Proteomics-based models of gene expression and cellular control of cotton fiber development

Running title:
Linking proteome dynamics with fiber development

Keywords: G*ossypium hirsutum*, trichoblast, proteomics, machine learning, cellulose synthase, multi-omics

Youngwoo Lee[1,2], Pengcheng Yang[3], Heena Rani[4], Gideon Miller[3], Corrinne E Grover[5], Sivakumar Swaminathan[6], Olga A Zabotina[6], Jonathan F Wendel[5], Jun Xie[3], and Daniel B Szymanski[1,*]

1 Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN, 47907
2 Center for Plant Biology, Purdue University, West Lafayette, IN, 47907
3 Department of Statistics, Purdue University, West Lafayette, IN, 47907
4 Cereal Crops Research Unit, USDA-ARS, 502 Walnut Street, Madison, WI, 53762
5 Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA, 50011
6 Roy J Carver Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA, 50011
*Correspondence to: szymandb@purdue.edu

Abstract

The cotton textile industry is based on the material properties of terminally differentiated trichoblasts. Tens of thousands emerge from the seed coat epidermis and march through a reproducible developmental program that includes interconversion between cell-autonomous and tissue-based morphogenesis. Proteomic analyses of purified fibers provide a way to associate molecules with cellular processes that directly affect morphological transitions and material properties of the harvested cells. However, insufficient temporal sampling and poor protein

coverage have limited the extent to which proteomic data can predict control mechanisms. Here we quantified the subcellular proteomes of purified fibers daily over a 20-day interval that includes diverse modes of elongation and cell wall remodeling. Thousands of reliable protein abundance profiles were assigned to expression groups and tested for associations with a broad array of phenotypes. These analyses were validated by the presence of proteins with known localization and stage-specific functions and also include novel functional predictions for hundreds of genes with no known function. Integration of proteomic and RNA-sequencing data revealed distinct modes of transcriptional and post-transcriptional control. Collectively proteomics data resource provides a valuable foundation for large-scale gene function predictions and provides a path to engineer cotton fiber material properties.

Introduction

Cotton (*Gossypium spp*.) is the world's most important fiber crop and the foundation of a multi-billion dollar renewable textile industry (Huang et al., 2021; Wilkins et al., 2010). Tens of thousands of unicellular seed coat trichoblasts synchronously differentiate and transform their lenth-to-width aspect ratios to ~3000:1 during development (Stewart, 1975). At maturity, each fiber contains a helically patterned cellulose-rich wall that desiccates after programmed cell death to generate harvestable thin, strong, spinnable fibers (Butterworth et al., 2009; Kim and Triplett, 2001; Seagull, 1992). Fiber size, spinnability, and strength are the output of a morphogenetic program of anisotropic diffuse growth (Ryser, 1977; Yanagisawa et al., 2022) with an often underestimated degree of complexity (Figure 1). Phenotyping daily reveals an early tapering phase that is completed by 2 days post anthesis (DPA), during which the hemispherical fiber initial is converted to a thin elongated trichoblast (Graham and Haigler, 2021; Yanagisawa *et al.*, 2022). After tapering the cell wall (CW) thickens as it enters a protracted phase of decelerating anisotropic diffuse growth (Schubert et al., 1973; Wilson et al., 2024), while simultaneously modulating plasmodesmata functions and turgor pressure (Ruan et al., 2001). Given progress in understanding the molecular and biomechanical control of trichoblast morphogenesis (Yanagisawa et al., 2015; Yanagisawa *et al.*, 2022; Yu et al., 2019), there are emerging opportunities to use biotechnological interventions to engineer fiber traits (Liu et al., 2025; Qin et al., 2022; Sun et al., 2025). The goal of this research is to integrate high-temporal resolution quantitative proteomics, RNA-seq, and phenotyping to broadly predict cellular networks that mediate key developmental transitions.

A complexity of studying fiber cell "omics" is their coordinated development with the organ-scale growth and physiology of the enclosing capsule (colloquially termed a "boll") chamber (Figure 1A. and 1B; Supplemental Video 1). Fiber development also includes reversible transitions between cell-autonomous growth and tissue-morphogenesis during which the cells are tightly adhered (Singh et al., 2009). Cotton has long served as a useful model for CW assembly and morphogenesis (Meinert and Delmer, 1977; Seagull, 1993; Tiwari and Wilkins, 1995), and components of the molecular control continue to emerge (Huang *et al.*, 2021). A highly conserved cellulose patterning module underlies polarized diffuse growth in the plant kingdom (Baskin, 2005), and the details of how interactions among cortical microtubules, patterned cellulose synthase (CESA) activity, and CW stress-sensing operate to control cell shape changes

are beginning to emerge (Coen and Cosgrove, 2023; Li et al., 2023). Strong lateral interactions between microfibrils (Yanagisawa *et al.*, 2015; Yanagisawa *et al.*, 2022; Zhang et al., 2021) and interactions between fiber and matrix components generate domains of the CW that deform asymmetrically in response to turgor-generated mechanical stress in the CW (Yanagisawa *et al.*, 2015; Yanagisawa *et al.*, 2022; Zhang *et al.*, 2021). Spatially patterned cortical microtubules influence microfibril deposition throughout development to maintain cell diameter and pattern a helical secondary CW (SCW) (Seagull, 1992; Tiwari and Wilkins, 1995; Wilson *et al.*, 2024; Yanagisawa *et al.*, 2022). Important cotton microtubule-associated proteins are being discovered using forward and reverse genetic approaches (Chen et al., 2024; Rajangam et al., 2008; Zang et al., 2021). The microtubule-CESA module does not function in isolation; actin-based intracellular transport, vesicle trafficking, metabolism and redox control, and hormone signaling function in concert during fiber development (Huang *et al.*, 2021).

Proteomic profiling provides a way to obtain direct measurements of the protein machineries and post-translational controls that enable effective morphogenetic work and coordinate diverse functions, some of which reside in different subcellular compartments. Historically, proteomic analyses of cotton fibers have identified proteins with altered abundances as a function of domestication (Bao et al., 2011; Hu et al., 2013; Qin et al., 2017), hormone or inhibitor responses (Wang et al., 2020a; Yang et al., 2023a), and developmental stages (Jiang et al., 2022; Mujahid et al., 2016; Zhou et al., 2019). In general, the number of discovered proteins is low, and temporal sampling of fiber development is limited to 5-day intervals at best. Gel-free workflows and increasingly sensitive instrumentation enable the quantification of thousands, rather than dozens, of proteins in a typical experiment (Hu et al., 2015; Lee and Szymanski, 2021; Yang *et al.*, 2023a). This presents an opportunity to gain broad insights into the protein dynamics that underlie the numerous developmental transitions during fiber development that are not known.

Here we combined daily sampling and clean cell fractionation procedures to analyze how protein abundances change in distinct subcellular compartments during fiber development. Our analytical pipeline allowed us to generate thousands of reliable protein abundance profiles spanning a broad period of fiber development that encompasses primary cell wall synthesis and the transition into secondary wall synthesis. Clustering methods identified unique protein expression groups that reflected specialized functions of the apoplast, cytosol, and membrane-

associated proteins. Multiple groups had clear temporal transitions and broad memberships across the cell fractions. Correlation analysis with a wide range of phenotypes quantified across the same timeline provided a way to predict suites of genes that coordinate the unidirectional progression of fiber development. Integration of RNA-seq and proteomic data can provide large-scale predictions about general modes of gene expression control (Manzoni et al., 2018; Nie et al., 2006). These results were validated by the presence of proteins with known localizations and stage-specific functions in secondary cell wall synthesis. We used combinations of cross-correlation analysis and machine learning methods to globally predict modes of transcriptional and post-transcriptional control. Collectively, our resource article provides valuable data and analytical methods that provide an entry point to broadly analyze the types of proteins and cellular processes that may control morphology in cottons. Because many of these processes and machineries are conserved across the plant kingdom, the data have broad utility in the analysis of cellular growth control.

Results

Protein profiling of developing cotton fiber by proteomics

Fiber development is reproducible across a wide range of environmental conditions (Jan et al., 2022; Schubert *et al.*, 1973), including fibers that are cultured in vitro (Meinert and Delmer, 1977). A central goal was to generate reliable methods to collect time-series data of protein abundance in multiple subcellular fractions from purified fiber tissue (Figure 1). Methods to isolate highly purified fiber samples and generate high-quality cell fractions with limited cross-contamination with broken cells were developed (Lee et al., 2025). The method enables precursor ion-based quantification of a diverse population of proteins within and between cell fractions. Here we scaled up the procedure to increase coverage of the crude cytosolic (S200) fraction. Each replicate included multiple locules from a single boll and included three replicates daily from 5-24 DPA (Figure 1C). The entire developmental progression, from anthesis at 0 DPA to boll opening at 48 DPA, is shown in Figure 1A and Supplemental Video 1. Protein yields in the different cell fractions are included in Data S1. The earliest timepoint was chosen based on the minimum requirements for adequate protein precipitation from the soluble fractions, and the final time point was selected in order to capture a complete transition to SCW synthesis.

A crude apoplast (APO_T) fraction was generated after careful dissection and brief immersion of the intact locules in the isolation buffer. This protocol minimizes contamination from broken cells and has been shown to generate a distinct apoplast fraction with unique protein multimerization behaviors and composition compared to cytosol and microsome fractions (Lee *et al.*, 2025). After immersion, fibers were manually dissected, homogenized, and subjected to differential centrifugation to generate a crude microsomal (P200) and a soluble (S200) fraction (Figure 1C). Across the 5 to 24 DPA window, 4316 proteins were reproducibly identified and quantified in 2 of 3 biological replicates of at least one cell fraction, providing the most comprehensive cotton fiber proteome coverage to date (Figure 1D; Data S3).

Cotton (*G. hirsutum*) is an allopolyploid with two subgenomes (A, D), and as such the duplicated genes (homoeologs) are encoded by the A- and D-subgenomes, such that they potentially can be distinguished based on diverged peptide sequences (Hu *et al.*, 2015; Lee and Szymanski, 2021). Diagnostic homoeolog-specific peptides were detected for ~55% of the cases and these proteins are annotated with "A" and "D" suffixes following the Locus ID (Data S3). The mRNA and protein search databases (Supplemental Files S1 & S2) were generated based on aggregated “A” and “D” subgenome SNPs relative to a base diploid assembly (Grover et al., 2025). These databases enabled high coverage matches with the LC/MS spectra and accurate comparisons of transcript and protein abundance profiles for specific homoeologs (see below). The search databases and locus ID conventions used here were cross-referenced locus IDs across several commonly used cotton genome assemblies to facilitate gene matching (call out table exactly here??). The P200 fraction consistently had the most diverse protein populations over the time course (Data 3C), and, as observed previously (Lee *et al.*, 2025), was more similar to the APO_T compared to S200. The APO_T fraction primarily consists of membrane-enclosed particles with a highly heterogeneous composition (Lee *et al.*, 2025). Widespread overlap among the different cell fractions can occur because many proteins are dually localized (McBride et al., 2017). Our pipeline generated protein quantification across over 6 orders of magnitude (Figure 1E). The protein types amenable to proteomic analyses are extremely diverse (Data S3). The coverage of metabolism in this dataset is deep (Supplemental Figure 1) and provides broad experimental evidence for the specific gene products that operate in central metabolism, phenylpropanoid biosynthesis (Tuttle et al., 2015; Yang *et al.*, 2023a), fatty acid synthesis (Qin et al., 2007), and cell wall assembly (Pettolino et al., 2022).

