Proteomic profiling of developing cotton fibers from wild and domesticated *Gossypium barbadense*

Guanjing Hu[1], Jin Koh[2,3], Mi-Jeong Yoo[2], Kara Grupp[1], Sixue Chen[2,3,4] and Jonathan F. Wendel[1]

[1]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011, USA; [2]Department of Biology, University of Florida, Gainesville, FL 32610, USA; [3]Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, FL 32610, USA; [4]Genetics Institute, University of Florida, Gainesville, FL 32610, USA

Author for correspondence:
Jonathan F. Wendel
Tel: +1 515 290 9237
Email: jfw@iastate.edu

Received: *18 April 2013*
Accepted: *27 May 2013*

New Phytologist (2013) **200:** 570–582
doi: 10.1111/nph.12381

Key words: domestication, expression-level bias, fiber development, *Gossypium barbadense* (Pima cotton), homoeolog expression, iTRAQ, polyploidy, proteomics.

Summary

- Pima cotton (*Gossypium barbadense*) 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 an 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 isobaric tag for relative and absolute quantification (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 degree 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* (Upland cotton) and *Gossypium barbadense* (Pima or Egyptian cotton), account for *c.* 99% of the world's cotton production. *G. barbadense*, which comprises a relatively small proportion of US plantings (4%), carries a 50–80% price premium compared with *G. hirsutum* fiber, owing to its superior fiber properties (longer staple length and higher strength), which can generate stronger and softer threads, yarns, and fabrics (http://www.cotton.org).

Cotton plants belong to *Gossypium*, which includes *c.* 45 diploids that represent *c.* 10 million yr 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 yr 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 *c.* 5000 yr ago in the Yucatan Peninsula and the intermontane Peruvian Andes areas, respectively (Wendel & Cronn, 2003; Wendel *et al.*, 2012). Over 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 (ROS) that control cellular redox concentrations (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 degree 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 executors for most cellular activities, for example, 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 isobaric tag for relative and absolute quantification (iTRAQ) 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 coanalysis 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 cv Pima S-7 and a wild accession K101 from Bolivia were chosen to represent domesticated and wild accessions of *G. barbadense* L., respectively (Fig. 1). Plants were grown in the Bessey Hall Glasshouse at Iowa State University. Flowers were tagged at anthesis and harvested at four key developmental stages, that is, 5, 10, 20 and 25 d post-anthesis (dpa), representing primary wall synthesis (5 dpa) and elongation (10 dpa), and the transition to (20 dpa) and continuation (25 dpa) 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) polyvinylpolypyrrolidone 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) sodium dodecyl sulfate (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 (8 M urea, 25 mM triethylammonium bicarbonate (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 three times with cold 80% acetone and solubilized in the protein buffer. Protein assays were performed using an EZQ® Protein Quantitation Kit (Invitrogen).

iTRAQ labeling, strong cation exchange and LC-MS/MS

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,

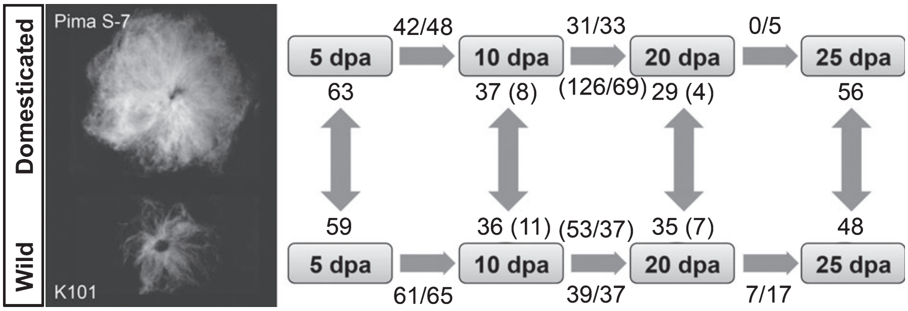

Fig. 1 Number of proteins differentially expressed during fiber development within and between wild and domesticated *Gossypium barbadense*. A representative image of a single seed with attached trichomes, that is cotton fibers, is shown for each accession. Arrows represent comparisons conducted in the 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 d post-anthesis (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.

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 high-performance liquid chromatography (HPLC) system 1260 with a polysulfoethyl A column (2.1 × 100 mm, 5 μm, 300 Å; PolyLC, Columbia, MD, USA; Supporting Information, Fig. 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 the area size of each fraction was calculated for the percentage coefficient of variation among three biological replicates (Fig. 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). A hybrid quadrupole time-of-flight QSTAR Elite MS/MS system (AB Sciex, Inc.) 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 nonredundant *Gossypium* protein database (122 785 entries) using the recently sequenced genome (Paterson *et al.*, 2012) of the diploid D-genome species *Gossypium raimondii* and a cotton SNP index (Page *et al.*, 2013) generated between *G. raimondii* and the A-genome diploid *Gossypium 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.). To examine homoeolog-specific expression, that is, 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 < 5% local false-positive identification rate (false discovery 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 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 with general Gene Ontology 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 catalog (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; Table S5). To examine the degree of concordance between transcript and protein levels, Pearson correlation tests were conducted using expression ratios of 20 vs 10 dpa in both accessions and Pima S-6 vs K101 at both time points, respectively.

