Transcriptome and proteome profiling revealed the key genes and pathways involved in the fiber quality formation in brown cotton

Zhenzhen Wang  
Chinese Academy of Agricultural Science (ICR, CAAS)

Abdul Rehman  
Zhengzhou University

Yinhua Jia  
Chinese Academy of Agricultural Science (ICR, CAAS)

Panhong Dai  
Chinese Academy of Agricultural Science (ICR, CAAS)

Shoupu He  
Chinese Academy of Agricultural Science (ICR, CAAS)

Xiaoyang Wang  
Chinese Academy of Agricultural Science (ICR, CAAS)

Hongge Li  
Chinese Academy of Agricultural Science (ICR, CAAS)

Linu Wang  
Chinese Academy of Agricultural Science (ICR, CAAS)

Abdul Qayyum  
Bahauddin Zakariya University

Zhen Peng  
Chinese Academy of Agricultural Science (ICR, CAAS)

Xiongming Du  
duxiongming@caas.cn  
Chinese Academy of Agricultural Science (ICR, CAAS)

Research Article

Keywords: brown cotton, proteomics, transcriptome, carbohydrate active enzymes, transcription factor

Posted Date: September 23rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2044629/v1
Abstract

Background

Colored cotton is also called eco-cotton because of its natural color fiber. However, the yield and quality of natural colored cotton are generally lower than white cotton. Now, little is known about the regulatory mechanisms involved in biochemical and metabolic pathways and critical genes between fiber quality and pigment synthesis. The present study used a pair of brown cotton near-isogenic lines (brown cotton ‘Z1-61’ and white cotton ‘RT’) to determine the transcriptomic and proteomic data of fiber (ovule) at 0DPA, 10DPA, and 20DPA.

Results

Integration analysis of the two-omics data demonstrated that the differential proteins with the same expression trend in the adjacent developmental stages were significantly and positively correlated with fold change. Enrichment analysis of metabolic pathway revealed that Z1-61, enriched in fiber elongation genes related to a flavonoid biosynthesis, phenylalanine metabolism, glutathione metabolism and many more genes (proteins) are up-regulated. Short Time-series Expression Miner analysis inferred that up-regulated genes of Z1-61 were specifically involved in the metabolism of carbohydrates, energy, and lipid, suggesting that these pathways play an important role in fiber elongation and pigmentation. We further analyzed the 164 glycosyltransferases genes were predominantly expressed in the early stage of Z1-61 fiber elongation while they were predominantly expressed during the later period of RT fibers elongation. And 15 MYB-bHLH-WD40 triple complex genes and other transcription factor genes such as C2H2 (12), ERF (11), and NAC (7) were also preferentially expressed in Z1-61. Weighted correlation network analysis exhibited that fatty acid synthesis and energy metabolism were the active metabolic pathways in both cotton genotypes during fiber development. The 10 hub genes obtained from the final screening can be considered to play a critical role in the process of fiber pigment deposition while ensuring that fiber quality does not degraded.

Conclusion

This is the first complementary analyses of the transcriptome and proteome related with the quality of brown cotton fiber. Those studies can be understood as better playing a major role in fiber quality enhancement and pigmentation during fiber elongation.

Background

Cotton having natural hues is known as naturally colored cotton (NCC). Hence, NCC reduced dyeing costs and harmful waste disposal [1, 2]. But NCC has few color types, instability, and low fiber quality, which severely restricts its cultivation. Brown and green cotton are the main types, but fiber quality is afflicted by
sunlight and wind [3]. Experiments were performed to improve NCC yield, but the results were not promising. Hence, understanding the fundamental principles of cotton fiber growth is a prerequisite to enhance fiber quality. Fiber development is divided into five phases: initiation (0–3 DPA), elongation (3–15 DPA), transitional cell-wall remodeling (15–20 DPA), secondary wall biosynthesis (20–40 DPA), and maturity (40–50 DPA) [4]. The first two phases are responsible for the length of the fibers. Whereas, the third and fourth phases are associated with fiber strength and fineness. Fiber cell formation is a complicated process comprising numerous routes, including signal transduction and transcriptional regulatory components [5]. Pigment metabolism differs between naturally colored and white cotton fibers [6–7]. Genetic and heterosis studies revealed that fiber color has a substantial negative link with lint percentage, fiber length, and strength [8]. Pigment synthesis genes may limit lint length and fineness; hence, color genotypes have worse fiber quality [2, 7].

It is a crucial scientific question to understand the molecular coordinated regulatory mechanism of color formation and fiber quality formation during growth. Various related research illustrated that the color of brown cotton is closely related to fiber quality. Feng et al. [8] showed that brown fiber negatively correlates with fiber yield and quality. Specific QTLs were identified for fiber length, strength, and fineness in the brown fiber genetic locus (LC1), which could reduce fiber length, strength, and fineness by 23%, 18%, and 16%, respectively [9]. Feng et al. [10] exhibited a negative correlation through QTL analysis of brown fiber color and quality and also found several QTLs related to fiber quality in the LC1 region. Sun et al. [11] demonstrated that the darker the fiber color of brown cotton, the worse the fiber quality. Xiao et al. [12] revealed that the synthesis of brown pigment in cotton fiber might directly or indirectly affect other metabolic and physiological pathways by comparing the difference in digital gene expression (DGE) between white cotton and brown cotton. Therefore, the biggest challenge is synthesizing and accumulating pigments in cotton fiber while ensuring the normal growth, development, quality, and yield of the fiber.