A protein was retained for abundance profiling across the time course if it had non-zero values in three or more fractions with at least two of them on 2 consecutive days. The protein population composed of those detected in at least 2 of 3 consecutive DPAs and the subset detected in 10 or more DPAs had similar intensity distributions (Figure 1E; Supplemental Figure 5E), indicating that signal intensity is not the primary determinant of expression profile variability. The raw and normalized intensity data and plotted distributions are provided in Data S3 and Supplemental Figures 2–4. Our deep proteome profiling provided a broad depiction of compartmentalized protein behaviors during development, identifying diverse groups of proteins that coordinate fiber development. However, the widespread occurrence of missing values generated between biological replicates is a widely recognized shortcoming of label-free quantitative proteomics (Hamid et al., 2022; Jörnsten et al., 2005; Wang et al., 2020b). We developed a highly stringent data filtering procedure to minimize the effects of artifactual missing values between DPA and eliminate potentially unreliable profiles and generate a restricted dataset containing informative expression profiles from three subcellular locales (see below and methods).

To identify expression groups, the informative profiles were analyzed using a machine learning-based clustering method termed Self-Organizing Map (SOM) analysis (Kohonen, 1990; Wehrens and Kruisselbrink, 2018). In an SOM, the input data are mapped onto a discrete topological grid formed by a set of groups. The groups are typically organized in a 2-dimensional grid, with rows and columns. As the model is trained on input data, the groups gradually adjust their weights to become more similar to the pattern of the input data in their neighborhood, resulting in clusters of the input profile data. The number of SOM groups was determined by evaluating cluster stability, ensuring that proteins consistently grouped with the same or neighboring clusters as the number of groups increased. This strategy minimized empty or single-member clusters, providing meaningful resolution and interpretability. Each SOM group has a “code” that reflects a central expression pattern of the members (Figure 2A; Data S4). A group was scored as unreliable if: 1) the code profile was punctuated by missing values across the timeline; 2) the profile pattern reflected variability in protein coverage in specific DPAs (Data S3D); or 3) the pattern was defined by strings of 3 adjacent zeros that could simply reflect random gaps in the data that were not filled by interpolation. Following this filtering, 39 groups were classified as unreliable, and the remaining 44 potentially informative groups were subjected

to further filtering to remove unreliable members with discordant features in their expression. A multi-step quality control procedure was then applied to each SOM cluster to exclude outliers, unreliable expression patterns, and low-abundance proteins, yielding high-confidence protein expression profiles (Supplemental Figures 2–5). For each group, member intensities were normalized to a 0–1 range based on their maximum values and averaged to generate representative expression patterns on a consistent scale (Figure 2B). This process resulted in 44 distinct cell fraction protein expression group patterns (12 in APO_T, 23 in P200, and 9 in S200), where a 10% change in the normalized scale corresponds to an about 4-fold difference in abundance (or 2-fold on the log2 scale) due to the high signal intensity of the included profiles. In the end, these 44 expression groups provide large-scale functional data on gene sets and cellular activities (Figure 2B; Data S5), and provide clues about gene functions that may operate during specific developmental windows in particular subcellular locales.

Many of the profiles were non-uniform, with sharp abundance transitions often reflecting a shift to or from consecutive days of non-detection. Those classified as “increasing” had a clearly defined timepoint at which the protein transitioned from below to above the detection limit (Figure 2B). The inverse was true for the profiles classified as “decreasing”. We hypothesize that these clearly varying expression patterns reflect rearrangements of suites of cellular machineries as the fiber population undergoes developmental transitions. Some non-uniform expression groups were defined by a specific subcellular compartment. Gene ontology (GO) enrichment analysis shows that P200 protein expression groups with an increasing pattern have hydrolase activity, guanyl nucleotide binding, and catalytic activity terms, whereas S200 was enriched in response to stress term (Data S6). Constitutively expressed proteins classified as non-uniform were flagged if there was at least a 4-fold increase or decrease in the mean protein abundance of the group members during development (Data S5). These “constitutive increase/decrease” groups may reflect broad reprogramming of gene expression relative to the transition to SCW synthesis (Grover *et al.*, 2025) and/or proteins with unequal activities in the elongation and SCW phases. A large number of proteins had flat relative abundances and were classified as uniform. These proteins may define part of the “constitutivome” that includes essential functions that operate across the developmental phases.

Integration of protein dynamics modules with fiber phenotypes

We next wanted to correlate protein expression profiles with phenotypic transitions. To generate a merged set of profiles aggregating protein abundance dynamics across the cell fractions, we turned again to SOM. After evaluating SOM configurations with 9, 20, 30, and 40 groups, we selected the 20-groups SOM as it minimized empty or single-member clusters while maintaining stable group memberships (Data S7A; Supplemental Figures 6–9). This configuration yielded 15 distinct merged protein expression modules (Figure 3A). 11groups contained clearly distinct and non-uniform profiles (Figure 3A; Data S7). The mean normalized intensities of the merged group members were used to estimate the behavior of the group and test for correlations with phenotypes. Aggregation of 7 developmentally variable fiber phenotypes from the same or similar developmental windows (Swaminathan et al., 2024; Wilson *et al.*, 2024) provided a way to test for significant associations between protein expression groups and phenotypic transitions using correlation analyses (Figure 3C; Data S8).

As an example, the M16 group was most strongly positively correlated with the transition to SCW synthesis, which is characterized by transitions in cellulose microfibril angle, increases in cell wall thickness, and an increasing proportion of cellulose in the cell wall (Figure 3C; Data S8B). M16 contained proteins present in all three subcellular compartments and included many known proteins involved in CW biosynthesis, as well as many previously uncharacterized proteins (Figure 3D; Data S8 and S9). M16 also contained *G. hirsutum* orthologs of each of the SCW-associated CESAs (Grover *et al.*, 2025; Kim et al., 2019) that were induced at ~16 DPA to similar levels (Figure 4D; Data S9A). M19 was negatively correlated with SCW transition and may contain proteins that are specifically involved in the fiber elongation and remodeling of the primary CW (Figure 3C; Data 7C and 8B). M14 was also negatively correlated with SCW transition, but this group was composed solely of APO_T members. By comparing these data with those in a purified cotton EV dataset (Lee *et al.*, 2025), we found that M14 members were either specific to the particulate fraction or dual-localized in the soluble and particulate fractions of the apoplast (Data 7D). M14 contained more than half of the total APO_T proteome. The particulate fraction was enriched in spherical stain-excluding compartments that could be visualized using negative staining and transmission electron microscopy (Figure 3B). To determine if the apoplastic particles (APO_{p200}) had a unique polysaccharide composition, their glycome profiles were compared to those of the P200 and APO_{s200} fractions. Several epitopes related to xylan and xyloglucan were specifically enriched in the APO_{p200} particles compared to the other cell

fractions and many other epitopes were selectively depleted (Data S10). These data provide further evidence for their unique composition and may reflect their mode of generation. M9 and M18 were also APO_T-specific expression groups with similar profiles (Figure 3C). These profiles may reflect the developmentally regulated biogenesis of extracellular vesicles.

M20 declined from maximal levels at 5 DPA and was most strongly correlated with the decelerating relative growth rate and a reduced proportion of pectin (Figure 3C). M20 may contain groups of proteins associated with a fiber state with a faster relative growth rate. M11 and M17 were anti-correlated with relative growth rate and auxin-response factors, ABA receptors, and vesicle trafficking machineries (Data S7C, column T, M11) that enable persistent axial growth and/or early components that prime the cell for the transition to tissue-based morphogenesis (Singh *et al.*, 2009) or SCW synthesis (Pettolino *et al.*, 2022). Many of the merged protein groups correlated with phenotypic profiles, and these analyses provide models of how diverse types of gene products function in multiple cellular locations to coordinate developmental transitions.

Integration of mRNA and protein dynamics: distinct modes of gene expression control

To globally analyze potential modes of gene expression control, we integrated the proteomic profile data with fiber transcript data obtained using the same conditions and developmental time points (Grover *et al.*, 2025). Pearson correlation coefficients (PCC) were calculated for 6809 genes in which protein profiles could be matched to a single homoeolog (see methods, case 1) or a pair of homoeologs (Figure 4A; Supplemental Figures 10-12). We initially focused on proteins with a non-uniform expression pattern, i.e., the mRNA/protein correlations in the “increasing” or “decreasing” categories that were stronger in the P200 and S200 fractions than APO_T fraction (Figure 4B; Data S12). About ~ 48 % of mRNA/protein pairs in P200 and S200 had PCC values > 0.75, and protein abundance could be explained by transcriptional control. In general, the APO_T proteins in both expression groups were weakly correlated with their mRNAs because this cell fraction was dominated by putative extracellular vesicle proteins that may have reduced abundance after ~20 DPA (Figure 3A). In addition to the APO_T groups, those showing a negative correlation could be explained by the spatial and functional compartmentalization of gene expression regulation, highlighting the importance of subcellular localization information to understand mRNA-protein correlations. However, there was evidence

for transcriptionally regulated proteins in the apoplast. A cotton ortholog of secreted signaling molecule STOMAGEN, usually associated with non-cell-autonomous cell proliferation (Cui et al., 2014; Sugano et al., 2010), was downregulated at ~14 DPA and correlated with transcript levels (Data S7).

To conduct a more open-ended test for commonly observed or non-linear patterns of gene expression control, a weighted SOM clustering was conducted by combining the mRNA and protein data for 1228 different P200 proteins with informative profiles (Figure 4C; Data S13). The full clustering result is shown graphically in Supplemental Figure 13. The analysis produced three groups (V8-V10) with a total of 48 members with coherently increasing protein and mRNA expression, and three groups (V81,91,92) with a total of 50 members with coherently decreasing expression (Figure 4C), of which 48 of 98 were captured in the PCC analysis described in Figure 4B. Among the large population of proteins with a more uniform expression, there were 79 genes in which mRNA levels fell by at least 8-fold over the time course (Figure 4C). These examples could reflect post-transcriptional controls in which the mRNA translation efficiency increases or protein stability increases during development. In V30, transcript levels increased 8-fold, but protein levels were stable (Figure 3C), indicating that translation of the mRNA may be regulated or the steady state of the protein may be the result of physical interactions with a stabilizing factor. The SOM-weighted cross-correlation analyses provide quantitative information about how diverse types and timings of gene expression control operate during fiber development.