Results

Identification of *G. barbadense* fiber proteins

Using our *Gossypium* protein database, a total of 1317 proteins were identified at a 95% confidence level and a 1% FDR (Tables S1–S3). 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 overrepresented, while nucleic acid binding (11.3%), transcription factor (2.0%), transporter (2.6%) and phosphatase (0.6%) proteins were underrepresented in the fiber proteomes (Fig. 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 (Tables S3, S4). Proteomic changes were first examined between adjacent developmental stages (5–10, 10–20, and 20–25 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 amount of developmental expressional variation (151 proteins, 11.5%) than did K101 (198 proteins, 15.0%). As shown in Fig. 1, the distribution of changes was biased toward the earliest developmental stage in both accessions ($P < 0.05$, chi-squared test): 90 (5–10 dpa), 64 (10–20 dpa), and five (20–25 dpa) differentially expressed proteins were identified in Pima S-7, and 126 (5–10 dpa), 76 (10–20 dpa) and 24 (20–25 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 (Fig. 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

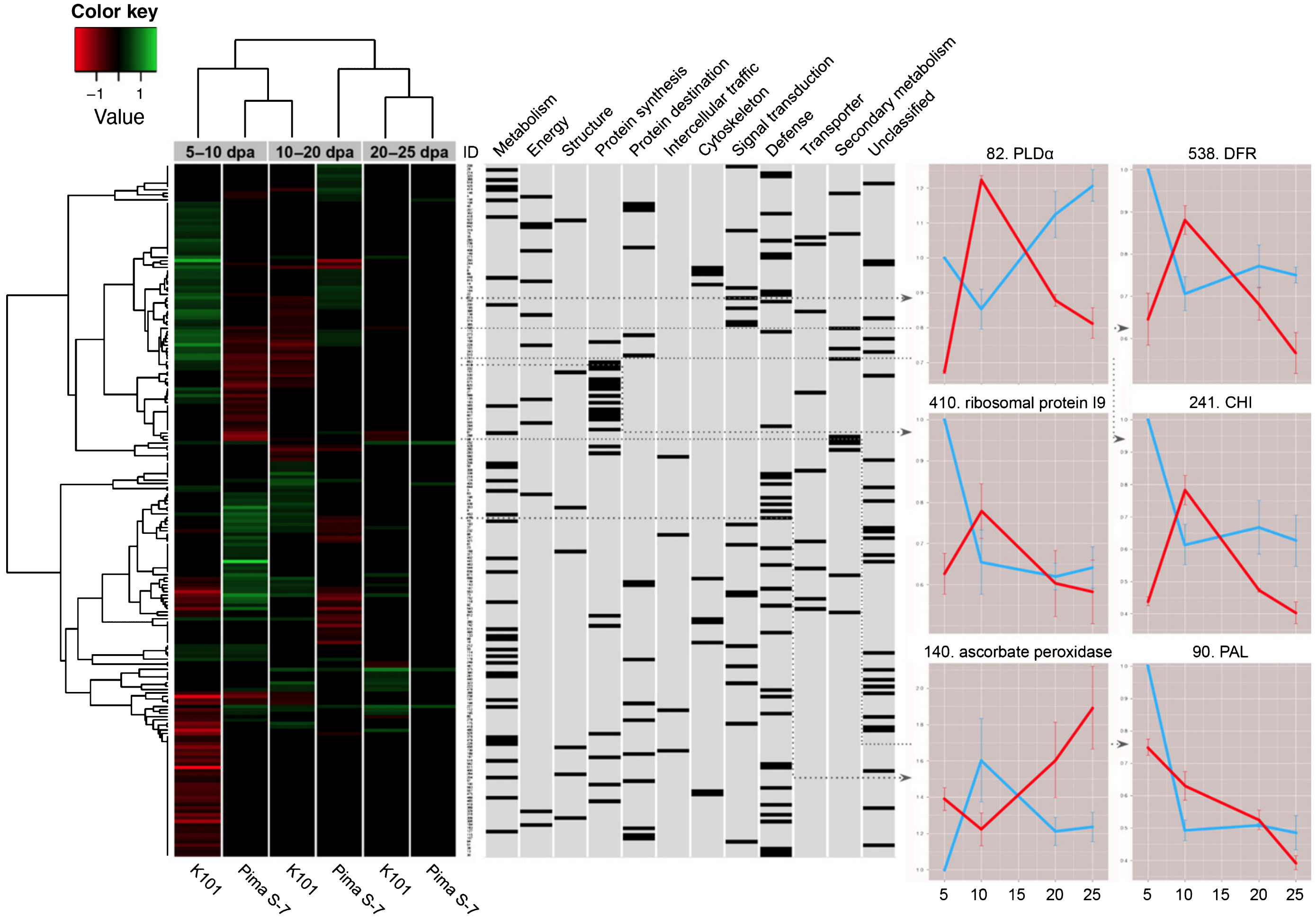

Fig. 2 Analysis of expression changes in the *Gossypium barbadense* fiber proteome accompanying development and domestication. Expression ratios of adjacent time points were calculated using the earlier stage as denominator, and plotted in a heatmap on a $\log_2$ scale. Differential expression patterns 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, respectively; black 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 in black in the central gay columns. Examples of protein expression profiles are shown on the right (wild, red line; domesticated, blue line). dpa, days post-anthesis; PLDα, phospholipase D alpha; DFR, dihydroflavonol 4-reductase; CHI, chalcone isomerase; PAL dihydroflavonol 4-reductase.

cluster was formed between 10 and 20 dpa in K101 and the earlier course of 5–10 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 (Fig. 2, left dendrogram). Proteins involved in similar or relevant cellular activities in most functional classes displayed diverse expression patterns (Fig. 2, grey columns; Table S4). 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 (Table S4). 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-squared 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 Table S4). 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

