation are shown in green and red colors, respectively; black color corresponds to no significant change. For a given protein represented by a row, its assigned category for total expression of both homoeologs is shown on the right columns. * Gorai.001G259400, class III peroxidase.
Supporting Information

**Fig. S1** Images of 2-DE gels
Fig. S2 Panther protein classification. (a) Classification of the fiber proteins. (b) Over- and under-representation of protein categories. Fractional differences between identified fiber proteins and encoded *Gossypium* proteins were calculated for each category.

*Significant difference, based on binomial tests with bonferroni correction.
(b)

| Protein Category                  | Percentage | Percentage | Percentage | Percentage | Percentage | Percentage | Percentage | Percentage | Percentage | Percentage |
|-----------------------------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| * oxidoreductase                  | 3.70%      | 3.10%      | 2.90%      | 2.70%      | 1.50%      | 0.50%      | 0.40%      | 0.30%      | 0.20%      | 0.10%      | 0.10%      |
| * isomerase                       |            |            |            |            |            |            |            |            |            |            |            |
| * cytoskeletal protein            |            |            |            |            |            |            |            |            |            |            |            |
| kinase                            | 1.50%      |            |            |            |            |            |            |            |            |            |            |
| calcium-binding protein           |            |            |            |            |            |            |            |            |            |            |            |
| enzyme modulator                  |            |            |            |            |            |            |            |            |            |            |            |
| structural protein                |            |            |            |            |            |            |            |            |            |            |            |
| cell adhesion molecule            |            |            |            |            |            |            |            |            |            |            |            |
| receptor                          |            |            |            |            |            |            |            |            |            |            |            |
| ligase                            |            |            |            |            |            |            |            |            |            |            |            |
| * phosphatase                     |            |            |            |            |            |            |            |            |            |            |            |
| * transporter                     |            |            |            |            |            |            |            |            |            |            |            |
| * nucleic acid binding            | -8.10%     |            |            |            |            |            |            |            |            |            |            |

* denotes protein classes.
Fig. S3 Pairwise comparison of developmental expression changes among diploid and polyploid cotton. Log$_2$ expression ratios were calculated for 20 dpa versus 10 dpa within each genome using 2-DE (a) and iTRAQ datasets (b). Lower-left panels of the comparison matrix show scatterplots while upper-right panels give Pearson’s correlation coefficients, with the font size proportional to the absolute value of the correlation. Asterisks indicate significance of the adjacent correlation coefficient: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. 
Table S1  Protein identification and False discovery rate (FDR) analysis

a. Number of proteins identified at 95% confidence level

<table>
<thead>
<tr>
<th></th>
<th>run 1</th>
<th>run 2</th>
<th>run 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Spectra</td>
<td>114654</td>
<td>83352</td>
<td>110173</td>
<td></td>
</tr>
<tr>
<td>Spectra Identified</td>
<td>54020</td>
<td>29555</td>
<td>46871</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(47.1%)</td>
<td>(35.5%)</td>
<td>(42.5%)</td>
<td></td>
</tr>
<tr>
<td>Distinct Peptides</td>
<td>19652</td>
<td>10749</td>
<td>17250</td>
<td></td>
</tr>
<tr>
<td>Proteins</td>
<td>1483</td>
<td>886</td>
<td>1339</td>
<td>1652</td>
</tr>
</tbody>
</table>

b. Number of proteins estimated according to FDR analysis

<table>
<thead>
<tr>
<th>Critical FDR</th>
<th>Local FDR</th>
<th>Global FDR</th>
<th>Global FDR from Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0%</td>
<td>1388[836]</td>
<td>1085</td>
<td>1539[935]</td>
</tr>
<tr>
<td>5.0%</td>
<td>1449[853]</td>
<td>1270</td>
<td>1683[1014]</td>
</tr>
<tr>
<td>10.0%</td>
<td>1490[887]</td>
<td>1331</td>
<td>1817[1102]</td>
</tr>
</tbody>
</table>

* Protein numbers from triplicated experiments were listed; numbers in bold are close to that from protein identification with above 95% confidence level.
**Table S2** Normalized spot volumes of 2-DE gels