To partially validate our analyses with knowns, we queried our list of genes predicted to be subject to temporal transcriptional control with those known to be involved in CW assembly or the transition to SCW synthesis (Data S14). For example, among the 67 proteins present in the M16 module, 44 were predicted to be transcriptionally regulated. Several orthologs of documented players in SCW synthesis (e.g., cell wall CESAs, COBRAs, and WAVE-DAMPENED-LIKE2 proteins) were reliably detected in the P200 fraction from fiber cell extracts (Figure 4D) (Chen *et al.*, 2024; He et al., 2024; Wen et al., 2022). GhCESA4-B, 7-A/B, and 8-B proteins were first detected with a ~2-day lag compared to when the corresponding mRNAs first increased. A database of cotton orthologs to predicted Arabidopsis cell wall genes (Cell Wall Genomics database; https://cellwall.genomics.purdue.edu/) was created, and a subset of those were present in the set of predicted transcriptionally controlled genes (Data S14). These data summarized in this resource article provide a glimpse into the regulatory complexities of

fiber development and a source of promoters and gene products that appear to be subject to transcriptional control and more easily modulated in transformed lines.

Discussion

In this resource article, we combine sensitive and accurate quantitative proteomics with the synchronized process of cotton fiber morphogenesis to broadly predict protein networks that orchestrate fiber development. The data are derived from a single cell type that develops in the insulated environment of the developing capsule. Daily sampling revealed an unexpected complexity: an extended phase of growth deceleration in early development, followed by waves of gene activation and suppression that precede the transition to SCW synthesis and multiple phases of cellulose microfibril and cell wall matrix remodeling (Grover *et al.*, 2025; Lee *et al.*, 2025; Swaminathan *et al.*, 2024; Wilson *et al.*, 2024). These transitions are not simply anchored to a general transition to SCW synthesis. Our proteomic dataset provides a conceptual model of how networks of proteins in multiple subcellular locations coordinate cellular functions and fiber phenotypes during development.

The proteomic resources generated here provide experimental support for over 4,300 fiber proteins, and reliable abundance profiles over a wide window of development for more than 2,900. This dataset provides useful time-series data on the localization and abundance of a functionally diverse population of enzymes, hormone signaling proteins like auxin response factors and ABA-receptors, as well as proteins that span the full breadth of molecular functions in the cell. There are many enzymes that are specifically associated with fiber elongation, including those involved in flavonoid, anthocyanin, and wogonin metabolism (Supplemental Figure 1; Data S5 and S6). The roles of these small molecules in fiber morphogenesis merit further analyses. All of the proteomic and associated RNA-seq data are publicly available in a simple search interface (link to be provided).

The SOM clustering method and cross-correlation with time-resolved phenotypes were used to generate models of how protein groups in multiple cellular compartments are linked to specific developmental transitions (Figure 3). The merged profiles captured constitutively expressed proteins with uniform (M2,M3,M4,M5), increasing (M6,M7), and decreasing (M9,M10) abundances across the timeline. These proteins likely provide essential functions that

may be more likely to be subject to post-translational controls. Many proteins were assigned to groups with clear time points of abundance change (M11,M14,M16-M20). This approach enabled us to identify groups of up- and down-trending proteins that transition at a similar time point. There were expression group pairs that were most strongly associated with growth rate variability (M20,M17) and others with the transition to SCW synthesis (M16,M19). Gene knockout experiments will be needed to determine how accurately these gene products are linked to these specific developmental transitions.

In the case of the transition to SCW synthesis, many key players are conserved across species and cell types with a thickened SCW (Kumar et al., 2016). Our dataset contains orthologs of each of the three CESA clades (Wen *et al.*, 2022) and COBRA-like (He *et al.*, 2024) proteins that are known to influence SCW synthesis in cotton, and many of them appear to be transcriptionally regulated (Figures 3D and 4D). SCW patterning is a complex process that includes a radical reorganization of the cytoskeletal, vesicle trafficking, and cell wall assembly systems (Hoffmann et al., 2021). This proteomics analysis provides new insights into the particular CESAs and systems of proteins that operate prior to (M19) and after (M16) the transition to SCW synthesis. Some of expression groups may reflect reversible transitions in cell-cell adhesion as fibers interconvert between cell-autonomous and tissue-like morphogenesis at ~5 and ~20 DPA (Singh *et al.*, 2009). These different modes of growth may partially explain relative growth rate deceleration and the functional roles of protein expression groups correlated with growth rate control (Figure 3C). A prior publication on the apoplast proteome (Lee *et al.*, 2025) and the time-series analyses here point to the existence of heterogeneously sized membrane-enclosed compartments in the extracellular space (Figure 3B). The merged expression group profiles M18 and M14 reflect time points of transition to and from tissue-based morphogenesis (Figure 3A). We speculate that these vesicle-enclosed compartments cannot simply diffuse from the plasma membrane to the extracellular space but are generated when cell-to-cell contact structures that include cytoplasmic components experience nanoscale mechanical failure as bundled fibers grow at different rates or directions.

Integration of proteomic and RNA-seq data can reveal distinct modes of gene expression control. These data also define novel temporal relationships between transcriptional upregulation, protein detection, and associated phenotypes (e.g., increases in cellulose content at 17 DPA). We identified a large set of transcriptionally activated and repressed genes at different

stages of development (Figure 4B and 4C). These lists are likely reliable as they identified specific *G. hirsutum* loci encoding orthologs of known players (Chen *et al.*, 2024; He *et al.*, 2024; Wen *et al.*, 2022) as well as a broader set of unknown genes that have been annotated to have CW-related functions (Figure 4D). These gene sets provide a potential toolkit of promoters and genes that may be effectively used to modulate fiber traits. Other potentially useful gene sets with constitutive expression or discordant transcript and protein profiles likely reflect unknown post-translational control mechanisms and will require further research to develop methods to alter their activities. The validation of these findings is an important next step, but it is not addressed in this resource article.

Conclusions and future challenges

Fiber development is an experimentally tractable system that has enormous potential in the context of architectural engineering for enhanced fiber traits. This potentially is likely to be increased by future improvements to proteomic pipelines. Only ~10% of the expressed genes were detected here (Figure 4A), and there are too many missing values in the daily data matrix. In this respect, improved instrumentation and organelle separation methods would facilitate new insights into the overall proteomic composition and transformation states during cotton fiber development. Finite element biomechanical models of growing trichoblasts predict key CW material properties and CW geometry/force-sensing mechanisms underlie morphogenesis (Li *et al.*, 2023; Yanagisawa *et al.*, 2022). Discovery of key controlling genes can be accelerated by finding better ways to integrate time-series mRNA, proteome, and glycome data with genetic maps and gene function predictions. The barriers to cotton fiber transformation are diminished, so it is now possible to validate these predictions at a reasonable scale (Kangben et al., 2024). These improvements increase the prospects for crop enhancement via cotton fiber engineering. We anticipate growing proteomic datasets and additional high-dimensional transcriptomic and metabolomic data (Qin *et al.*, 2022; Sun *et al.*, 2025; Tuttle *et al.*, 2015) will be fully aggregated in public databases (Yang et al., 2023b; Yu et al., 2021) to fuel machine learning and AI-driven approaches that broadly predict relationships among genetic maps, molecular features, cellular functions, and important fiber traits.

Methods

Cotton growth conditions

Gossypium hirsutum cv. TM1 plants were cultivated in a Conviron® E15 growth chamber (Conviron, Pembina, ND, USA) at the Plant Growth Center in Purdue University as previously described (Lee *et al.*, 2025). Seeds were sown in a 3-gallon pot with a soil mixture composition prepared as 4:2:2:1 soil:perlite:bark:chicken grit. The growth chamber was controlled to retain 50-70 % of relative humidity and remain with a day/night setting of 28/23 °C and 16/8 h at a light intensity of 500 μmol m^{-2} s^{-1} (fluorescent lamps of 28 Sylvania F72T12/CW/VHO 100W and 4 Sylvania F24T12/CW/HO 35W; incandescent bulbs of GE 60W light 130V A19). The 30 mins two-step ramp-up and 30 mins two-step ramp-down periods (15 minutes at 166 μmol photons + 15 minutes at 336 μmol photons) were programmed at the daytime beginning and the ending, respectively. At anthesis, flowers were marked as 0 day post anthesis (0 DPA). Three bolls were collected daily from 5 to 24 DPA for a total of 60 samples (20 days × 3 replicates). The harvested cotton bolls were immediately maintained on ice and carefully dissected to get four intact ovules from each boll.

Cell fractionation into apoplast (APO_T), microsome (P200), and cytosolic (S200) fractions

The APO_T, P200, and S200 fractions were prepared as previously described (Lee *et al.*, 2025). One intact ovule (700 mg) in each locule was dipped in 5 mL of pre-chilled microsome isolation buffer (MIB) [50 mM Hepes/KOH (pH 7.5), 250 mM sorbitol, 50 mM KOAc, 2 mM $Mg(OAc)_2$, 1 mM EDTA, 1 mM EGTA, 1 mM dithiothreitol (DTT), 2 mM PMSF and 1% (v/v) protein inhibitor cocktail (160 mg/mL benzamidine-HCl, 100 mg/mL leupeptin, 12 mg/mL phenanthroline, 0.1 mg/mL aprotinin, and 0.1 mg/mL pepstatin A)]. After 10 minutes of incubation under gentle shaking, the resulting buffer solution was recovered through 2 layers of cheesecloth. Then, the entire APO_T fraction was used for protein precipitation using an acetone precipitation method.

For P200 and S200 fractions, fibers were directly isolated from the ovule material used for the above APO_T collection (Lee *et al.*, 2025). Purified fresh fiber tissues were homogenized under the cold MIB (sample to MIB ratio was 1:4) using Polytron homogenizer (Brinkmann Instruments, New York, NY, USA) with chilled blade tip with 10 seconds grinding, 1 minute rest on ice, and another 10 seconds grinding. The homogenate was filtered with 4 layers of

cheesecloth pre-soaked in cold MIB, and the cheesecloth was further squeezed to get a residual sample in the cheesecloth. Then the filtered homogenate was further centrifuged on an Allegra X-30R centrifuge (Beckman Coulter Life Sciences, Indianapolis, IN, USA) at 1,000 x g for 10 min to remove debris. The supernatant was fractionated at 200 k x g for 20 minutes at 4°C using a Beckman Optima Ultracentrifuge with TLA110 rotor (Beckman Coulter Life Sciences, Indianapolis, IN, USA) into a cytosolic (S200) fraction and pellet. The resulting pellet was washed with cold MIB, incubated on ice for 10 minutes, and ultracentrifuged at the same condition as shown above, and the washing step was repeated. The pellet was defined as a microsome fraction (P200) and directly subjected to protein precipitation using an acetone precipitation method. Protein yield information for the three cell fractions is provided in Data S1.

Determination of protein concentration

For the APO_T and S200 cellular fractions, dried pellets in the acetone precipitation were dissolved and denatured in 100 μL of 8 M Urea for 1 hour at room temperature. For the P200 fraction, 200 μL of 8 Urea was directly added over the final pellet and incubated for 1 hour at room temperature to denature proteins from membranes. Undissolved debris was removed by centrifugation at 12 k x g for 15 min using an Allegra X-30R centrifuge (Beckman Coulter Life Sciences, Indianapolis, IN, USA). Protein assay was performed using a BCA assay kit, according to the manufacturer's protocol (Thermo Fisher Scientific, Waltham, MA, USA).