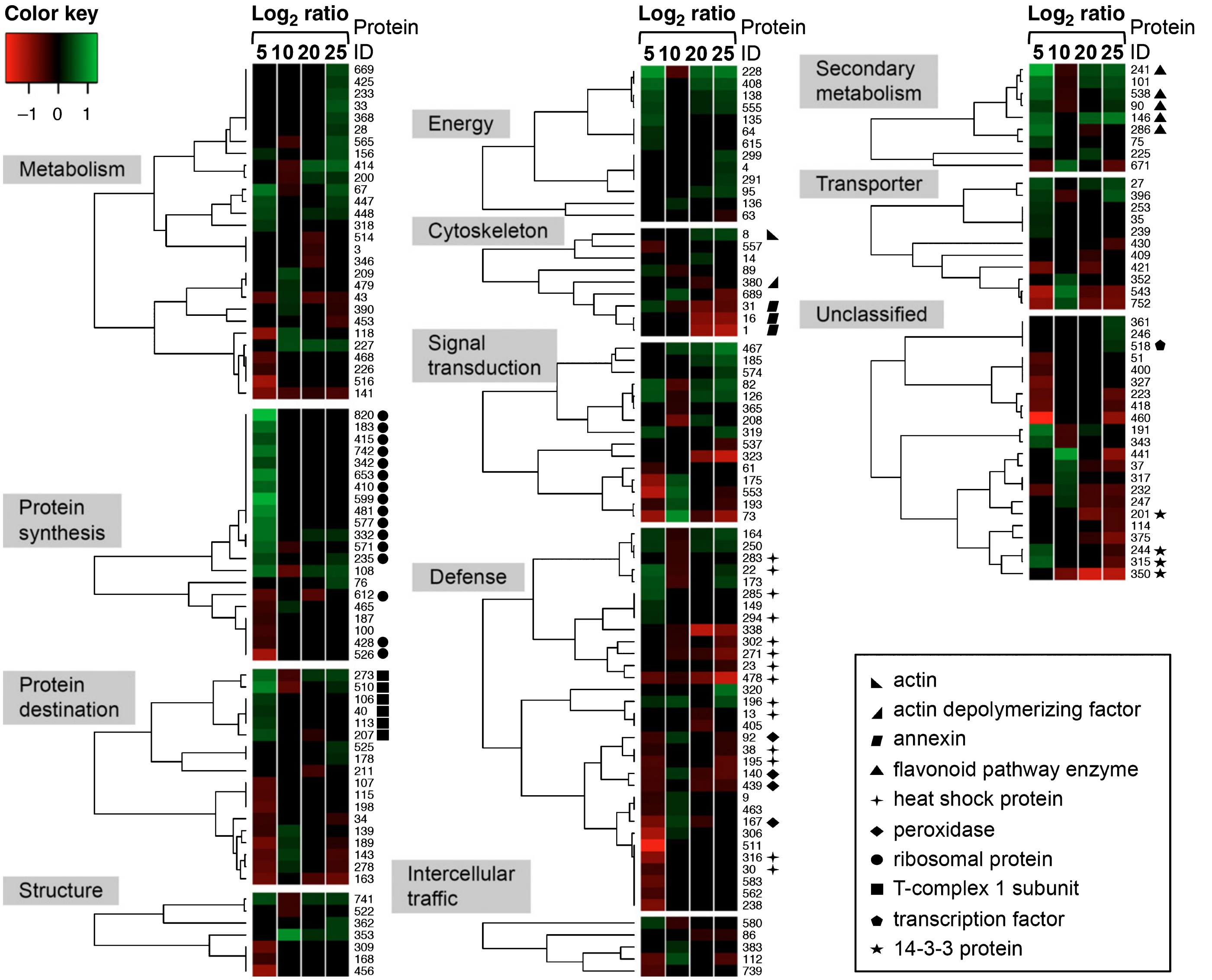

Fig. 3 A total of 190 proteins differentially expressed in domesticated *Gossypium barbadense* relative to its wild counterpart at one or more developmental stages. Expression ratios were plotted in a heatmap on a $\log_2$ scale. The green and red colors indicate up- and down-regulation in domesticated Pima S-7, respectively, relative to the wild form. Black represents no significant expression change.

10 dpa but 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, Hsp70, Hsp83 and Hsp90, 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 (second 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' (third 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 (second 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 (Table S5), allowing a direct comparison of transcript and protein expression during cotton fiber elongation. Concordance tests revealed poor correlations ($r_{\mathrm{Pearson}} = 0.06$–$0.24$) between mRNA and protein ratios (Fig. 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 (Fig. 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 17 cases of concordant expression were evident, that is, where protein accumulation was significantly correlated with transcript abundance (Fig. 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 (Table S6). 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 as a result of the technical limitations of 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-squared 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, that is, 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; Table S5), 27 exhibited the same directional bias at the protein level (Table 1, ratios marked with 'c'), 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