Proteomic research has emerged as a robust tool for studying gene expression at the protein level on a large scale [13]. Comparative transcriptome sequencing is a powerful tool for analyzing differential expression to find candidate genes linked to certain features. In the past decade, transcriptome analyses of cotton identified many potential genes linked with high-quality fiber in chromosomal segment substitutions lines (CSSLs) generated from near-isogenic lines (NILs) [14–15], wild and domesticated *G. hirsutum* [16–17], and *G. hirsutum × G. barbadense* [18–20].

However, the mechanism of coordinated regulation of pigmentation by brown cotton fiber development to achieve the balance between quality and color formation has not been revealed at the transcriptional and translation levels. Therefore, the purpose of this study was to determine the critical biosynthesis and metabolism processes that coordinated the development and color of brown cotton fiber and to explore the feasibility of using biotechnology to improve the fiber quality of colored cotton through RNA sequencing, KEGG metabolic pathway, weighted correlation network analysis (WGCNA), iTRAQ, and other technologies. The brown NIL genotypes were analyzed to determine the essential genes (proteins) that balance fiber color and quality and regulate metabolic pathways. Because of its connection with pigment
production, the fiber formation of natural colored cotton is more complex. Therefore, transcriptomics and proteomics research is essential for classifying proteins involved in complex biological networks, which assist to understand the formation of natural brown cotton fiber and pigment biosynthesis.

Results

Transcriptome data analysis of brown cotton and white cotton fiber development period

Transcriptome data analysis of 18 samples recorded 118.86 Gb of clean data, whereas each sample reached 6.37 Gb of clean data, and the percentage of Q30 bases was more than 87.57%. The evaluation results of other indicators are shown in Table S1 [Additional file 1]. For further assembly analysis, the transcriptome sequencing findings satisfy the quality standards. The clean reads of each sample were aligned with the reference genome of upland cotton TM-1, and the alignment efficiency ranged from 95.41% to 97.63% (Table S2) [Additional file 2]. The purpose of transcriptome sequencing was to compare the gene differential expression changes of Z161 and RT in adjacent developmental stages. This study screened 36,327 DEGs in Z161 and RT at adjacent developmental stages, and RT-10/0 had 24,936 DEGs. While 18,346, 22,836 and 14,877 DEGs were exhibited at RT-20/10, Z161-10/0 and Z161-20/10, respectively (Fig. 1a). In the three developmental stages, the number of up-regulated DEGs in the Z161-10/0 comparison group was more than that of down-regulated DEGs compared with RT-10/0, RT-20/10 and Z161-20/10. The number of up-regulated DEGs in the other groups was slightly less than down-regulated (Fig. 1a).

Correlation between gene transcription and protein expression during fiber development in brown cotton and white cotton

According to the protein abundance level, when the fold change is > 1.2 (q-value<0.05), it is regarded as a differential protein. Results depicted that the differentially abundance proteins in RT-10/0 and RT-20/10 were 1,751 and 622, respectively. Whereas the DAPs in Z161-10/0 and Z161-20/10 were 1901 and 453, respectively (Fig. 1b). In all developmental stages, the number of up-regulated DAPs in the RT-20/10 comparison group was higher than that of down-regulated DAPs. In contrast, RT-10/0, Z161-10/0 and Z161-20/10, the up-regulated DAPs in the comparison group were less than the down-regulated DAPs (Fig. 1b). We cross-analyzed the four differentially expressed comparison groups’ differentially expressed genes and proteins and found that the RT-10/0 and Z161-10/0 comparison groups shared 1,101 and 1,164 genes (proteins), respectively. RT-20/10 and Z161-20/10 comparison groups shared 374 and 229 genes (proteins), respectively (Fig. 1c). We believed that the low number was due to tissue specificity, as the 0 DPA samples were ovules and the 10 DPA samples were fibrous tissue. In this study, we quantified 12,324 proteins and 2,368 differentially expressed proteins (Table S3) [Additional file 3]. The number of expressed genes was 48,259, and the differentially expressed genes were 36,327 (Table S4) [Additional file 4]. Venn diagram of Fig. 1d was finally divided into six groups, namely Group I: protein and gene expression gene sets with the same trend; Group II: Gene sets with opposite directions in protein and gene expression; Group III: Gene sets with no changes in protein and differentially expressed genes; Group IV:
Gene sets with only differentially expressed genes; Group V: No changes in genes, protein differences expression gene set; Group VI: only protein differentially expressed gene set (Fig. 1d).

In this study, the association analysis between differential expression proteins and genes was carried out. All quantitative proteins and genes were positively correlated, and correlation was low, with correlation coefficients of 0.487 (10/0) and 0.159 (20/10), while the expression trends were the same for the differential proteins and genes (Fig. 2). The expression was positively correlated, the correlation was high with correlation coefficients 0.831 (10/0) and 0.815 (20/10), the expression trend was opposite, the differential protein and gene expression were negatively correlated, and the correlation coefficient was -0.54 (10/0) and 0.642 (20/10).

**KEGG metabolic pathway enrichment analysis of differentially expressed genes and proteins**

Group I described the collection of genes with the same protein and gene expression trend. Significant enrichment analysis can determine the main biochemical, metabolic and signal transduction pathways of Group I, Group II, Group III, and Group V genes (proteins). This study found 2,210 differentially expressed genes (proteins) \((P\text{-value} < 0.05)\) annotated into 50 classified metabolic pathways (Fig. 3a). In the 10/0 comparison group, as compared to white cotton RT, brown cotton Z161 enriched significantly like ubiquinone and other terpenoid quinones biosynthesis (ko00130), flavonoid biosynthesis (ko00941), phenylalanine metabolism (ko00360), histidine metabolism (ko00340), glutathione metabolism (ko00480), tyrosine metabolism (ko00350), and many genes (proteins) were up-regulated (Fig. 3b). So revealed that brown cotton mainly involved more genes in synthesizing secondary metabolites and amino acid metabolism. In addition, in the 20/10 comparison group, the main significantly enriched pathways in Z161 were alanine, aspartate and glutamate metabolism (ko00053), glyoxylate and bismuth carboxylic acid metabolism (ko00630), carbon metabolism (ko01200), carbon fixation in photosynthetic organisms (ko00710), amino acid biosynthesis (ko01230) and nitrogen metabolism (ko00910) and most of the genes were down-regulated (Fig. 3b). The brown cotton fiber was regulated through genes related to amino acid metabolism and energy metabolism than white cotton.