Electron and fluorescence microscopy of isolated extracellular vesicles (APO_{p200})

A freshly prepared APO_{p200} pellet isolated from a 9 DPA apoplastic wash solution was isolated as described above and diluted 100 times in phosphate buffered saline (pH 7). One microliter of the resuspended vesicles representing APO_{p200} was negatively stained with 1% uranyl acetate on plasma-treated copper grids. Imaging was performed using an accelerating voltage of 40 to 100kV on a Phillips CM-100 Transmission Electron Microscope (Philips/FEI Corporation, Eindhoven, Holland).

Glycome profiling

APO_T was fractionated into APO_{S200} and APO_{P200} fractions by ultracentrifugation as described previously (Lee *et al.*, 2025). The supernatant (APO_{S200} fraction) after

ultracentrifugation was lyophilized completely. The pellet containing apoplastic vesicles (APO_{P200} fraction) were lysed by sonication and ultracentrifuged to collect the soluble fraction with the released polysaccharides (Okekeogbu et al., 2019). The lyophilized APO_{S200} and APO_{P200} fractions were redissolved in water, the sugar contents were measured, and 2 µg of polysaccharides were spotted in each well of a 96-well plate and probed with 71 different antibodies as described previously (Swaminathan *et al.*, 2024). The crude microsomes obtained from lysed fiber cells (P200) were processed identically and used as an additional control.

LC-MS/MS sample preparation

For LC-MSMS analysis, proteins were digested using trypsin as described previously (Lee *et al.*, 2025; McBride et al., 2019). We used all protein samples we gained from the APO_T, and S200 preparations, while 50 µg of proteins in the P200 cell fractions. Denatured protein samples were reduced in 10 mM DTT for 45 min at 60°C and then alkylated with 20 mM iodoacetamide for 45 min at room temperature under the dark. The urea concentration of the solution was brought down to below 1.5 M by adding 50 mM NH_4HCO_3 before trypsin T6567 (Sigma-Aldrich, St. Louis, MO, USA)) digestion. The digestion was proceeded at an enzyme-to-protein ratio of 1 to 50 at 37 °C. After overnight digestion, trifluoroacetic acid (TFA) was added to end digestion. The digested peptides were purified using C18 Micro Spin Columns 74-4601 (Harvard Apparatus, Holliston, MA, USA). One µg of each peptide sample was injected onto the LC-MS/MS system.

LC-MS/MS data acquisition

LC-MS/MS analysis was carried out as described previously (Barabas et al., 2019). Peptide samples were analyzed by a reverse-phase LC-ESI-MS/MS system using a Dionex UltiMate 3000 RSLCnano System coupled with the Orbitrap Fusion Lumos Tribrid Mass Spectrometer (Thermo Fisher Scientific, Waltham, MA, USA). Peptides were loaded onto a trap column (300 µm ID × 5 mm) packed with 5 µm 100 Å PepMap C18 medium and then separated on a reverse phase column (75 µm ID × 50 cm) packed with 2 µm 100-Å PepMap C18 medium (Thermo Fisher Scientific, Waltham, MA, USA). Peptides were resolved at a flow rate of 200 nL/min over a 160-min LC gradient. All the data were acquired in the positive ion mode in the Orbitrap mass analyzer using a higher energy collisional dissociation (HCD) fragmentation

scheme. The MS scan range was from 375 to 1,500 m/z at a resolution of 120,000. The automatic gain control target was set as standard, maximum injection time (50 ms), dynamic exclusion 60 s, and intensity threshold 5.0e3. MS data were acquired in data-dependent mode with a cycle time of 3 s/scan.

Peptide Identification and Quantification

Andromeda search engine on MaxQuant (version 1.6.14.0) was used for relative protein abundance quantification and protein identification (Tyanova et al., 2016). The search was conducted as described (Lee et al., 2021). To minimize locus identity (Locus ID) assignment errors between cell fractions LC/MS raw files were searched in a single MaxQuant job, and signal intensities were normalized (Cox et al., 2014; Tyanova *et al.*, 2016) across all time points (Data S2). The three different parameter groups for the three cellular fractions were defined to specify different parameters for each cell fraction. The search parameters were as follows: Match between runs was set with the default setting; one percent of the Protein and Peptide false discovery rate (FDR) was chosen; 1 was chosen for Min. unique peptides; Label-free quantification was selected within each parameter group; all other parameters were set as default. The species-specific and homoeolog-diagnostic cotton reference was built (Grover *et al.*, 2025) using *G. raimondii* genome annotation to facilitate transcriptome (Grover *et al.*, 2025) and proteome data integration. The reference (Supplemental File S2) includes species/homoeolog-specific SNP and had been validated as performing well in *G. hirsutum* (Hu et al., 2021). To promote consistent locus ID assignments between annotated *Gossypium hirsutum* genome annotations, a locus ID conversion table is provided. Identified peptides, proteins, and a locus ID conversion table can be found in Data S2.

Data filtering for protein identification in each cellular fraction (APO_T, P200, S200)

Two layers of filtering strategies were applied for each cellular fraction dataset to identify proteins in each cellular fraction as described before (Lee *et al.*, 2025). First, those proteins identified with MS/MS count(s) ≥ 1 in at least 2 out of 3 replicates were chosen as identified proteins in each cellular fraction. Secondly, quantified proteins with valid LFQ values in at least 2 out of 3 replicates in each cellular fraction were selected from the above count-based datasets.

These reproducible proteins defined by MS/MS count(s) and LFQ criteria are provided in Data S3A and S3B, respectively.

Profile interpolation

Up to two missing values, which are denoted as zeros, at the two ends or in the middle of a profile were interpolated. In the middle of a profile, a single zero was interpolated by averaging its neighboring non-zero values. In cases where two consecutive zeros occur within an abundance profile, a linear interpolation method was applied. At the two ends of a profile, a single zero was replaced by its neighboring non-zero value, while two consecutive zeros were directly replaced by the nearest non-zero value. The interpolated profiles were used in downstream analyses and are available in Data S3C.

Self-organizing maps (SOM) for protein expression patterns

Self-organizing maps (SOM), a robust unsupervised machine learning method (Kohonen, 1990), were used for analyzing protein expression patterns. We applied a SOM algorithm (Wehrens and Kruisselbrink, 2018), based on an inner product distance metric, to cluster the interpolated protein profiles of each cell fraction. Choosing the number of groups in the SOM analysis involves a trade-off. A large number of groups can lead to empty clusters or clusters with only a single member, while too few groups may fail to capture the underlying clustering structure in the data. We selected a balanced number of groups that preserves meaningful cluster structure while avoiding empty or single-member clusters. The SOM clusters defined the corresponding protein expression profile patterns (Figure 2A). The SOM clustering results for the protein profiles in each group are available in Data S4 and Supplemental Figures 2–4. Cluster profiles were flagged as unreliable if they reflected the overall distribution pattern of protein coverage (Data S3D) or were dominated by patterns that can be explained by the interpolation, e.g., a gap of three consecutive zeros. After this inspection, 12 groups in APO_T, 23 groups in P200, and 9 groups in S200 fractions were selected as containing informative protein profiles (Data S4).

Quality control for reliable protein expression patterns

To ensure reliable protein expression patterns, a multi-step quality control procedure was applied to each SOM cluster to remove outliers, unreliable profiles, and low abundance profiles (Supplemental Figures 2–4).

Step 1: Outlier detection

Outliers were identified by calculating all pairwise similarities defined by the inner product of two profiles in a cluster:

$$\frac{x_i^T x_j}{\sqrt{x_i^T x_i}\sqrt{x_j^T x_j}}$$

If the similarity between a protein profile within a group and the group centroid was less than either the group's median similarity or 0.8, the protein was treated as an outlier and removed (Supplemental Figures 2–4, Step 1).

Step 2: Removal of unreliable profiles

Unreliable profiles were defined as those with isolated non-zeros, including profiles with a single non-zero value flanked by zeros or three consecutive zeros in the middle of the expression pattern. These profiles were removed from the clusters (Supplemental Figures 2–4, Step 2).

Step 3: Filtering based on abundance

To further validate the clustering and filtering steps, we conducted a detailed analysis to assess the relationship between protein expression classes–including non-reproducible proteins, outliers (Step 1), unreliable profiles (Step 2), and reliable profiles–and signal intensity (Supplemental Figure 5). Proteins with below median abundance values across DPAs were more likely to exhibit missing values and contain more zeros in their profiles. When the median abundance dropped below 20 (log2-transformed) for APO_T, 19 for P200, and 20 for S200, these proteins were considered low abundance and removed. These cutoffs were derived from the median abundance values of proteins excluded during the earlier steps (Supplemental Figures 2–4, Step 3).

Informative cell fraction protein expression group patterns

To generate representative expression patterns for each protein groups, member intensities were normalized to a 0–1 range based on their own maximum values within the group

(Figure 2B). The normalized profiles were then averaged to produce a consistent scale across all patterns. This procedure resulted in 44 distinct cell fraction protein expression group patterns, including 12 in APO_T, 23 in P200, and 9 in S200, as detailed in Data S5. In this dataset, a 10 % increase or decrease in a normalized (0–1) profile corresponds approximately to a 4-fold changed (equivalent to a 2-fold change in the log2 scale), because the reliable profiles used in these protein expression groups have averaged protein abundances of at least 20 in the log2-transformed values.

Merging protein expression patterns for data integration.

Similar protein expression patterns were merged into modules through additional SOM clustering analysis, using the 44 previously generated protein expression group patterns. We tested different SOM methods with 9, 20, 30, and 40 groups to optimize group number in the SOM analysis (Data 7A; Supplemental Figures 6–9). As the number of groups increased, the SOM produced more empty clusters or clusters containing only a single group pattern, typically with flat profile patterns, while cluster memberships remained unchanged. Therefore, we selected the 20-cluster SOM method, which defined 15 distinct clusters as merged protein expression modules (Figure 3A). The profiles of these merged modules, along with the proteins in each module, were used as protein phenotypes in data integration analyses with phenotypic and transcriptomic data, which are available in Data S7B and S7C.

Correlation analysis and classification of mRNA and protein pair

In this study, we used a published cotton fiber transcriptome dataset, which profiled transcripts present in developing fiber tissues obtained in the same time window (Grover *et al.*, 2025). We downloaded Supplementary Table 1 which includes transcript profiles of 74,776 genes and further filtered out low abundance transcripts with less than one transcript per kilobase million (TPM) to generate high-quality, reliable transcriptome information. Data S11 provides the filtered 50,512 protein-coding genes used in this comparison analysis.

MaxQuant assigns proteins into a single protein group if they share all their peptides with other proteins (Tyanova *et al.*, 2016), so ambiguously identified homoeologous or orthologous proteins within a protein group can limit the ability to link proteins to their corresponding mRNAs. To address this, we used unique peptide information to assign representative proteins to

each protein group. In this assignment, if proteins were detected by at least a quarter of the maximum number of unique peptides found in the leading protein of a protein group, those proteins were considered identified for that group. Isoform identifications were then condensed to representative gene model IDs used in the transcriptome dataset.