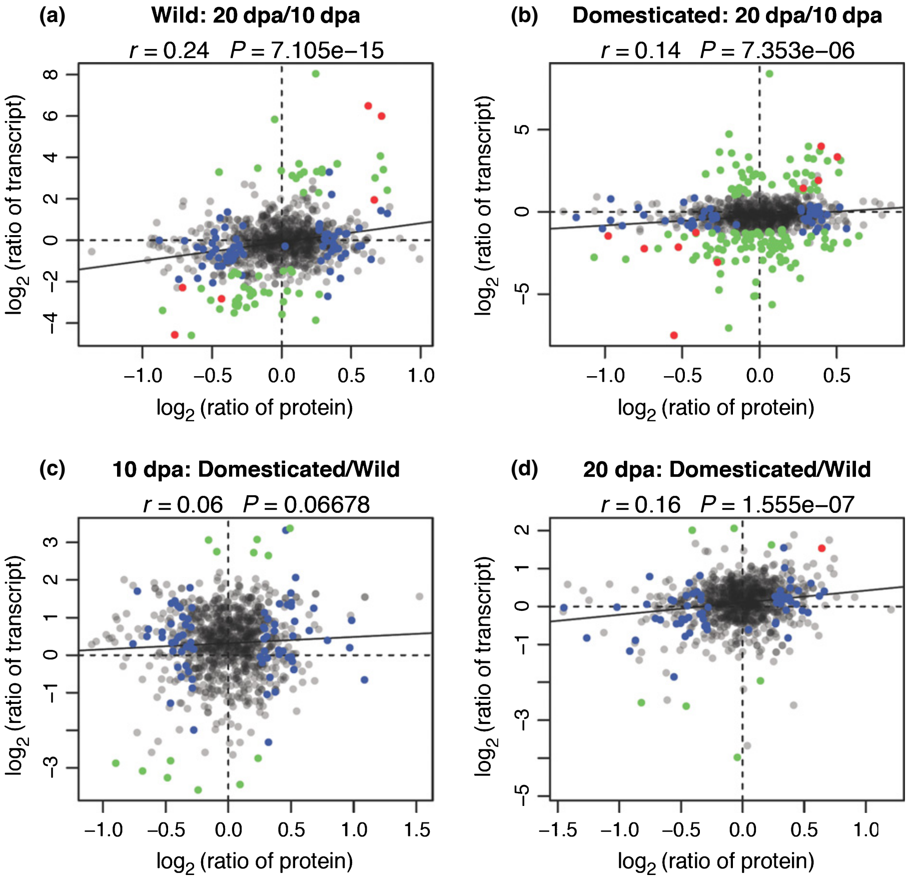

Fig. 4 Comparison of expression ratios from transcriptomic (*y*-axis) and proteomic (*x*-axis) profiling. Log$_2$ expression ratios were calculated from 20 d post-anthesis (dpa) vs 10 dpa within wild (a) and domesticated cotton (*Gossypium barbadense*) (b), and wild vs domesticated cotton at 10 dpa (c) 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) an osmotin-like protein precursor (Gorai.009G100600) and two proteins with unknown functions (Gorai.006G137000, Gorai.013G097200) were up-regulated at 10 dpa, while beta-alanine-pyruvate amino-transferase (Gorai.006G233600), a ferredoxin-like protein (Gorai.010G096500) and mlp-like protein 28 (Gorai.N010400) were up-regulated at 20 dpa; (b) annexin 2 (Gorai.007G239000), dehydroascorbate reductase (Gorai.012G068600), glutamine synthetase (Gorai.012G133400), a lipid transfer protein (Gorai.012G176200) and two unknown proteins (Gorai.013G097200, Gorai.013G130500) were up-regulated at 10 dpa, while 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) were up-regullated at 20 dpa; (d) UDP-glucose:flavonoid 3-o-glucosyltransferase (Gorai.012G009300)..

mechanisms by which traits arise. Over the past 20 yr, 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 & 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, has recently 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; M. J. Yoo *et al.*, unpublished). Nearly a quarter of the transcriptome, comprising 9465 genes, was differentially expressed between wild

Table 1 Homoeolog-specific expression at the transcript level for genes showing biased expression at the proteomic level