According to the KEGG enrichment results in the 10/0 comparison group, the number of genes was less in Group II, and Z161 consisted of ribosome (ko03010) and mRNA monitoring pathway (ko03015) (Fig. S1A) [Additional file 5]. In the 20/10 comparison group, Z161 compared with RT, the only identified pathway was ribosome (ko03010). More differential genes were observed in group III (Fig. S1B) [Additional file 5]. In RT 10/0, metabolic pathways were ascorbic acid and uronic acid metabolism (ko00053), butyric acid metabolism (ko00650), histidine metabolism (ko00340), phenylalanine metabolism (ko00360), tropane, piperidine, and pyridine alkaloid biosynthesis (ko00960) and tyrosine metabolism (ko00350). Metabolic pathways unique to Z161 10/0 were lipoic acid metabolism (ko00785), glycine, serine, and threonine metabolism (ko00260), lysine biosynthesis (ko00300), sesquiterpenes and triterpenes biosynthesis (ko00909), and sulfur metabolism (ko00920). Metabolic pathways exceptional to RT 20/10 contained basal transcription factor (ko03022), plant circadian rhythm (ko04712), glycosylphosphatidylinositol (GPI)-anchor biosynthesis (ko00563), ribosome (ko03010), and splicing
body (ko03040). Pathways unique to Z161 20/10 intricated amino acid biosynthesis (ko01230), carbon metabolism (ko01200), glyoxylate and dicarboxylic acid metabolism (ko00630), peroxisomes (ko04146), ribosomes in eukaryotes biogenesis (ko03008), and RNA polymerase (ko03020). Hence, these genes were only differentially expressed at the transcriptional level during the development of brown cotton and white cotton fibers, mainly enriched in amino acid metabolism pathways at the early stage of elongation and mainly enriched in energy metabolism genetic information conversion in the later stage.

In Group V, the number of differential proteins was relatively less, but the significantly enriched pathways were more common. A comparison of KEGG enrichment results between brown cotton and white cotton (Fig. S1C) [Additional file 5] revealed glycolysis/gluconeogenesis (ko00010), secondary metabolite biosynthesis (ko01110), flavonoid biosynthesis (ko00941), amino acid biosynthesis (ko01230), starch and sucrose metabolism (ko00500), flavonoid and flavonol biosynthesis (ko00944), fructose and mannose metabolism (ko00051) and pentose phosphate pathway (ko00030). These results disclosed that the pigment accumulation metabolic proteins pathways were more active during the development of brown cotton fibers. However, these metabolic proteins pathways that require much energy during fiber development were enriched and ultimately led to brown cotton fiber quality might be better than white cotton.

**Comparative analysis of STEM expression trends in the transcriptome of brown cotton and white cotton during fiber development**

The STEM was used to analyze and identify the expression patterns of DEGs in the four fiber development stages of Z161 and RT (Fig. 4). The differentially expressed genes were divided into the eight most representative expression trend clusters, and then three significantly enriched expression profiles were screened according to P<0.05: expression profile 0 (P0; continuous down-regulation), expression profile 1 (P1: early down-regulation) and expression profile 3 (P3: late down-regulation). Interestingly, the highest number of differential genes of both lines were expressed in profile 1 (Fig. 4a). We further screened out the standard and unique DEGs of the same expression profile of the two lines through Venn diagram (Fig. 4b). In addition to RT in P6 (up-regulated in the early stage) and P7 (continuously up-regulated), the number of unique DEGs was less than Z161, and more than Z161 in other expression profiles.

To further understand the biological functions of the genes, we performed KEGG enrichment analysis on the respective unique DEGs contained in the eight expression profiles of Z161 and RT. The continuously up-regulated in the P7 involved in glyoxylic acid and dicarboxylic acid metabolism (ko00630) and carbon metabolism (ko01200) of Z161, amino sugar and nucleotide sugar metabolism (ko00520), starch and sucrose metabolism (ko00500), arginine and proline metabolism (ko00330) and biosynthesis of cutin, suberine and wax (ko00073) (Fig. 4c). Most of them also appeared in the P4 (late up-regulated) of Z161. These were involved in carbohydrate metabolism, energy metabolism and lipid metabolism, suggesting that Z161 extends along with fiber elongation. The relative activity of these two pathways may be responsible for the difference between Z161 fiber quality and RT. In addition, the metabolic pathways that
were consistently down-regulated in P0, Z161 relative to RT were unique to ribosome biogenesis (ko03008), RNA degradation (ko03018), and mismatch repair (ko03430). As well as early down-regulated metabolic pathways in P1 were unique terpenoid biosynthesis (ko00900), thiamine metabolism (ko00730) and lipoic acid metabolism (ko00785) metabolic pathways (Fig. 4d). These results explained that the expression activity of genes of the metabolic pathway in Z161 was fully inhibited from retaining more energy metabolites for normal fiber development and pigment synthesis.