Based on the homoeolog/ortholog assignment results, protein and mRNA pairs were classified into three cases for Pearson correlation coefficient (PCC) analyses. Case 1 includes homoeologs that were identified unambiguously, where unique peptides map to only one homoeolog. Case 2 involves cases in which unique peptides map to both homoeologs; in this case, we retained the homoeolog with the higher PCC. Case 3 includes homoeologs and orthologs identified by unique peptides that map to multiple homoeologs and more than one ortholog within a protein group. Using this protein assignment (annotated in Data S4), we integrated the protein profiles (in protein expression groups shown in Figure 2B) with their mRNA profiles and calculated PCCs between mRNA-protein pairs in each of the cell fractions (Supplemental Figures 10–12; Data S12).

RNA-protein relationship patterns via SOM

To identify relationships between mRNAs and proteins that exhibit constitutively expressed profile patterns, P200 proteins in Case 1 with flat profiles were paired with their corresponding transcripts. We applied SOM again to cluster these pairs of protein and mRNA profiles. In this analysis, 100 groups (10 rows and 10 columns) were used (Supplemental Figure 13). The SOM clusters defined the relationship patterns between mRNAs and proteins over cotton development. Details of the clustering results and the profiles of paired mRNAs and proteins are provided in Data S13.

Statistic tests and data analysis

Statistical analysis was performed using R version 4.2.0 (R Core Team, 2018) on RStudio 2022.07.1 (RStudio Team, 2018). The Flexible Self-Organizing Maps in Kohonen 3.0 package for R (Wehrens and Kruisselbrink, 2018). Microsoft Excel on Office 365 for Mac was used to organize and display the analyzed data.

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository (Perez-Riverol et al., 2021) with the dataset identifier PXD051704.

Code availability

The source codes for the protein SOM clustering analyses and the weighted cross-correlation SOM clustering analysis are publicly available on GitHub (https://github.com/yangpengchengstat/Proteomics-based-models-of-cotton-fiber-protein-dynamics-and-gene-expression-control).

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1951819 to D.B.S. We would like to thank Dr. Uma Aryal at the Purdue Proteomics Facility for running the LC/MS samples. We are grateful to Eileen L. Mallery for growing cotton plants and to Dr. Pragya Barua for helping with sample preparation.

Contributions

Y.L.: Methodology, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization

P.Y.: Methodology, Data Curation, Software, Formal analysis, Writing - Review & Editing

H.R.: Investigation

C.E.G.: Investigation, Writing - Review & Editing

S.S.: Investigation, Writing - Review & Editing

G.M.: Data Curation

J.X.: Methodology, Formal analysis, Writing - Review & Editing, Supervision, Funding acquisition

O.A.Z.: Writing - Review & Editing, Supervision, Funding acquisition

J.F.W.: Writing - Review & Editing, Supervision, Funding acquisition

D.B.S.: Conceptualization, Methodology, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

Competing interests

The authors declare no competing interests.

Figures

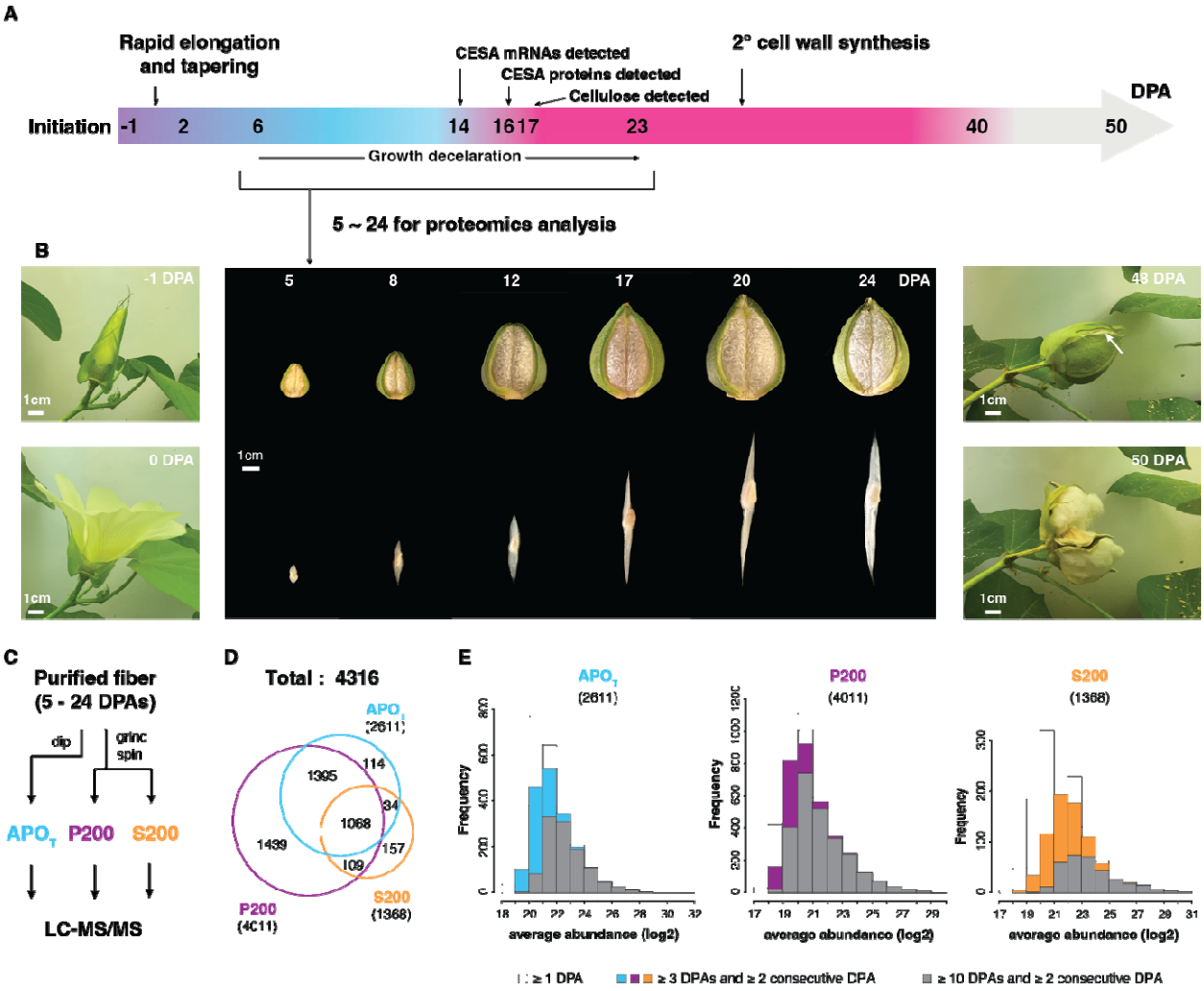


Figure 1. Deep proteome profiling across fiber development in cotton (*G. hirsutum*) plants: cell fractionation and quantitative proteomics.

(A) Key developmental stages of cotton fibers adapted and re-illustrated based on published descriptions (Grover *et al.*, 2025; Swaminathan *et al.*, 2024; Wilson *et al.*, 2024). Flower opening day is defined as 0 days post anthesis (DPA).

(B) Phenotypes of flowers, dissected bolls, and combed fibers at selected time points for proteomics analysis (5 to 24 DPAs). The white arrow indicates the opening of the boll at 48 DPA. One sepal was removed from the flower for visualization. A movie of cotton flower and boll development is available in Supplemental Video 1.

(C) Cell fractionation of triplicate samples for quantitative proteomics analysis (Lee *et al.*, 2025).

(D) Proteome coverages across the three different subcellular compartments.

(E) Distributions of average protein abundances across DPAs showing missing values.

APO_T: apoplast; P200: crude microsome; and S200: soluble fractions. Parentheses show numbers of proteins in each cellular proteome.

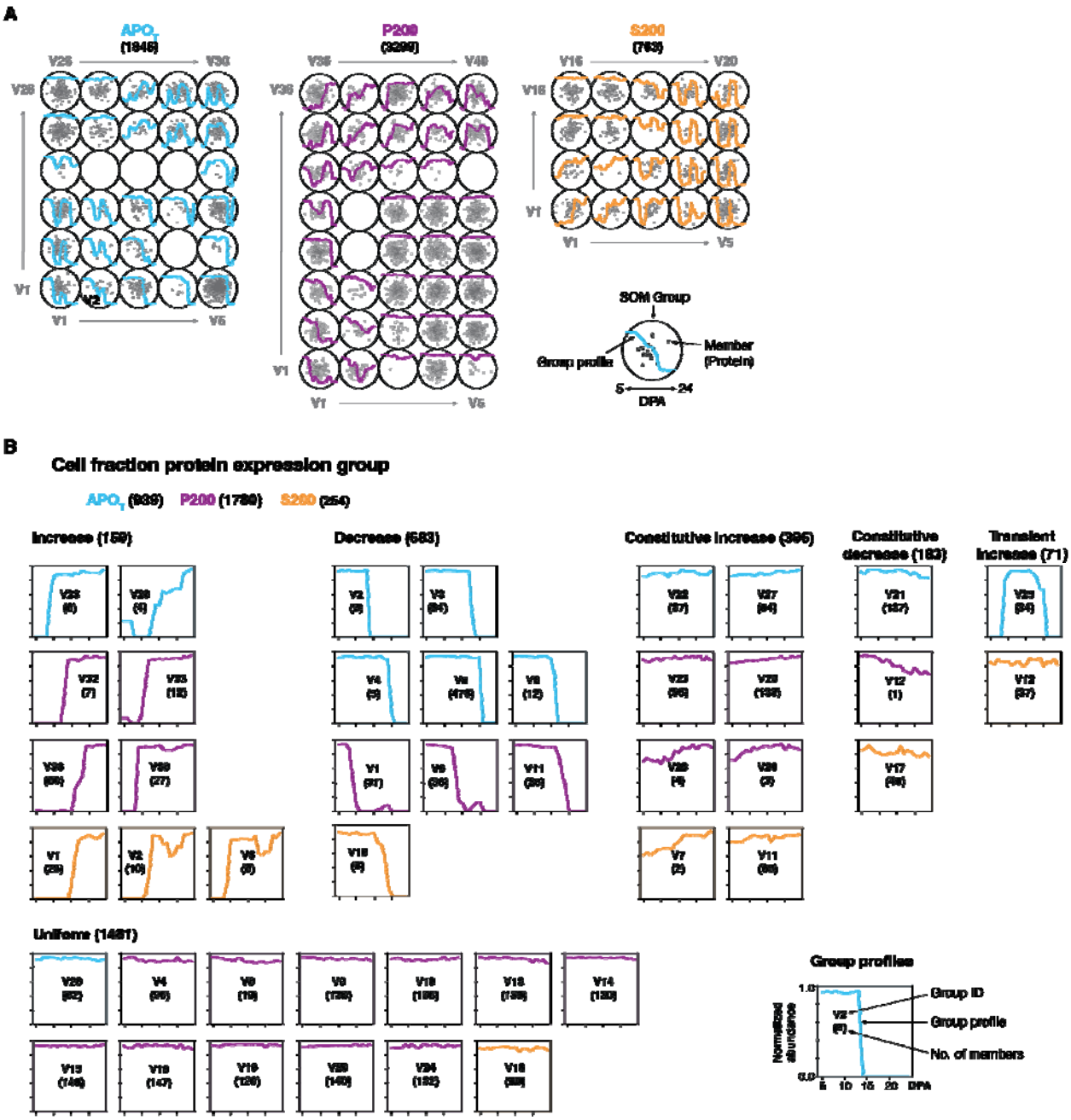


Figure 2. Generating a core set of protein expression modules during developing cotton fibers.