Gene ID	Description	K101[a]		Pima[a]	
		10	20	10	20
Detection of A-specific peptides only: ratios of A_T/D_T transcripts					
Gorai.001G016900	Chaperonin 20	0.9	0.6	0.5	0.5[b]
Gorai.001G032300	Leucine aminopeptidase 1	14[c]	5.5[c]	21[c]	33.6[c]
Gorai.001G261100	Aldehyde dehydrogenase family 2 member B4, mitochondrial	0.8	0.7	0.4[b]	1.1
Gorai.001G261400	Formate dehydrogenase	1.5	1.8	2.5	1.7
Gorai.002G062100	4-methyl-5(b-hydroxyethyl)-thiazole monophosphate biosynthesis	1.1	1.6	1.8	1.7
Gorai.002G119600	Dehydrin	2.8	1.2	3.2[c]	2.6[c]
Gorai.002G233700	Unknown	1.2	1.2	1.6	1.3
Gorai.004G056700	nadh-ubiquinone oxidoreductase 24 kDa subunit	1.7	1.3	1.3	1.3
Gorai.004G211400	Glutathione peroxidase	1.2	0.8	1.4	0.8
Gorai.004G246300	Auxin-induced protein PCNT115 isoform 1	7.0	10.1[c]	11.8[c]	8.7[c]
Gorai.004G260100	26S proteasome nonATPase regulatory subunit RPN12A-like	1.0	0.9	1.1	1.1
Gorai.005G096500	Aspartic proteinase nepenthesin-1-like	6.9	8.4[c]	10.4[c]	22.6[c]
Gorai.005G187700	Pyruvate dehydrogenase alpha subunit	1.3	3.4[c]	1	3.8[c]
Gorai.006G172100	Oxysterol-binding family protein	0.8	0.6	0.5	0.7
Gorai.007G090400	Chloroplast transketolase	1.6	2.3	1.8	1.8
Gorai.007G168600	Unknown	0.5	0.1[b]	0.5	0.1[b]
Gorai.007G175100	Glutathione s-transferase	0.9	2.6	1.6	2.3[c]
Gorai.007G257500	Methionine synthase	7.3	12.4[c]	8.4[c]	11.0[c]
Gorai.007G307000	20S proteasome beta subunit PBG1	1.6	0.8	0.9	1.0
Gorai.008G099900	Unknown	21.6[c]	4.6	7.5[c]	1.8
Gorai.009G035800	Fiber protein GLP1	0	0.2	0.2[b]	0.2[b]
Gorai.009G203100	Deoxyuridine 5-triphosphate nucleotidohydrolase-like	1.4	1.7	1.2	4.1[c]
Gorai.009G234600	14-3-3 protein	1.6	0.9	1.2	1.0
Gorai.009G418800	Syntaxin-71	6.1	2.9	2.3	2.8[c]
Gorai.010G159200	Ketol-acid reductoisomerase	0.8	0.8	0.9	0.9
Gorai.010G185300	Unknown	16.9	13.3[c]	6.0[c]	8.9[c]
Gorai.012G062300	Hop-interacting protein THI141	5.1	3.4	5.6[c]	1.9
Detection of D-specific peptides only: ratios of D_T/T_T transcripts					
Gorai.001G038600	Glutathione peroxidase	0.6	1.0	0.8	1.1
Gorai.001G103700	Quinone-oxidoreductase-like protein	0.6	1.5	2.0	2.0
Gorai.002G036600	Elongation factor ef-2	1.9	1.9	1.9	2.1[c]
Gorai.002G063300	Mitochondrial nad-dependent malate dehydrogenase	1.5	1.5	1.5	1.6
Gorai.002G252300	Glutathione S-transferase	71.2[c]	91.3[c]	86.1[c]	91.7[c]
Gorai.003G097400	nad-malate dehydrogenase	1.8	1.8	2.2	2.4[c]
Gorai.004G009000	60s acidic ribosomal protein	10.1	14.8[c]	6.7[c]	8.5[c]
Gorai.004G048300	utp:alpha-D-glucose-1-phosphate uridylyltransferase	0.7	0.7	0.7	1.0
Gorai.004G063900	6-phosphogluconate dehydrogenase	2.7	2.4	2.3	1.9
Gorai.004G096800	Glutaredoxin-C4	1.5	1.5	1.7	1.6
Gorai.004G166200	Unknown	0.4	0.3	0.5	0.5[b]
Gorai.004G226600	RPM1-interacting protein 4-like	2.4	0.7	2.7[c]	0.9
Gorai.005G028100	rna-binding glycine-rich protein	1.3	1.3	1.3	1.1
Gorai.005G050200	Malate dehydrogenase	1.4	1.4	1.4	1.4
Gorai.006G070400	*N*-carbamoyl-L-amino acid hydrolase	14.1[c]	13.7[c]	23.6[c]	108.5[c]
Gorai.006G156700	60s acidic ribosomal protein	0.6	0.4	0.6	0.4[b]
Gorai.007G022600	Elicitor-responsive protein 3-like	3.7	1.4	2.0	1.0
Gorai.008G048100	Glycine decarboxylase complex h-protein	1.1	0.9	1.0	0.8
Gorai.008G155400	Fasciclin-like arabinogalactan protein 9	9.2	5.0[c]	8.8[c]	5.3[c]
Gorai.008G169300	Glutamate decarboxylase	2.8	2.9	4.1[c]	1.2
Gorai.008G285900	Aspartic proteinase nepenthesin-1	68[c]	31.0[c]	150.5[c]	8.6[c]
Gorai.009G060500	Aspartate aminotransferase	6.2	5.4[c]	1.2	1.6
Gorai.009G078400	Beta-galactosidase 8	2.6	1.2	4.5[c]	1.6
Gorai.009G432600	Uridine 5′-monophosphate synthase	1.0	0.7	0.3[b]	0.7
Gorai.010G221900	Probable leucine-rich repeat receptor-like protein kinase at5 g49770-like	0.7	1.0	0.7	1.2
Gorai.011G026000	Clathrin light chain protein	1.2	1.2	1.0	1.0
Gorai.011G035400	GDSL esterase/lipase	1.7	1.9	3.8[c]	3.5[c]
Gorai.012G142600	2-nitropropane dioxygenase	34.8[c]	41.8[c]	2.8[c]	4.4[c]
Gorai.012G146500	Dolichyl-diphosphooligosaccharide–protein glycosyltransferase 48 kDa subunit-like	1.3	1.3	1.5	1.6
Gorai.013G216000	Leucine aminopeptidase 3	1.1	3.8[c]	1.3	1.0

[a] RNA-seq datasets of 10 and 20 dpa were separately analyzed.

[b] Significant transcriptional bias toward the opposite direction of protein-level bias.

[c] Significant transcriptional bias toward the same direction of protein-level bias.

and domesticated fiber phenotypes in *G. hirsutum* (Rapp *et al.*, 2010), and a less dramatic alteration of *c.* 4200 genes was observed in domesticated *G. barbadense* compared with 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 (Fig. 3), which account for 14.4% of the proteins we profiled. This proteomic change is of the same order of magnitude as that previously observed at the transcriptional level, 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 (Fig. 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 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 a 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 Fig. 3, Table S4); 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, as a result of either 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 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 twodimensional 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 probably precludes our ability to discover additional proteins. Indeed, the fiber proteins we identified were overrepresented 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* (Fig. 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 *c.* 10 dpa in domesticated *G. hirsutum*, whereas this phase was delayed until *c.* 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 (Table S4). 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 (Fig. 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 coregulated 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 coregulated 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 Fig. 2. Relevant to this finding is the earlier suggestion that regulation of hydrogen peroxide (H_2O_2) and other ROS is a key process in both cotton fiber development and evolution (Hovav *et al.*, 2008). H_2O_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 concentrations 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 concentrations 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 (Fig. 2, Table S4). 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 (Fig. 3). As implicated in a previous proteomic analysis, a *G. hirsutum* cytosolic ascorbate peroxidase (*GhAPX1*) functions to detoxify H_2O_2 produced during fast fiber elongation, as evidenced by the fact that transcript abundance and enzymatic activity of *GhAPX1*, as well as fiber length, can be promoted by exogenous H_2O_2 (Li *et al.*, 2007). Therefore, it is reasonable to speculate that accumulation of peroxidase was unknowingly targeted by humans so that lower concentrations of H_2O_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; Fig. 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 is also 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 the biosynthesis of polyphenol compounds were identified in this group – phenylalanine ammonia-lyase (PAL, catalyzing the first committed step in the phenylpropanoid pathway), chalcone isomerase (CHI) and dihydroflavonol 4-reductase (DFR; Fig. 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 earlier. It was also reported that the products of phenylpropanoid pathway could be deposited in the 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, coordinated 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 with other analyses using 2-DE methods (Yao *et al.*, 2006; Yang *et al.*, 2008; Zhang *et al.*, 2013), iTRAQ data resolved at least fivefold 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 to 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 as to 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 that is 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, and 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 the LC-MS ⁄MS analysis. The LC-MS/MS system was funded by National Institutes of Health grant 1S10RR025418-01 to S.C. This work was funded by Cotton Incorporated grant 09-558 and by the NSF Plant Genome Research Program, both to J.F.W.