This study identified that the genes pathway of carbohydrate and energy metabolism were significantly enriched in the fiber development stages of the two varieties. To further investigate the regulatory mechanism of the significantly up-regulated genes related to fiber development in brown cotton compared with white cotton, we analyzed the transcriptional performance of carbohydrate-active enzymes and transcription factor family genes. Carbohydrate-Active enZymes (CAZy) of carbohydrate metabolic pathways were involved in the biotransformation of sugars or sugar derivatives. The catalytic activity of these enzymes can assemble monosaccharides into oligosaccharides or polysaccharides, as well as with nucleic acids, proteins, lipids substances, polyphenols and other compounds. They are responsible for carbohydrate synthesis (via glycosyltransferases [GTs]), polysaccharide lyases [PLs], degradation (glycoside hydrolases [GHs], active accessory enzymes [AAs]) carbohydrate esterases [CEs], and carbohydrates compound recognition (Carbohydrate-Binding Module [CBM]). Total 2,737 CAZy genes of 6 families in upland cotton were obtained from the database of co-expression networks with functional modules for diploid and polyploid Gossypium (ccNET: http://structuralbiology.cau.edu.cn/gossypium/). However, Transcription factors were essential in the four-cotton fiber cell growth and development stages. We obtained a gene list of all members of the carbohydrate-active enzyme family from the diploid and polyploid cotton functional module co-expression network database (ccNET) (http://structuralbiology.cau.edu.cn/gossypium/annotation_sum.php). A list of 5,022 member genes of 58 gene families of upland cotton was obtained from the Plant Transcription Factor Database (PlantTFDB v5.0) (http://planttfdb.gao-lab.org/index.php?sp=Ghi). Then crossover analysis was performed with gene sets that were up-regulated in at least one of the four comparison groups. Finally, 580 up-regulated carbohydrate-active enzymes and transcription factors were obtained. Among them, GTs genes were the most (164) abundant, followed by GHs (107), and the third is NAC family (31). We focused on analyzing the performance of 11 gene families in four comparison groups. We found that 99 GT genes had higher differential expression folds in the RT-20/10 comparison group than Z161-20/10 comparison group, and 92 GT genes had higher differential expression folds in the Z161-10/0 comparison group than RT-10/0 comparison group (Fig. 4e and Table S5) [Additional file 6]. The other three families, i.e., GHs, CEs and CBs (72 in total), behaved similarly (Fig. 4f and Table S5) [Additional file 6]. From these results, carbohydrate-active enzyme genes were predominantly expressed in the early stage of elongation of brown cotton fibers, and predominantly expressed in the later stages of elongation of white cotton fibers. For the transcription factor family, MYB transcription factors and other transcription factors form protein complexes, essential in fiber development. For example, MYB-type transcription factors can include a MYB-bHLH-WD40 protein complex with bHLH-type transcription
factors and WD40-type transcription factors to perform functions. This study found 15 R2R3-MYB transcription factor genes, 12 bHLH genes, and 24 WD40 genes. The differential heat map analysis found ten triplet genes whose differential expression folds in the Z161-10/0 comparison group were more significant than those in the RT-10/0 comparison group. There were five triplet genes with higher differential expression folds in the Z161-20/10 comparison group, but no significant difference in the RT-10/0 comparison group (Fig. 4g and Table S6) [Additional file 7]. The gene families of other transcription factors with higher fold changes in the two comparison groups were C2H2 (11), ERF (11) and NAC (23), while those in Z161 were 12, 11, and 7, respectively (Fig. 4h and Table S7) [Additional file 8]. Finally, we have identified several different transcription factor genes related to the development of brown cotton and white cotton fibers. Nevertheless, the regulatory roles of these transcription factors in the development of upland cotton fibers with different fiber colors and the improvement of cotton fiber quality the utilization value needs to be further investigated.

Construction of gene-weighted co-expression network and analysis of modules

By analyzing the transcriptomes of the four developmental stages of the two cotton varieties, we finally obtained 36,327 differentially expressed genes (|log2FC| > 1, FDR < 0.05) in the adjacent developmental stages of Z161 and RT. Before constructing the gene co-expression network, we filtered these genes to select only up-regulated genes in the three comparison groups. Then, we filtered the FPKM values matrix to remove the genes with less than 5. Ultimately, 9,999 genes were finally selected for WGCNA analysis. The weight values were first calculated using the WGCNA package to make the network conform to the scale-free network distribution. Co-expression network was constructed with a soft threshold = 12. The correlation of expression levels among genes was used to create a gene clustering tree. The dynamic cutting tree method was used to cut and distinguish the branches and generate different modules. According to the module similarity (75%), the modules with similar expression patterns were merged and divided into 12 gene co-expression modules (Fig. 5a and Table S8) [Additional file 9]. Each color represents a module, and grey modules represent genes that cannot be assigned to any module. The ME midnight-blue module had the largest number of genes (1,744), followed by ME-brown (896), and the ME-orange module had the least number of genes (37) (Fig. 5b). Six of the 12 modules were highly specific with the sample (and tissue) (|r|>0.60, p<0.05). Among them, ME green was highly correlated with Z161_20 (r=0.99, P-value=5e-15), ME greenyellow module was highly correlated with Z161_20 (r=0.77, P-value=2e-04), as well as ME lightgreen was highly correlated with fiber (r=0.92, P-value=6e-08) (Fig. 5b and Fig. S2) [Additional file 10].