(A) Grouping of similar protein profiles in the three different subcellular proteomes using a machine learning-based self-organizing map (SOM) approach. The number of SOM groups was selected based on group density and distribution to minimize empty groups or multiple groups with similar patterns.

(B) A set of informative protein expression groups across fiber development in the three different cellular fractions.

APO_T: apoplast; P200: crude microsome; and S200: soluble fractions. Parentheses show numbers of proteins in each cellular proteome or each group.

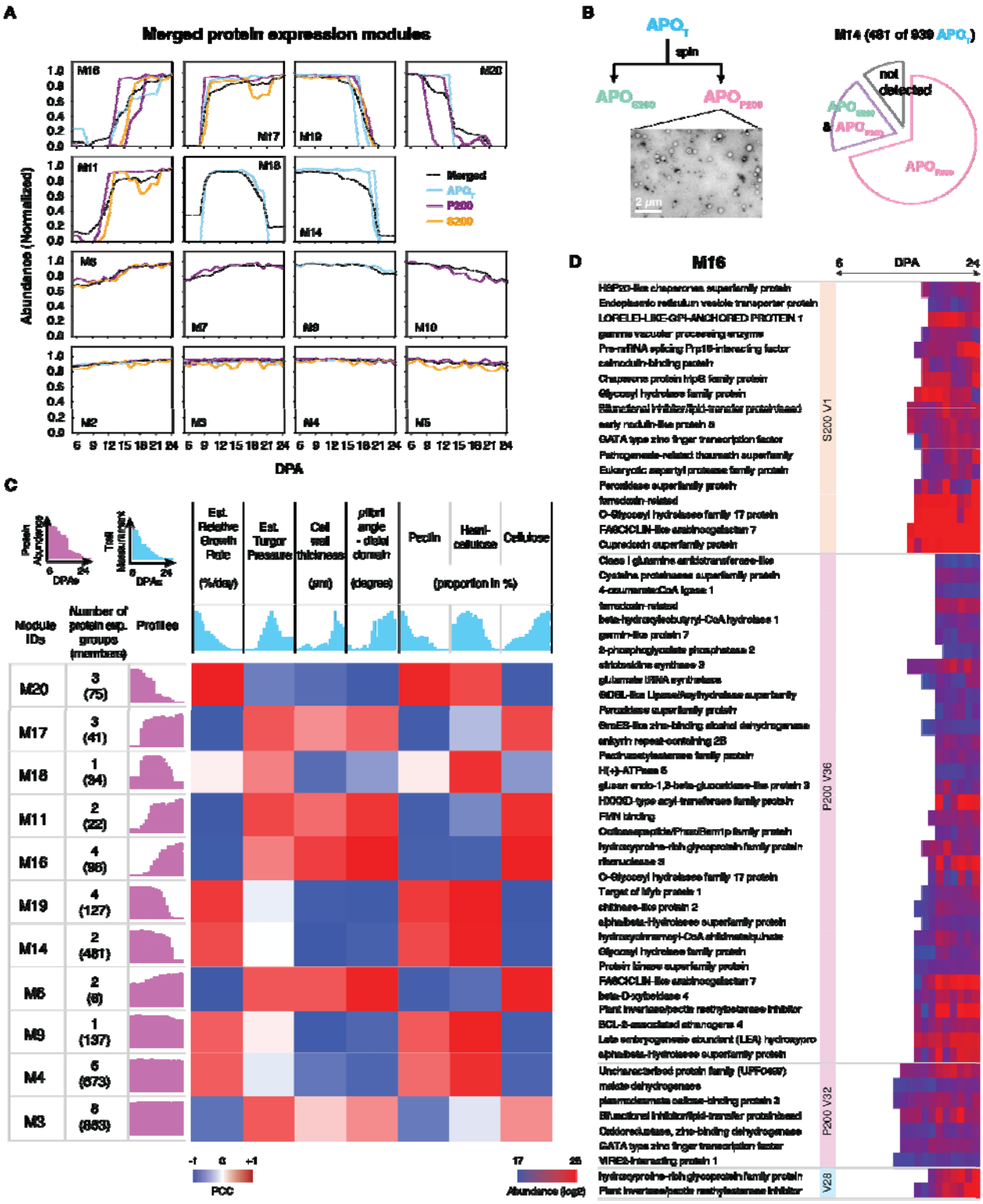


Figure 3. Integration of protein dynamics modules with fiber phenotypes.

(A) Protein expression modules merged from the APO_T, P200, and S200 cellular fraction protein expression groups. The number of SOM groups was selected based on group

density and distribution to minimize empty groups or multiple groups with similar patterns. Empty SOM groups were not visualized.

(B) M14 protein expression group reflects developmentally regulated extracellular vesicles. The proteomics pipeline used to identify APO_{P200} and APO_{S200} proteins (Lee *et al.*, 2025). TEM images of vesicles present in APO_{P200}. Overlap of APO_T proteins in M14 with APO_{P200} proteins (Data S7D).

(C) Integration analyses of fiber phenotypes and protein network modules. Pearson correlation coefficient (PCC) between protein patterns and phenotypes (Swaminathan *et al.*, 2024; Wilson *et al.*, 2024) was calculated. Module profiles show the merged protein expression profiles.

(D) M16 protein expression module shows similar and unique compartment specific profiles that are associated with cellulose content (proportional) increase across development. A full list of proteins in the M16 can be found in Data S9A.

APO_T: apoplast; P200: crude microsome; and S200: soluble fractions. Parentheses show numbers of proteins in each cellular proteome or each group.

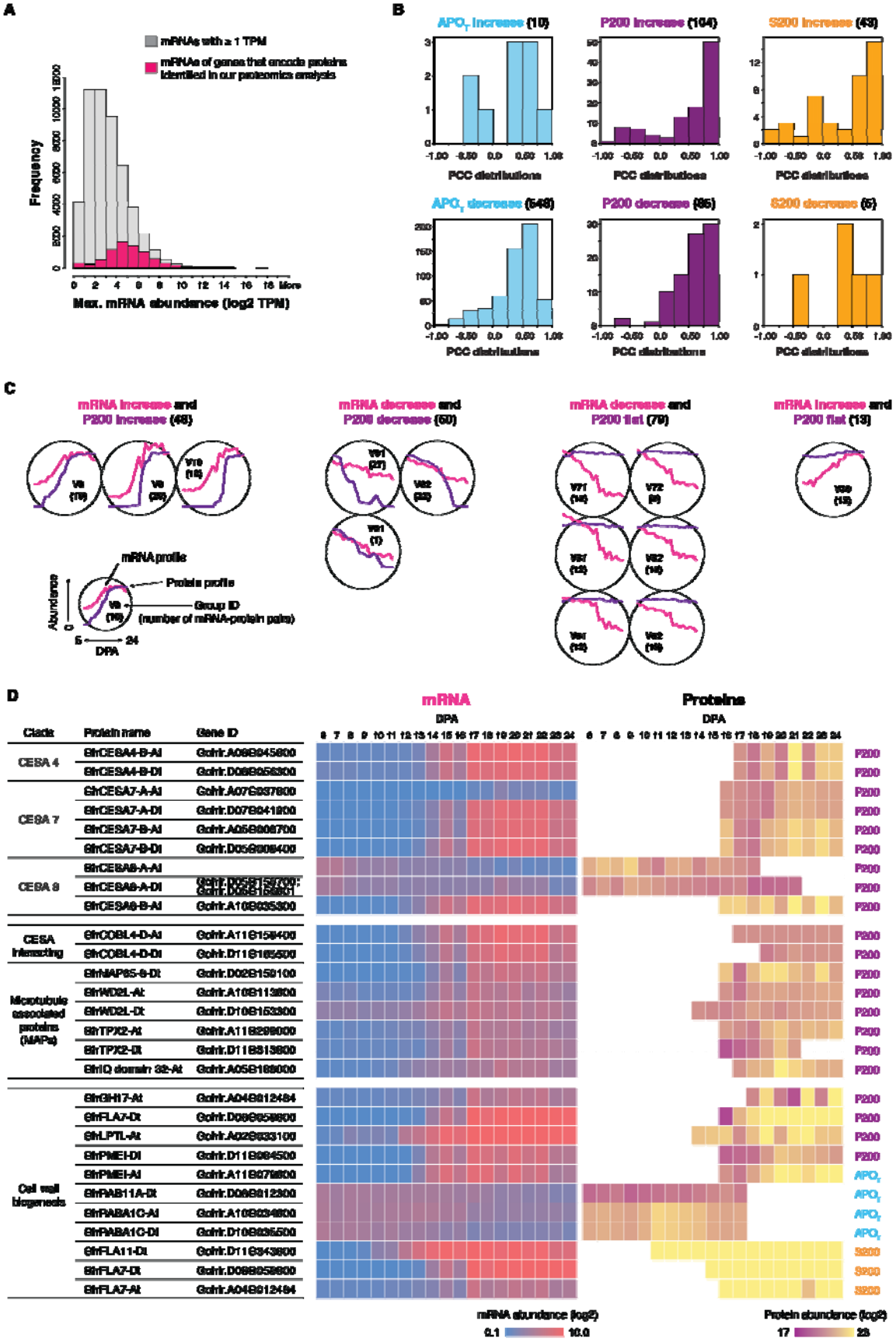

A
mRNAs with ≥ 1 TPM
mRNAs of genes that encode proteins identified in our proteomics analysis
Frequency
Max. mRNA abundance (log2 TPM)
B
APO$_T$ increase (10)
P200 increase (104)
S200 increase (43)
APO$_T$ decrease (548)
P200 decrease (85)
S200 decrease (5)
PCC distributions
C
mRNA increase and P200 increase (48)
mRNA decrease and P200 decrease (60)
mRNA decrease and P200 flat (79)
mRNA increase and P200 flat (13)
mRNA profile
Protein profile
Group ID
(number of mRNA-protein pairs)
Abundance
DPA
D
mRNA
Proteins
DPA
Clade
Protein name
Gene ID
CESA 4
CESA 7
CESA 8
CESA interacting
Microtubule associated proteins (MAPs)
Cell wall biogenesis
mRNA abundance (log2)
Protein abundance (log2)

Figure 4. Global analysis transcript and protein profiles across fiber development.

(A) Coverages of proteins present in the cell fraction protein expression groups over the cotton fiber transcriptome data (Grover *et al.*, 2025). Maximum mRNA and protein abundances across fiber development were visualized. The filtered transcriptome dataset is available in Data S11.

(B) Pearson correlation coefficient (PCC) between mRNAs and proteins (Case 1 proteins; please see the methods) present in the protein expression groups with up or down profiles in Figure 2B.

(C) SOM-based weighted cross-correlation of paired mRNA and proteins that have flat profiles across development. Protein patterns were defined in Figure 2B, and mRNA patterns were defined using fold change (8-times abundance change). The entire SOM result can be found in Data S13 and Supplemental Figure 13.