References

Achkor H, Diaz M, Fernandez MR, Biosca JA, Pares X, Martinez MC. 2003. Enhanced formaldehyde detoxification by overexpression of glutathione-dependent formaldehyde dehydrogenase from *Arabidopsis. Plant Physiology* **132**: 2248–2255.

Al-Ghazi Y, Bourot S, Arioli T, Dennis ES, Llewellyn DJ. 2009. Transcript profiling during fiber development identifies pathways in secondary metabolism and cell wall structure that may contribute to cotton fiber quality. *Plant and Cell Physiology* **50**: 1364–1381.

Applequist WL, Cronn R, Wendel JF. 2001. Comparative development of fiber in wild and cultivated cotton. *Evolution & Development* **3**: 3–17.

Arpat AB, Waugh M, Sullivan JP, Gonzales M, Frisch D, Main D, Wood T, Leslie A, Wing RA, Wilkins TA. 2004. Functional genomics of cell elongation in developing cotton fibers. *Plant Molecular Biology* **54**: 911–929.

Bao Y, Hu G, Flagel LE, Salmon A, Bezanilla M, Paterson AH, Wang Z, Wendel JF. 2011. Parallel up-regulation of the profilin gene family following independent domestication of diploid and allopolyploid cotton (*Gossypium*). *Proceedings of the National Academy of Sciences, USA* **108**: 21152–21157.

Bevan M, Bancroft I, Bent E, Love K, Goodman H, Dean C, Bergkamp R, Dirkse W, Van Staveren M, Stiekema W *et al.* 1998. Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana. Nature* **391**: 485–488.

Burger JC, Chapman MA, Burke JM. 2008. Molecular insights into the evolution of crop plants. *American Journal of Botany* **95**: 113–122.

Burke JM, Burger JC, Chapman MA. 2007. Crop evolution: from genetics to genomics. *Current Opinion in Genetics & Development* **17**: 525–532.

Chapman MA, Pashley CH, Wenzler J, Hvala J, Tang S, Knapp SJ, Burke JM. 2008. A genomic scan for selection reveals candidates for genes involved in the evolution of cultivated sunflower (*Helianthus annuus*). *The Plant Cell* **20**: 2931–2945.

Chaudhary B, Hovav R, Flagel L, Mittler R, Wendel JF. 2009. Parallel expression evolution of oxidative stress-related genes in fiber from wild and domesticated diploid and polyploid cotton (*Gossypium*). *BMC Genomics* **10**: 378.

Chaudhary B, Hovav R, Rapp R, Verma N, Udall JA, Wendel JF. 2008. Global analysis of gene expression in cotton fibers from wild and domesticated *Gossypium barbadense*. *Evolution & Development* **10**: 567–582.

Cho RJ, Campbell MJ. 2000. Transcription, genomes, function. *Trends in Genetics : TIG* **16**: 409–415.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**: 3674–3676.

Darwin C. 1859. *On the origin of the species by means of natural selection: or, the preservation of favoured races in the struggle for life.* London, UK: John Murray.

Diz AP, Martinez-Fernandez M, Rolan-Alvarez E. 2012. Proteomics in evolutionary ecology: linking the genotype with the phenotype. *Molecular Ecology* **21**: 1060–1080.

Doebley J. 2004. The genetics of maize evolution. *Annual Review of Genetics* **38**: 37–59.

Doebley J, Stec A, Hubbard L. 1997. The evolution of apical dominance in maize. *Nature* **386**: 485–488.

Doebley JF, Gaut BS, Smith BD. 2006. The molecular genetics of crop domestication. *Cell* **127**: 1309–1321.

Fan L, Shi WJ, Hu WR, Hao XY, Wang DM, Yuan H, Yan HY. 2009. Molecular and biochemical evidence for phenylpropanoid synthesis and presence of wall-linked phenolics in cotton fibers. *Journal of Integrative Plant Biology* **51**: 626–637.

Fisher RA. 1948. Questions and answers #14. *The American Statistician* **2**: 30–31.

Foreman J, Demidchik V, Bothwell JH, Mylona P, Miedema H, Torres MA, Linstead P, Costa S, Brownlee C, Jones JD *et al.* 2003. Reactive oxygen species produced by NADPH oxidase regulate plant cell growth. *Nature* **422**: 442–446.

Frary A, Nesbitt TC, Grandillo S, Knaap E, Cong B, Liu J, Meller J, Elber R, Alpert KB, Tanksley SD. 2000. fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. *Science* **289**: 85–88.

Fry SC. 1998. Oxidative scission of plant cell wall polysaccharides by ascorbate-induced hydroxyl radicals. *Biochemical Journal* **332**(Pt 2): 507–515.

Gou JY, Wang LJ, Chen SP, Hu WL, Chen XY. 2007. Gene expression and metabolite profiles of cotton fiber during cell elongation and secondary cell wall synthesis. *Cell Research* **17**: 422–434.

Gross BL, Olsen KM. 2010. Genetic perspectives on crop domestication. *Trends in Plant Science* **15**: 529–537.

Gross BL, Strasburg JL. 2010. Cotton domestication: dramatic changes in a single cell. *BMC Biology* **8**: 137.