Results revealed that the gene expression levels in each module were highly correlated. While the ME expression levels were also highly correlated with the overall expression levels of the modules, indicating that the ME of the target module could adequately represent the prevalent genes in its module (Fig. 5c). To obtain the hub genes in the above three modules, we rank the genes in the powerful module by the degree of connectivity and select the top five genes as the hub genes of the module (Table S9) [Additional file 11]. The heatmap analysis of the expression of these 15 hub-genes found that each corresponding module's expression was the highest in the related highly correlated samples or tissues (Fig. 5d). Such as
Gh_D03G0021 (CAT2) and Gh_A13G2171 (ACX4) were involved in the redox process; Gh_D11G2916 (APRR1), Gh_Sca133868G01 (PETE) and Gh_D05G2375 (AGL11) were involved in the regulation of DNA templated transcription (Table S9) [Additional file 11]. Finally, the 15 genes and their related candidate genes were used to draw the WGCNA regulation map by Cytoscape software (Fig. 6 and Table S10) [Additional file 12]. Each node in the network represents a gene, and the connecting line represents the regulatory relationship between genes.

Firstly, the ME green was significantly associated with brown cotton 20 DPA fibers. The main metabolic pathways were enriched in Carbon fixation in photosynthetic organisms (ko00710), carbon metabolism (ko01200), glyoxylate and dicarboxylic acid metabolism (ko00630), pentose phosphate pathway (ko00030), glycolysis/gluconeogenesis (ko00010) (Fig. S3A) [Additional file 13]. ME greenyellow was also associated with brown cotton 20 DPA fibers. Metabolic pathways were enriched in SNARE interactions in vesicular transport (ko04130), RNA transport (ko03013), Oxidative phosphorylation (ko00190), Ribosome (ko03010), Fatty acid degradation (ko00071) (Fig. S3B) [Additional file 13]. The ME lightgreen module was significantly associated with fiber tissue. The main metabolic pathways were enriched in Amino sugar and nucleotide sugar metabolism (ko00520), Phagosome (ko04145), Oxidative phosphorylation (ko00190), Metabolic pathways (ko01100), Fatty acid degradation (ko00071) (Fig. S3C) [Additional file 13]. The most active metabolic pathways were related to fiber growth and development, such as oxidative phosphorylation and fatty acid degradation, regardless of brown cotton. For 20 DPA, different fiber color varieties have different metabolic pathways involved in core genes that play an essential role. Brown cotton was biased towards energy metabolism and oxidative phosphorylation, which could be understood as better in the fiber elongation period. It plays a significant role in fiber quality and pigment synthesis.

**Discussions**

**Transcriptome and proteomics association analysis**

Biotechnology and high-throughput sequencing technology have been widely used in plant genomics and transcriptomics in recent years. Hence, it provides a valuable tool for studying the molecular regulatory mechanism of plant growth and development and uncovering the gene functions. Proteins are the embodiment and executor of gene functions, and the synthesis of a protein may also involve the transcription and post-transcriptional modification of multiple genes. Therefore, the spatiotemporal relationship and network regulation of gene expression during plant growth and development cannot fully reflect the gene transcription regulation level. Through the association analysis of transcriptomic and proteomic data, the intrinsic links between genes and proteins can be deeply explored. These results are also validated by several prior transcriptome and genome research based on similar research [21].

The molecular pathways of pigment production in brown colored cotton have previously been investigated using transcriptome or proteome analysis [12, 22–24]. To reveal the molecular regulatory mechanism of brown cotton fiber development from the gene transcription level and protein expression
level, the fibers (ovules) of brown cotton and white cotton at four stages were simultaneously analyzed by transcriptome and proteome sequencing. The sequencing results were correlated at three levels: identification, quantification and differential expression of transcripts and proteins. Proteins have a low correlation with the transcriptome at the quantitative level (Fig. 2a). However, the correlation was higher at the differential expression level (Fig. 2b) and slightly lower when the trend was reversed (Fig. 2c). The previous reports suggested that all quantified and differentially expressed proteins exhibited a low correlation between transcriptome and protein changes [25–26].

In contrast, the highest correlation was also reported between transcriptome and proteome profiles [27–28]. Post-transcriptional regulatory mechanisms are one of the biological reasons for the low correlation between transcriptome and proteome data expression profiles [29–31]. It can also be seen that the correlation between the transcriptome and proteome profiles was affected by different developmental stages, and similar findings were also reported by Peng et al. [32]. The low correlation between transcriptome and proteome may be affected by the following factors, such as translation efficiency and mRNA expression variability [33]. This also reflected that protein levels were not only determined by their transcriptional levels but also regulated by developmental stages, post-transcriptional translation, and other processes. In addition, post-transcriptional modification and protein turnover also influenced protein levels. Hence, the transcriptome–proteome study revealed a skewed relationship between transcriptional activity and protein synthesis, highlighting the importance of plant post-transcriptional control [34]. On the other hand, by comparing the transcriptome and proteome of white cotton in parallel, we could screen and identify critical factors specific to brown cotton fiber development (Fig. 3). Hence, the molecular regulation mechanism of brown cotton fiber development can be described more comprehensively.