(D) Transcriptionally controlled known protein markers that function in secondary cell wall synthesis.

Example mRNA and protein abundance profiles of transcriptionally regulated genes present both in the M16 module and assigned to the V8, V9, V10, and V92 groups of the weighted cross-correlation SOM in Figure 4C. A broader list of the predicted cell-wall transcriptionally regulated dataset of the abundance profiles of proteins across fiber development and PCC values are provided in Data S14. At and Dt indicate homoeologs for the A and D genomes.

APO_T: apoplast; P200: crude microsome; and S200: soluble fractions.

Supplemental information

Supplemental Figure 1. Enzyme coverages of metabolic pathways.

Supplemental Figure 2. SOM grouping and profile filtering (APO_T).

Supplemental Figure 3. SOM grouping and profile filtering (P200).

Supplemental Figure 4. SOM grouping and profile filtering (S200).

Supplemental Figure 5. Analysis of proteins removed at each data filtering step while identifying the subset of reliable members.

Supplemental Figure 6. Merged SOM protein expression modules (9 cluster model).

Supplemental Figure 7. Merged SOM protein expression modules (20 cluster model).

Supplemental Figure 8. Merged SOM protein expression modules (30 cluster model).

Supplemental Figure 9. Merged SOM protein expression modules (40 cluster model).

Supplemental Figure 10. Profiles of paired mRNA and proteins (APO_T).

Supplemental Figure 11. Profiles of paired mRNA and proteins (P200).

Supplemental Figure 12. Profiles of paired mRNA and proteins (S200).

Supplemental Figure 13. Weighted cross-correlation SOM of paired mRNA and proteins (P200).

Supplemental Video 1. A movie of the cotton flower and boll development (-5 to 53 DPAs).

Supplemental File 1. The mRNA search databases (AD1_AtDt.transcripts.fa).

Supplemental File 2. The protein search databases (AD1_AtDt.pep.fasta).

Data S1. Fiber protein yields in the APO_T, P200, and S200 fractions.

- The protein amount yielded in each DPA from three different cellular fractions was measured by BCA assay.

Data S2. Raw abundances of peptides and proteins (APO_T, P200, and S200).

- Raw peptide and protein abundance information obtained from a MaxQuant search were reported.
- Locus ID converter from Gorai to Gohir ID system

Data S3. Raw and interpolated protein profiles across fiber development (APO_T, P200, and S200).

- The averaged abundances of proteins identified by MS/MS count(s) > 0 in at least 2 out of the 3 replicates and quantified by LFQ > 0 in at least 2 out of the 3 replicates
- **Interpolated protein profiles**
- Numbers of proteins at each filtering step, i.e. raw (unfiltered), MS/MS, and LFQ.

Data S4. SOM clustering results and filtering for generating protein expression group profiles (APO_T, P200, and S200).

- SOM group profiles in each cell fraction
- **Profiles of members in each SOM group and data filtering were summarized.**

Data S5. SOM protein expression group profiles (APO_T, P200, and S200).

- **The SOM groups with their profiles in the three cell fractions were reported and visualized in Figure 2B.**

Data S6. Gene ontology enrichment analysis of P200 and S200 protein expression groups with increasing patterns.

Data S7. Merged SOM protein expression modules.

- Optimization of the number of groups defined for SOM-based clustering
- **Informative merged SOM groups with their profiles and profiles of members in each merged SOM group were reported.** The reliable profiles reported in Data S4 were pooled, so localization and cell fraction group information can be found in column O.
- M14 APO_T analysis

Data S8. Integration analyses of fiber phenotypes and merged protein expression modules.

- Phenotype estimation data (profiles of each phenotype from 6 to 24 DPA)
- PCC and p-values of correlation result between phenotypes and merged modules. This was visualized in Figure 3C.

Data S9. Gene ontology enrichment analysis of the M16 protein expression module.

- This was visualized in Figure 3D.

Data S10. Glycome profiling of polysaccharide epitopes (APO_{p200} vs. APO_{s200} vs. P200)
.

Data S11. Abundance profiles of transcripts in cotton fiber across development (Grover *et al.*, 2025).

- Filtered mRNA dataset downloaded from Grover et al, 2025 paper.

Data S12. Analyses of Pearson correlation coefficient (PCC) between mRNAs and proteins present in cell fraction protein expression groups with trend-up or -down patterns.

Data S13. Weighted cross-correlation SOM of paired mRNA and proteins (P200).

- SOM group profiles with fold changes in their expression profiles. This was used for data visualization in Figure 4C.
- mRNA and protein profiles were assigned in each SOM group

Data S14. Integrative analysis of mRNAs and proteins annotated as CW-related gene families.

- The list of Cell Wall Related Gene Families was downloaded at https://cellwall.genomics.purdue.edu/
- Data integration and correlation analysis results of CW-related gene families with their mRNA and protein profiles were reported.

References

Bao, Y., Hu, G., Flagel, L.E., Salmon, A., Bezanilla, M., Paterson, A.H., Wang, Z., and Wendel, J.F. (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 **108**:21152-21157.

Barabas, A.J., Aryal, U.K., and Gaskill, B.N. (2019). Proteome characterization of used nesting material and potential protein sources from group housed male mice, Mus musculus. Scientific Reports **9**:17524. 10.1038/s41598-019-53903-x.

Baskin, T.I. (2005). Anisotropic expansion of the plant cell wall. Annu. Rev. Cell Dev. Biol. **21**:203-222. 10.1146/annurev.cellbio.20.082503.103053.

Butterworth, K.M., Adams, D.C., Horner, H.T., and Wendel, J.F. (2009). Initiation and early development of fiber in wild and cultivated cotton. Int. J. Plant Sci. **170**:561-574.

Chen, B., Tian, Z., Fu, G., Zhang, A., Sun, Y., Wang, J., Pan, Z., Li, H., Hu, D., Xia, Y., et al. (2024). GhWDL3 is involved in the formation and development of fiber cell morphology in upland cotton (Gossypium hirsutum L.). Journal of Cotton Research **7**:6. 10.1186/s42397-024-00167-0.

Coen, E., and Cosgrove, D.J. (2023). The mechanics of plant morphogenesis. Science **379**:eade8055. 10.1126/science.ade8055.

Cox, J., Hein, M.Y., Luber, C.A., Paron, I., Nagaraj, N., and Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol Cell Proteomics **13**:2513-2526. 10.1074/mcp.M113.031591.

Cui, H.C., Kong, D.Y., Wei, P.C., Hao, Y.L., Torii, K.U., Lee, J.S., and Li, J. (2014). SPINDLY, ERECTA, and Its Ligand STOMAGEN Have a Role in Redox-Mediated Cortex Proliferation in the Root. Mol Plant **7**:1727-1739. 10.1093/mp/ssu106.

Graham, B.P., and Haigler, C.H. (2021). Microtubules exert early, partial, and variable control of cotton fiber diameter. Planta **253**:47. 10.1007/s00425-020-03557-1.

Grover, C.E., Jareczek, J.J., Swaminathan, S., Lee, Y., Howell, A.H., Rani, H., Arick, M.A., Leach, A.G., Miller, E.R., Yang, P., et al. (2025). A high-resolution model of gene expression during Gossypium hirsutum (cotton) fiber development. BMC Genomics **26**:221. 10.1186/s12864-025-11360-z.

Hamid, Z., Zimmerman, K.D., Guillen-Ahlers, H., Li, C., Nathanielsz, P., Cox, L.A., and Olivier, M. (2022). Assessment of label-free quantification and missing value imputation for proteomics in non-human primates. BMC Genomics **23**:496. 10.1186/s12864-022-08723-1.

He, Q., Yu, Y., Qin, Z., Duan, Y., Liu, H., Li, W., Song, X., Zhu, G., Shang, X., and Guo, W. (2024). COBRA-LIKE 9 modulates cotton cell wall development via regulating cellulose deposition. Plant Physiol **197**10.1093/plphys/kiae675.

Hoffmann, N., King, S., Samuels, A.L., and McFarlane, H.E. (2021). Subcellular coordination of plant cell wall synthesis. Dev Cell **56**:933-948. 10.1016/j.devcel.2021.03.004.

Hu, G., Koh, J., Yoo, M.J., Chen, S., and Wendel, J.F. (2015). Gene-Expression Novelty in Allopolyploid Cotton: A Proteomic Perspective. Genetics 10.1534/genetics.115.174367.

Hu, G., Koh, J., Yoo, M.J., Grupp, K., Chen, S., and Wendel, J.F. (2013). Proteomic profiling of developing cotton fibers from wild and domesticated *Gossypium barbadense*. New Phytol **200**:570-582. 10.1111/nph.12381.

Hu, G., Grover, C.E., Arick, M.A., II, Liu, M., Peterson, D.G., and Wendel, J.F. (2021). Homoeologous gene expression and co-expression network analyses and evolutionary inference in allopolyploids. Briefings in Bioinformatics **22**:1819-1835. 10.1093/bib/bbaa035.

Huang, G., Huang, J.Q., Chen, X.Y., and Zhu, Y.X. (2021). Recent Advances and Future Perspectives in Cotton Research. Annu Rev Plant Biol **72**:437-462. 10.1146/annurev-arplant-080720-113241.

Jan, M., Liu, Z., Guo, C., and Sun, X. (2022). Molecular Regulation of Cotton Fiber Development: A Review. International Journal of Molecular Sciences.

Jiang, J., Shi, Z., Ma, F., and Liu, K. (2022). Identification of key proteins related to high-quality fiber in Upland cotton via proteomics analysis. Plant Cell Rep **41**:893-904. 10.1007/s00299-021-02825-y.

Jörnsten, R., Wang, H.-Y., Welsh, W.J., and Ouyang, M. (2005). DNA microarray data imputation and significance analysis of differential expression. Bioinformatics **21**:4155-4161. 10.1093/bioinformatics/bti638.

Kangben, F., Kumar, S., Li, Z.G., Sreedasyam, A., Dardick, C., Jones, D., and Saski, C.A. (2024). Phylogenetic and functional analysis of tiller angle control homeologs in allotetraploid cotton. Frontiers in Plant Science **14**ARTN 1320638 10.3389/fpls.2023.1320638.

Kim, H.J., and Triplett, B.A. (2001). Cotton fiber growth in planta and in vitro. Models for plant cell elongation and cell wall biogenesis. Plant Physiol **127**:1361-1366.

Kim, H.J., Thyssen, G.N., Song, X., Delhom, C.D., and Liu, Y. (2019). Functional divergence of cellulose synthase orthologs in between wild Gossypium raimondii and domesticated G. arboreum diploid cotton species. Cellulose **26**:9483-9501. 10.1007/s10570-019-02744-y.

Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE **78**:1464-1480. 10.1109/5.58325.

Kumar, M., Campbell, L., and Turner, S. (2016). Secondary cell walls: biosynthesis and manipulation. Journal of Experimental Botany **67**:515-531. 10.1093/jxb/erv533.

Lee, Y., and Szymanski, D.B. (2021). Multimerization variants as potential drivers of neofunctionalization. Sci Adv **7**10.1126/sciadv.abf0984.