**Table S3** Protein identification and quantification by ProteinPilot iTRAQ analyses

**Table S4** Significant expression changes profiled by iTRAQ

**Table S5** Significant expression changes profiled by 2-DE

**Table S6** Homoeolog-specific peptides and expression changes
CHAPTER 6
GENERAL CONCLUSION

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

1. **To develop technology and tools for describing and studying the cotton fiber and seed proteomes.**

   With the aim of establishing a comprehensive platform for proteomic research in cotton evolution, I chose to apply two-dimensional gel electrophoresis (2-DE) and isobaric tag relative and absolute quantification (iTRAQ) methods in our studies, as representatives of two major types of modern proteomic strategies, gel-based and mass spectrometry (MS)-based, respectively. In Chapters 2 and 3, I present their separate application for comparative and quantitative analyses of cotton proteomes, and our results demonstrate that both 2-DE and iTRAQ are promising methods to understand the protein level consequences of evolutionary processes. By bringing together their analytic strengths in Chapters 4 and 5, I confirmed that the technological and methodological differences between 2-DE and iTRAQ make them complementary methods for exploring the complexity of protein level effects with respect to polyploidization and crop domestication. The large amount of data generated will serve as a rich proteomic database for the benefit of functional analyses of cotton biology as well as for evolutionary understanding.
2. To describe the cotton proteome from the standpoint of fiber development, which will allow us to assess the changes that accompany fiber evolution and domestication, and how this correlates with existing information on the transcriptome.

Chapters 3 and 4 each address these questions by studying one important crop species of cotton, and profile expression changes for thousands of fiber proteins with respect to developmental dynamics and crop domestication. In Chapter 3, I show that human-mediated selection may have shifted the timing of developmental modules, such that some important biological processes involved in fiber elongation occur earlier in the domesticated form of *G. barbadense*. In Chapter 4, I find that domestication of *G. hirsutum* appears to have targeted primary wall synthesis through increased regulatory dynamics, as evidenced by over twice as many proteins being differentially expressed in the domesticated form. The proteomic changes observed are of the same order of magnitude as those previously observed at the transcriptomic level; however, as in Chapter 3, there is poor correlation between protein and transcript expression change, which again highlight the indispensable and complementary role of proteomics in studying plant biology and evolution.

3. To understand how the proteome responds to genome doubling; that is, what is novel about polyploid cotton fiber and seed relative to that of its antecedent diploids?
Chapters 2 and 5 address this question for cotton seeds and fibers, respectively. In Chapter 2, I find an unexpectedly high level of divergence between diploid and polyploid seed proteomes, and that in allopolyploid cotton there is a biased accumulation of seed storage proteins favoring D-genome parent. In Chapter 5, the amount of interspecific variation of protein expression in fibers is much lower than that observed in seeds; with respect to allopolyploidization, both cotton allopolyploids show unbalanced protein expression level dominance, with *G. hirsutum* consistently favoring the parental A-genome, while *G. barbadense* displays opposite directions at different stages of fiber development. These findings lead us to a general conclusion that genomic merger and doubling have consequences that extend beyond the transcriptome into the realm of the proteome, with the magnitude and direction of impacts, with respect to the diploid parents, varying among allopolyploid species, tissue types and the developmental stages studied.

The other major finding regarding proteomic consequences of polyploidy is evident in all of the research chapters contained in this thesis. Specifically, homoeologous copies of proteins can be differentially regulated in allopolyploids, thereby generating new expression space for functional and phenotypic novelty and targeted human and natural selection.

4. **To detail proteomic consequences of cotton fiber evolution and domestication; for example, to catalog the key proteins associated with and therefore possibly responsible for phenotype changes and important traits relevant to crop improvement.**
The data generated in Chapters 3 and 4 present the comparative expression profiles for thousands of fiber proteins, and more importantly provide identification of those differentially expressed during fiber development and by domestication, such as the important signal protein flowering locus T, which accumulates high abundance but has not before been reported in cotton fibers, and enzymes in the flavonoid biosynthesis pathway that are coordinately regulated in response to both fiber development and domestication. As candidates for future functional analyses, these cataloged proteins will provide a useful resource that can be mined to expand our understanding of cotton biology and yield insight into crop improvement.