The dynamic changes in energy metabolism are closely related to the elongation process of brown cotton fibers

Carbohydrates are necessary for living cells and an essential source of energy and carbon during their biological reactions. Energy/carbon metabolism is active during cotton fiber development, providing energy and intermediates for fiber cell elongation and secondary wall synthesis. In this study, the consistently up-regulated genes in brown cotton fiber elongation (10 DPA-20 DPA) were mainly specifically enriched in glyoxylic acid and dicarboxylic acid metabolism (ko00630), carbon metabolism (ko01200), glucosamine and riboside sugar metabolism (ko00520) and starch and sucrose metabolism (ko00500) (Fig. 4c). The glyoxylic acid cycle is a replenishing pathway of the TCA cycle, enabling the TCA cycle to have high-efficiency production functions and provide relevant intermediate metabolites for many important biosynthetic reactions. The decomposition of sucrose into fructose and glucose can increase the intracellular osmotic pressure, which is beneficial to the rapid protrusion and elongation of the epidermal cells of the cotton fiber, which has been proved to be closely related to the extension of fiber cells [35]. Sucrose synthase (SuSy) plays a crucial role in plant sucrose catabolism and catalyses and reversibly converts sucrose to UDP-fructose and UDP-glucose [36]. Studies have shown that glucosamine and nucleoside sugar metabolism is highly expressed during fiber development, proving that UDP- rhamnose can significantly accelerate fiber elongation, and the pectin precursors UDP-
arabinose and UDP-galacturonic acid promote fiber elongation [37]. Three essential enzymes involved in glucose metabolism in fibre cells are invertase (INV), SuSy, and sucrose phosphate synthase (SPS) [38].

Carbohydrate-active enzymes were involved in many catalytic activities in the biotransformation of sugars or sugar derivatives that assemble monosaccharides into oligosaccharides or polysaccharides and bind nucleic acids and proteins lipids, polyphenols, and other compounds. In addition, they are responsible for carbohydrate synthesis. In our study, a large number of carbohydrate-active enzyme genes, especially glycosyltransferases [GTs] genes, were found to be highly expressed in brown cotton, followed by glycoside hydrolases [GHs] (Fig. 4e and f). Glycosyltransferase is mainly involved in synthesizing the main component of the cotton fiber cell wall, the xylan backbone. Research also revealed that glycosyltransferases participate in xylan synthesis [39], regulation of pectin biosynthesis [40] as well as the synthesis of xyloglucan and play an essential part in sustaining the plant cell wall's normal morphology [41]. Simultaneously, glycoside hydrolases are involved in glycan biosynthesis and remodulation, defence, energy mobilization, signalling, symbiosis, glycolipid metabolism and secondary plant metabolism [42]. The xylose transferase activities of IRX9 and IRX14 mutants were drastically reduced, proving that IRX9 and IRX14 were glycosyltransferases involved in xylan backbone synthesis [43]. Further studies using the tobacco BY2 cell system demonstrated that IRX9 and IRX14 were jointly involved in synthesizing the xylan backbone [44]. In addition, glycosyltransferase GT47 family members IRX10 and IRX10-L also have xylan xylosyltransferase activity, and there is some functional redundancy between them [45].

**MBW triplet complex and transcription factors may be involved in the regulation of genes related to brown cotton fiber development**

In this study, by using STEM and WGCNA analysis to compare the differentially expressed genes in two adjacent periods of cotton fiber development of two different fiber colors (0 DPA, 10 DPA, 20 DPA), it was found that the genes of the transcription factor family, NAC, MYB (especially R2R3-MYB type), C_{2}H_{2} zinc finger proteins, and WD40 were involved in early fiber development and fiber elongation (Fig. 4g and h). For MBW triplet complex genes, many genes were up-regulated in the 10/0 comparison group of the two varieties simultaneously (some were down-regulated in the 20/10 comparison group). On the contrary, the 20/10 comparison group was up-regulated. In terms of expression profiles of other transcription factors, more MYB (non-R2R3-MYB type) and NAC transcription factor genes were predominantly expressed in the later developmental stages of white cotton. MYB transcription factors were involved in multiple links and plant growth and development processes, especially in the regulation of hair trichome development. They had their specific regulatory model, namely the MYB-bHLH-WD40 (MBW) triplet complex [46]. The structural and developmental similarity between Arabidopsis epidermal hairs and cotton fibers provides a good reference for cotton fiber research [47]. The whole-genome sequencing of *A. thaliana* has been completed, and many genes closely related to the development of epidermal hair have been cloned. GL1-GL3-TTG1 controlled epidermal hairs by forming an MBW triplet complex, while anaphase growth is controlled by the HD-ZIP protein GL2 [48]. Therefore, researchers cloned genes homologous to members of the MBW triplet complex related to the development of Arabidopsis epidermal hairs from cotton fiber.
tissue and discovered the role in cotton fiber growth and development. *GhMYB25* [44, 49], *GhMYB1-6* [50], *GhMYB109* [51–52] and *GhHOX1*, *GhHOX3* [53] were reported for their role in fiber elongation and development.

It is suggested that these genes may be involved in regulating essential pathways of fiber elongation of brown cotton. Low correlation between transcript and protein, indicating complex post-transcriptional regulation. Similar genes need to be explored more by transforming *Arabidopsis* and cotton. The phenotypic changes of *A. thaliana* and transgenic cotton were observed whether these transgenic lines produced early changes in plant type, reproductive growth, and changes in fiber quality. Combined with the function of the gene and expression analysis, the molecular mechanism of the gene can be discussed and utilized.

**Conclusions**

The brown cotton near-isogenic lines Z1-61 and RT was comprehensively analyzed with transcriptomic and proteomic approaches. The expression patterns of genes/proteins related to amino acid metabolism and energy metabolism were found through the association analysis of DEGs and DAPs analysis. By comparing the transcriptome co-expression patterns of the two lines, the results showed that the expression patterns of carbohydrate synthesis related enzyme genes (CAZy) and fiber development related MYB-bHLH-WD40 protein complex genes were significantly different among the lines. Furthermore, we found that oxidative phosphorylation, carbon metabolism, sugar metabolism and fatty acid metabolism pathways were enriched in brown cotton fiber elongation stage and secondary wall thickening transition stage (20DPA), which can be understood as better playing a major role in fiber quality and pigment synthesis during fiber elongation via WGCNA analysis. Altogether, our study provides novel insight into the metabolic mechanism of fiber quality enhancement and pigmentation during fiber elongation.