Grover CE, Gallagher JP, Szadkowski EP, Yoo MJ, Flagel LE, Wendel JF. 2012. Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytologist* **196**: 966–971.

Hovav R, Udall JA, Chaudhary B, Hovav E, Flagel L, Hu G, Wendel JF. 2008. The evolution of spinnable cotton fiber entailed prolonged development and a novel metabolism. *PLoS Genetics* **4**: e25.

Hu G, Houston NL, Pathak D, Schmidt L, Thelen JJ, Wendel JF. 2011. Genomically biased accumulation of seed storage proteins in allopolyploid cotton. *Genetics* **189**: 1103–1115.

Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W *et al.* 2012. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**: 497–501.

Hufford MB, Xu X, van Heerwaarden J, Pyhajarvi T, Chia JM, Cartwright RA, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM *et al.* 2012. Comparative population genomics of maize domestication and improvement. *Nature Genetics* **44**: 808–811.

Jin J, Huang W, Gao JP, Yang J, Shi M, Zhu MZ, Luo D, Lin HX. 2008. Genetic control of rice plant architecture under domestication. *Nature Genetics* **40**: 1365–1369.

Karr TL. 2008. Application of proteomics to ecology and population biology. *Heredity* **100**: 200–206.

Kinoshita T, Ono N, Hayashi Y, Morimoto S, Nakamura S, Soda M, Kato Y, Ohnishi M, Nakano T, Inoue S *et al.* 2011. *FLOWERING LOCUS T* regulates stomatal opening. *Current Biology* **21**: 1232–1238.

Koh J, Chen S, Zhu N, Yu F, Soltis PS, Soltis DE. 2012. Comparative proteomics of the recently and recurrently formed natural allopolyploid *Tragopogon mirus* (Asteraceae) and its parents. *New Phytologist* **196**: 292–305.

Konishi S, Izawa T, Lin SY, Ebana K, Fukuta Y, Sasaki T, Yano M. 2006. An SNP caused loss of seed shattering during rice domestication. *Science* **312**: 1392–1396.

Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, He W, Qin N, Wang B *et al.* 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics* **42**: 1053–1059.

Lepiniec L, Debeaujon I, Routaboul JM, Baudry A, Pourcel L, Nesi N, Caboche M. 2006. Genetics and biochemistry of seed flavonoids. *Annual Review of Plant Biology* **57**: 405–430.

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. *ScientificWorldJournal* **2012**: 694740.

Li C, Zhou A, Sang T. 2006. Rice domestication by reducing shattering. *Science* **311**: 1936–1939.

Li HB, Qin YM, Pang Y, Song WQ, Mei WQ, Zhu YX. 2007. A cotton ascorbate peroxidase is involved in hydrogen peroxide homeostasis during fibre cell development. *New Phytologist* **175**: 462–471.

Liszkay A, van der Zalm E, Schopfer P. 2004. Production of reactive oxygen intermediates ($O_2^{\cdot-}$, H_2O_2, and $^{\cdot}OH$) by maize roots and their role in wall loosening and elongation growth. *Plant Physiology* **136**: 3114–3123; discussion 3001.

Liu H, Sadygov RG, Yates JR 3rd. 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical Chemistry* **76**: 4193–4201.

McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, Zeller G, Clark RM, Hoen DR, Bureau TE *et al.* 2009. Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proceedings of the National Academy of Sciences, USA* **106**: 12273–12278.

Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. 2010. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Research* **38**: D204–D210.

Millar AA, Dennis ES. 1996. The alcohol dehydrogenase genes of cotton. *Plant Molecular Biology* **31**: 897–904.

Millar AA, Olive MR, Dennis ES. 1994. The expression and anaerobic induction of alcohol dehydrogenase in cotton. *Biochemical Genetics* **32**: 279–300.

Mimida N, Kidou S, Iwanami H, Moriya S, Abe K, Voogd C, Varkonyi-Gasic E, Kotoda N. 2011. Apple FLOWERING LOCUS T proteins interact with transcription factors implicated in cell growth and organ development. *Tree Physiology* **31**: 555–566.

Mittler R, Vanderauwera S, Suzuki N, Miller G, Tognetti VB, Vandepoele K, Gollery M, Shulaev V, Van Breusegem F. 2011. ROS signaling: the new wave? *Trends in Plant Science* **16**: 300–309.

Olsen KM, Wendel JF. 2013. A bountiful harvest: genomic insights into crop domestication phenotypes. *Annual Review of Plant Biology* **64**: 47–70.

Page JT, Gingle AR, Udall JA. 2013. PolyCat: a resource for genome categorization of sequencing reads from allopolyploid organisms. *G3 (Bethesda)* **3**: 517–525.

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J *et al.* 2012. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* **492**: 423–427.

Potikha TS, Collins CC, Johnson DI, Delmer DP, Levine A. 1999. The involvement of hydrogen peroxide in the differentiation of secondary walls in cotton fibers. *Plant Physiology* **119**: 849–858.

Rapp RA, Haigler CH, Flagel L, Hovav RH, Udall JA, Wendel JF. 2010. Gene expression in developing fibres of Upland cotton (*Gossypium hirsutum* L.) was massively altered by domestication. *BMC Biology* **8**: 139.

Reumann S, Babujee L, Ma C, Wienkoop S, Siemsen T, Antonicelli GE, Rasche N, Luder F, Weckwerth W, Jahn O. 2007. Proteome analysis of *Arabidopsis* leaf peroxisomes reveals novel targeting peptides, metabolic pathways, and defense mechanisms. *The Plant Cell* 19: 3170–3193.

Rodriguez AA, Grunberg KA, Taleisnik EL. 2002. Reactive oxygen species in the elongation zone of maize leaves are necessary for leaf extension. *Plant Physiology* **129**: 1627–1632.