**Materials And Methods**

**Material source and sampling**

The material used in this study was a pair of near-isogenic lines (NILs) selected from the cross progeny of the upland cotton standard lines "TM-1" and "richmondi", Zong 1-61 (Z161) with brown fiber and RT (RT white fiber) with white fiber, respectively. The tissue materials for this study were obtained from the experimental field of Cotton Research Institute, Chinese Academy of Agricultural Sciences in Baibi Town, Anyang, Henan, China. The day of flowering was marked with a label, and 0 DPA ovules, 10 DPA, 20 DPA, and 30 DPA fibers (days post-anthesis, DPA) were collected after flowering, immediately frozen in liquid nitrogen, and stored in a -80°C for protein, RNA and metabolites extraction. Due to the difficulty of extracting total RNA from 30 DPA samples, only three-stage samples (0 DPA ovules and 10 DPA, 20 DPA fibers) were used for the transcriptome study.
**Fiber (ovule) RNA extraction**

RNAplant Plus plant total RNA extraction reagent (DP441) (Tiangen, Beijing, China) was used to extract complete RNA from leaf samples according to the manufacturer’s protocols. The cDNA was obtained by reverse transcription using TaKaRa Co. (Dalian, China). There were three biological replicates for each period in both varieties. The libraries’ construction and sequenced were entrusted to the BGI Gene Technology Co. (Shenzhen, China).

**Fiber (ovule) transcriptome data analysis**

The raw image data obtained by sequencing was converted into raw sequence data, and the conversion results were saved in the FASTQ file format. The original sequence data is modified with adaptor sequence, unknown base N ratio > 10% and low quality, and a series of de-cluttering data processing such as the removal of various reads, and finally, clean reads are obtained. The clean reads were compared to the reference genome of upland cotton (www.cottongen.org/species/Gossypium_hirsutum/nbi-AD1_genome_v1.1) using HISAT2 [54] for sequence alignment. Aligned reads were assembled and quantified using StringTie software [55]. Gene expression levels were calculated by FPKM (fragments per kilobase of exons per million) values. For samples with biological replicates, DESeq2 was useful for evaluating differential expression analysis between sample groups to acquire differentially expressed gene sets between two biological conditions [56]. Differently expressed genes (DEGs) were defined as FDR < 0.05 and fold change ≥ 2.0. GO functional analysis (cellular component, molecular function, biological process) and KEGG pathway analysis was performed on selected DEGs.

**Fiber (ovule) proteome data analysis**

Fiber (or ovule) protein extraction and concentration determination were performed according to Qin, et al. [57]. For each developmental stage, three biological replicates were used. Each replicate was harvested from 5-8 bolls, stripped of the fibers (or ovules), and their proteins extracted. iTRAQ proteomics data were quantitatively analyzed with IQuant software [58]. Proteins were filtered again with FDR 1% (protein-level FDR<=0.01) to remove false positives, and the strategy used was Picked protein FDR. Differentially abundant proteins (DAPs) were used for comparison with at least two occurrences of biological triplicates with a q-value < 0.05 and a fold-change ratio > 1.2 or < 0.833 to be considered significant. DAP’s functional properties were analyzed by the enzymatic code of the DAP protein and uploaded to a commercial database (KEGG; http://www.kegg.jp/), and annotated sequences and metabolic KEGG pathways were used to map the data. PXD020168 is the dataset identifier for mass spectrometry proteomics data that has been submitted to the protein exchange consortium via the iProX partner repository.

**Correlation analysis of transcriptome and proteome data**

Except for the proteomic information associated with the 30 DPA samples, only three-stage samples (0 DPA ovules, 10 DPA, and 20 DPA fibers) were used for combined proteomic and transcriptomic analysis.
The method of Lan, et al. [59] was used to calculate the Pearson correlation coefficient at the protein and transcription level to determine the strength of the correlation between protein and mRNA. The content of the association analysis between the proteome and the transcriptome includes (1) qualitative association analysis of protein and mRNA; (2) quantitative association analysis of protein and mRNA; (3) analysis of the expression trend of protein and mRNA.

Differentially expressed genes (proteins) expression patterns and KEGG enrichment analysis

The Short Time-series Expression Miner software (STEM) analyzed and identified the expression patterns of DEGs in the three fiber developmental stages of Z161 and RT. To investigate the signal transduction pathways and metabolic pathways involved in DEGs (proteins), KEGG significant enrichment analysis was performed (or proteins) (Q-value ≤ 0.05) from https://www.omicshare.com/tools/.

Construction of weighted gene co-expression network (WGCNA)

Weighted gene co-expression network construction was performed using the WGCNA software package in the R program [60]. A total of 18 transcriptome samples (3-time points, 2 species, 3 replicates) were assessed with FPKM >1 and normalized gene expression reads counts matrix. The top 50% of genes with the highest variations were selected for WGCNA. After threshold screening, β = 12 was finally selected to perform power processing on the original scaled relationship matrix to obtain an unscaled adjacency matrix. The dynamic shearing algorithm was used for gene clustering and module division. The minimum number of genes in a module was 30 (minModuleSize=30), and the threshold for merging similar modules is 0.16 (cut Height=0.16). The value of principal component 1 (PC1) was referred to as the Module Eigengene after Principal Component Analysis (PCA) was conducted on all genes in each module (ME). The correlation coefficient r and corresponding p-value between the module eigenvectors of each module and different samples were calculated for screen-specific modules related to disease resistance. Whereas, r > 0 reflects positive correlation and r < 0 represents a negative correlation. In this study |r| > 0.6 and p modules < 0.05 were further analyzed as specificity modules. After screening out the target genes with the highest correlation coefficient among modules, the genes corresponding to their expression correlation's top 5 weight values were extracted. The Gephi-0.91 software was used to draw the WGCNA graph. The gene regulatory relationship network map can help us accurately screen candidate genes that have a potential regulatory relationship with target genes and can use the functions of known genes to predict the functions of unknown genes.