Sang Y, Cui D, Wang X. 2001. Phospholipase D and phosphatidic acid-mediated generation of superoxide in *Arabidopsis*. *Plant Physiology* **126**: 1449–1458.

Schopfer P. 1996. Hydrogen peroxide-mediated cell-wall stiffening *in vitro* in maize coleoptiles. *Planta* **199**: 43–49.

Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. 2011. Global quantification of mammalian gene expression control. *Nature* **473**: 337–342.

Shalit A, Rozman A, Goldshmidt A, Alvarez JP, Bowman JL, Eshed Y, Lifschitz E. 2009. The flowering hormone florigen functions as a general systemic regulator of growth and termination. *Proceedings of the National Academy of Sciences, USA* **106**: 8392–8397.

Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP, Hunter CL, Nuwaysir LM, Schaeffer DA. 2007. The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Molecular & Cellular Proteomics* **6**: 1638–1655.

Simons KJ, Fellers JP, Trick HN, Zhang Z, Tai YS, Gill BS, Faris JD. 2006. Molecular characterization of the major wheat domestication gene Q. *Genetics* **172**: 547–555.

Suzuki N, Koussevitzky S, Mittler R, Miller G. 2012. ROS and redox signalling in the response of plants to abiotic stress. *Plant, Cell & Environment* **35**: 259–270.

Swanson-Wagner R, Briskine R, Schaefer R, Hufford MB, Ross-Ibarra J, Myers CL, Tiffin P, Springer NM. 2012. Reshaping of the maize transcriptome by domestication. *Proceedings of the National Academy of Sciences, USA* **109**: 11878–11883.

Taliercio EW, Boykin D. 2007. Analysis of gene expression in cotton fiber initials. *BMC Plant Biology* **7**: 22.

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 Physiology.* **162**: 86–95

Tang WH, Shilov IV, Seymour SL. 2008. Nonlinear fitting method for determining local false discovery rates from decoy database searches. *Journal of Proteome Research* **7**: 3661–3667.

Tatematsu K, Kamiya Y, Nambara E. 2008. Co-regulation of ribosomal protein genes as an indicator of growth status: comparative transcriptome analysis on axillary shoots and seeds in *Arabidopsis*. *Plant Signaling & Behavior* **3**: 450–452.

Tian F, Stevens NM, Buckler ESt. 2009. Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. *Proceedings of the National Academy of Sciences, USA* **106**(Suppl 1): 9979–9986.

Vogel C, Marcotte EM. 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics* **13**: 227–232.

Wang H, Nussbaum-Wagler T, Li B, Zhao Q, Vigouroux Y, Faller M, Bomblies K, Lukens L, Doebley JF. 2005. The origin of the naked grains of maize. *Nature* **436**: 714–719.

Wang P, Tang H, Zhang H, Whiteaker J, Paulovich AG, McIntosh M. 2006. Normalization regarding non-random missing values in high-throughput mass spectrometry data. *Pacific Symposium on Biocomputing* **11**: 315–326.

Wendel JF, Cronn RC. 2003. Polyploidy and the evolutionary history of cotton. In: Sparks DL ed. *Advances in agronomy.* New York, NY, USA: Academic Press, 139–186.

Wendel JF, Flagel LE, Adams KL. 2012. Jeans, genes, and genomes: cotton as a model for studying polyploidy. In: Soltis PS, Soltis DE, eds. *Polyploidy and genome evolution.* Berlin, Heidelberg, Germany: Springer, 181–207.

Wigge PA. 2011. FT, a mobile developmental signal in plants. *Current Biology* **21**: R374–R378.

Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS. 2005. The effects of artificial selection on the maize genome. *Science* **308**: 1310–1314.

Xu P, Wu X, Wang B, Luo J, Liu Y, Ehlers JD, Close TJ, Roberts PA, Lu Z, Wang S *et al.* 2012. Genome wide linkage disequilibrium in Chinese asparagus bean (*Vigna. unguiculata* ssp. *sesquipedialis*) germplasm: implications for domestication history and genome wide association studies. *Heredity* **109**: 34–40.

Yamasaki M, Tenaillon MI, Bi IV, Schroeder SG, Sanchez-Villeda H, Doebley JF, Gaut BS, McMullen MD. 2005. A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *The Plant Cell* **17**: 2859–2872.

Yang YW, Bian SM, Yao Y, Liu JY. 2008. Comparative proteomic analysis provides new insights into the fiber elongating process in cotton. *Journal of Proteome Research* **7**: 4623–4637.

Yao Y, Yang YW, Liu JY. 2006. An efficient protein preparation for proteomic analysis of developing cotton fibers by 2-DE. *Electrophoresis* **27**: 4559–4569.

Zdobnov EM, Apweiler R. 2001. InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**: 847–848.

Zhang B, Yang YW, Zhang Y, Liu JY. 2013. A high-confidence reference dataset of differentially expressed proteins in elongating cotton fiber cells. *Proteomics.* **13**: 1159–1163.

Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J *et al.* 2011. Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nature Communications* **2**: 467.

Zhu M, Dai S, Zhu N, Booy A, Simons B, Yi S, Chen S. 2012. Methyl jasmonate responsive proteins in *Brassica napus* guard cells revealed by iTRAQ-based quantitative proteomics. *Journal of Proteome Research* **11**: 3728–3742.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Representative example of the iTRAQ cation exchange chromatograph.

Fig. S2 Panther protein family classification.

Table S1 Analyses of false discovery rates (FDRs)

Table S2 Protein identifications at 95% confidence level

Table S3 Protein identification and quantification by ProteinPilot iTRAQ analyses

Table S4 Significant protein expression changes

Table S5 RNA-seq analysis of gene differential expressions

Table S6 Homoeolog-specific peptides identified in *G. barbadense* proteomes

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.