Abbreviations

DPA: Days post anthesis; STEM: Short Time-series Expression Miner; NCC: naturally colored cotton; DGE: digital gene expression; CSSLs: chromosomal segment substitutions lines; NILs: near-isogenic lines; FPKM: fragments per kilobase of exons per million; DEGs: differently expressed genes; DAPs: differentially abundant proteins; PC1: principal component 1; PCA: Principal Component Analysis; ME: module eigengene; MBW: MYB-bHLH-WD40; TM-1: Texas Marker-1; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes.
Declarations

Acknowledgments

We thank Jindong Shang of the National Supercomputing Center in Zhengzhou, Zhengzhou University, for providing the computing resources. the National Medium-term Gene Bank for cotton in the Institute of Cotton Research of the Chinese Academy of Agricultural Sciences (ICR, CAAS) provided us with cotton seeds. We are grateful for their kind helps.

Author’s contributions

Z.P. conceived and designed the experiments. P.D. and S.H. directed the methodology and software analysis. Z.W. analyzed and summed all the data. Y.J. managed and collected the plant tissues. and all the other authors collected the data evidence and revised the manuscript. Z.W., A.R. wrote the manuscript with inputs and guidance from Z.P. and X.D. All authors read and approved the nal manuscript.

Funding

This work was supported by the Major Projects for Transgenic Breeding of China (2016ZX08005-005).

Availability of data and materials

The RNA-seq raw data set used in this study has been uploaded to Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) under BioProject PRJNA766762. The mass spectrometry data have been deposited to the ProteomeXchange Consortium (http://proteomecentral. proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD020168.

Ethics approval and consent to participate

The experiments did not involve endangered or protected species. No specific permits were required for these activities. because the cotton seeds used in this study were obtained from the National Medium-term Gene Bank of Cotton in China., which is managed by the Institute of Cotton Research of the Chinese Academy of Agricultural Sciences. All methods were carried out under relevant guidelines.

Consent for publication

Not applicable

Competing interests

The authors declare no conflicts of interest.

References


**Figures**

![Figure 1](image-url)

**Figure 1**

Data analysis of differentially expressed genes and proteins during fiber development of Z161 and white cotton RT. (a) Comparison of the number of up and down regulated genes in different comparison groups. (b) Comparison of the number of up and down regulated proteins in different comparison groups. (c)
Cross comparison of differentially expressed genes and proteins in different comparison groups. (d) Meaning and quantity of 6 groups (Group I-VI).

**Figure 2**

**Correlation expressed quantitative proteins and genes.** (a/d) Correlation of all quantitative proteins and genes; (b/e) The correlation between protein and gene with same change trend; (c/f) The correlation between proteins and genes with the opposite change trend.
**Figure 3**

**Enrichment analysis of KEGG metabolic pathway of four protein and gene expression patterns.** (a) The KEGG enrichment result of the same expression trend of protein and gene. (b) Heatmap of 11 representative KEGG pathway genes (proteins) in four comparison groups. AAGM: Alanine, aspartate and glutamate metabolism; AAM: Ascorbate and aldarate metabolism; BAA: Biosynthesis of amino acids; GDM: Glyoxylate and dicarboxylate metabolism; PHM: Phenylalanine metabolism; UTQB: Ubiquinone and
Figure 4

STEM expression trend analysis and important gene expression of brown cotton and white cotton fiber transcriptome. (a) Trend and statistical chart of differential gene expression between Z1-61 and RT; (b)
Gene cross analysis of expression profile between Z1-61 and RT; (c) KEGG Enrichment analysis of gene specifically up-regulated by expression profiles of Z1-61 and RT; (d) KEGG Enrichment analysis of gene specifically down regulated by expression profiles of Z1-61 and RT; (e) Heat map of up-regulated genes of glycosyltransferase [GTS] family in four comparison groups; (f) Heatmap of up-regulated genes of glycoside hydrolase [GHS], carbohydrate esterase [CES] and carbohydrate binding module [CBM] in four comparison groups; (g) Heat map of up-regulated genes of MYB-bHLH-WD40 protein complex in four comparison groups; (h) Heat map of up-regulated genes of other transcription factor families MYB, C2H2, ERF and NAC in the four comparison groups.
Figure 5

The construction of weighted gene co-expression network and identification of core genes. (a) Gene co-expression network gene clustering number and modular cutting; (b) Heat map of the correlation between module and traits; (c) The expression levels of all genes and corresponding ME of different modules in each sample. (d) Heat map of gene expression of 15 hub genes in 18 samples of four significance modules.
Figure 6

Gene co-expression networks of the specific modules which significantly associated with different sample (tissue). Gene co-expression network of the green module (a), the greenyellow module, (b), and the lightgreen module (c).

Supplementary Files
This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.xlsx
- Additionalfile10.pptx
- Additionalfile11.xlsx
- Additionalfile12.xlsx
- Additionalfile13.pptx
- Additionalfile2.xlsx
- Additionalfile3.xlsx
- Additionalfile4.xlsx
- Additionalfile5.pptx
- Additionalfile6.xlsx
- Additionalfile7.xlsx
- Additionalfile8.xlsx
- Additionalfile9.xlsx