Lee, Y., Okita, T.W., and Szymanski, D.B. (2021). A co-fractionation mass spectrometry-based prediction of protein complex assemblies in the developing rice aleurone-subaleurone. The Plant Cell **33**:2965-2980. 10.1093/plcell/koab182.

Lee, Y., Rani, H., Mallery, E.L., and Szymanski, D.B. (2025). A cell fractionation and quantitative proteomics pipeline to enable functional analyses of cotton fiber development. The Plant Journal **121**:e17246. https://doi.org/10.1111/tpj.17246.

Li, J., Szymanski, D.B., and Kim, T. (2023). Probing stress-regulated ordering of the plant cortical microtubule array via a computational approach. BMC Plant Biol **23**:308. 10.1186/s12870-023-04252-5.

Liu, Z., Fan, L., Shu, S., Qanmber, G., Chen, E., Huang, J., Li, F., and Yang, Z. (2025). Cotton metabolism regulatory network: Unraveling key genes and pathways in fiber development and growth regulation. Plant Communications **6**10.1016/j.xplc.2024.101221.

Manzoni, C., Kia, D.A., Vandrovcova, J., Hardy, J., Wood, N.W., Lewis, P.A., and Ferrari, R. (2018). Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief Bioinform **19**:286-302. 10.1093/bib/bbw114.

McBride, Z., Chen, D., Reick, C., Xie, J., and Szymanski, D.B. (2017). Global analysis of membrane-associated protein oligomerization using protein correlation profiling. Molecular & Cellular Proteomics:mcp.000276.002017. 10.1074/mcp.RA117.000276.

McBride, Z., Chen, D., Lee, Y., Aryal, U.K., Xie, J., and Szymanski, D.B. (2019). A label-free mass spectrometry method to predict endogenous protein complex composition. Molecular & Cellular Proteomics **18**:1588. 10.1074/mcp.RA119.001400.

Meinert, M.C., and Delmer, D.P. (1977). Changes in biochemical composition of the cell wall of the cotton fiber during development. Plant Physiol **59**:1088-1097.

Mujahid, H., Pendarvis, K., Reddy, J.S., Nallamilli, B.R., Reddy, K.R., Nanduri, B., and Peng, Z. (2016). Comparative Proteomic Analysis of Cotton Fiber Development and Protein Extraction Method Comparison in Late Stage Fibers. Proteomes **4**10.3390/proteomes4010007.

Nie, L., Wu, G., and Zhang, W. (2006). Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in Desulfovibrio vulgaris: a quantitative analysis. Genetics **174**:2229-2243. 10.1534/genetics.106.065862.

Okekeogbu, I.O., Pattathil, S., González Fernández-Niño, S.M., Aryal, U.K., Penning, B.W., Lao, J., Heazlewood, J.L., Hahn, M.G., McCann, M.C., and Carpita, N.C. (2019). Glycome and Proteome Components of Golgi Membranes Are Common between Two Angiosperms with Distinct Cell-Wall Structures. The Plant Cell **31**:1094-1112. 10.1105/tpc.18.00755.

Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., Kundu, Deepti J., Prakash, A., Frericks-Zipper, A., Eisenacher, M., et al. (2021). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Research **50**:D543-D552. 10.1093/nar/gkab1038.

Pettolino, F.A., Yulia, D., Bacic, A., and Llewellyn, D.J. (2022). Polysaccharide composition during cotton seed fibre development: temporal differences between species and in different seasons. Journal of Cotton Research **5**ARTN 27 10.1186/s42397-022-00136-5.

Qin, Y., Wei, H.L., Sun, H.R., Hao, P.B., Wang, H.T., Su, J.J., and Yu, S.X. (2017). Proteomic Analysis of Differences in Fiber Development between Wild and Cultivated Gossypium hirsutum L. Journal of Proteome Research **16**:2811-2824. 10.1021/acs.jproteome.7b00122.

Qin, Y., Sun, M., Li, W., Xu, M., Shao, L., Liu, Y., Zhao, G., Liu, Z., Xu, Z., You, J., et al. (2022). Single-cell RNA-seq reveals fate determination control of an individual fibre cell initiation in cotton (Gossypium hirsutum). Plant Biotechnology Journal **20**:2372-2388. https://doi.org/10.1111/pbi.13918.

Qin, Y.M., Hu, C.Y., Pang, Y., Kastaniotis, A.J., Hiltunen, J.K., and Zhu, Y.X. (2007). Saturated very-long-chain fatty acids promote cotton fiber and Arabidopsis cell elongation by activating ethylene biosynthesis. Plant Cell **19**:3692-3704. 10.1105/tpc.107.054437.

R Core Team (2018). R: A language and environment for statistical computing. Version 3.5.1. http://www.R-project.org/.

Rajangam, A.S., Kumar, M., Aspeborg, H., Guerriero, G., Arvestad, L., Pansri, P., Brown, C.J., Hober, S., Blomqvist, K., Divne, C., et al. (2008). MAP20, a microtubule-associated protein in the secondary cell walls of hybrid aspen, is a target of the cellulose synthesis inhibitor 2,6-dichlorobenzonitrile. Plant Physiol. **148**:1283-1294. pp.108.121913 [pii] 10.1104/pp.108.121913 [doi].

RStudio Team (2018). RStudio: Integrated Development Environment for R. Version 1.1.463. https://rstudio.com/.

Ruan, Y., Llewellyn, D., and Furbank, R. (2001). The control of single-celled cotton fiber elongation by developmentally reversible gating of plasmodesmata and coordinated expression of sucrose and k(+) transporters and expansin. Plant Cell **13**:47-60.

Ryser, U. (1977). Cell-Wall Growth in Elongating Cotton Fibers - Autoradiographic Study. Cytobiologie **15**:78-84.

Schubert, A.M., Benedict, C.R., Berlin, J.D., and Kohel, R.J. (1973). Cotton Fiber Development - Kinetics of Cell Elongation and Secondary Wall Thickening. Crop Science **13**:704-709. DOI 10.2135/cropsci1973.0011183X001300060035x.

Seagull, R.W. (1992). A Quantitative Electron-Microscopic Study of Changes in Microtubule Arrays and Wall Microfibril Orientation during Invitro Cotton Fiber Development. Journal of Cell Science **101**:561-577.

Seagull, R.W. (1993). Cytoskeletal Involvement in Cotton Fiber Growth and Development. Micron **24**:643-660. Doi 10.1016/0968-4328(93)90042-Y.

Singh, B., Avci, U., Inwood, S.E., Grimson, M.J., Landgraf, J., Mohnen, D., Sorensen, I., Wilkerson, C.G., Willats, W.G., and Haigler, C.H. (2009). A Specialized Outer Layer of the Primary Cell Wall Joins Elongating Cotton Fibers Into Tissue-like Bundles. Plant Physiol.

Stewart, J.M. (1975). Fiber initiation on the cotton ovule. Am J Bot **62**:723-730.

Sugano, S.S., Shimada, T., Imai, Y., Okawa, K., Tamai, A., Mori, M., and Hara-Nishimura, I. (2010). Stomagen positively regulates stomatal density in Arabidopsis. Nature **463**:241-244. 10.1038/nature08682.

Sun, X., Qin, A., Wang, X., Ge, X., Liu, Z., Guo, C., Yu, X., Zhang, X., Lu, Y., Yang, J., et al. (2025). Spatiotemporal transcriptome and metabolome landscapes of cotton fiber during initiation and early development. Nature Communications **16**:858. 10.1038/s41467-025-55869-z.

Swaminathan, S., Grover, C.E., Mugisha, A.S., Sichterman, L.E., Lee, Y., Yang, P., Mallery, E.L., Jareczek, J.J., Leach, A.G., Xie, J., et al. (2024). Daily glycome and transcriptome profiling reveals

polysaccharide structures and correlated glycosyltransferases critical for cotton fiber growth. The Plant Journal **120**:1857-1879. https://doi.org/10.1111/tpj.17084.

Tiwari, S.C., and Wilkins, T.A. (1995). Cotton (*Gossypium hirsutum*) seed trichomes expand via diffuse growing mechanism. Can. J. Bot. **73**:746-757.

Tuttle, J.R., Nah, G., Duke, M.V., Alexander, D.C., Guan, X.Y., Song, Q.X., Chen, Z.J., Scheffler, B.E., and Haigler, C.H. (2015). Metabolomic and transcriptomic insights into how cotton fiber transitions to secondary wall synthesis, represses lignification, and prolongs elongation. Bmc Genomics **16**ARTN 477
10.1186/s12864-015-1708-9.

Tyanova, S., Temu, T., and Cox, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nature Protocols **11**:2301-2319. 10.1038/nprot.2016.136.

Wang, L., Liu, C., Liu, Y., and Luo, M. (2020a). Fumonisin B1-Induced Changes in Cotton Fiber Elongation Revealed by Sphingolipidomics and Proteomics. Biomolecules **10**10.3390/biom10091258.

Wang, S., Li, W., Hu, L., Cheng, J., Yang, H., and Liu, Y. (2020b). NAguideR: performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses. Nucleic Acids Research **48**:e83-e83. 10.1093/nar/gkaa498.

Wehrens, R., and Kruisselbrink, J. (2018). Flexible Self-Organizing Maps in kohonen 3.0. Journal of Statistical Software **87**:1 - 18. 10.18637/jss.v087.i07.

Wen, X., Zhai, Y., Zhang, L., Chen, Y., Zhu, Z., Chen, G., Wang, K., and Zhu, Y. (2022). Molecular studies of cellulose synthase supercomplex from cotton fiber reveal its unique biochemical properties. Sci China Life Sci **65**:1776-1793. 10.1007/s11427-022-2083-9.

Wilkins, O., Brautigam, K., and Campbell, M.M. (2010). Time of day shapes Arabidopsis drought transcriptomes. Plant J **63**:715-727. 10.1111/j.1365-313X.2010.04274.x.

Wilson, M.C., Howell, A.H., Sood, A., Lee, Y., Yang, P., Yu, E., Rani, H., Mallery, E.L., Swaminathan, S., Grover, C.E., et al. (2024). Developmental variability in cotton fiber cell wall properties linked to important agronomic traits. bioRxiv:2024.2008.2019.607249.
10.1101/2024.08.19.607249.

Yanagisawa, M., Keynia, S., Belteton, S., Turner, J.A., and Szymanski, D.B. (2022). A conserved cellular mechanism for cotton fibre diameter and length control. . *in silico* Plants **4**10.1093.

Yanagisawa, M., Desyatova, A.S., Belteton, S., Mallery, E.M., Turner, J.A., and Szymanski, D.B. (2015). Patterning mechanisms of cytoskeletal and cell wall systems during leaf trichome morphogenesis. Nature Plant **1**:15014. doi:10.1038/nplants.2015.14.

Yang, Y., Lai, W., Long, L., Gao, W., Xu, F., Li, P., Zhou, S., Ding, Y., and Hu, H. (2023a). Comparative proteomic analysis identified proteins and the phenylpropanoid biosynthesis pathway involved in the response to ABA treatment in cotton fiber development. Sci Rep **13**:1488.
10.1038/s41598-023-28084-3.

Yang, Z., Wang, J., Huang, Y., Wang, S., Wei, L., Liu, D., Weng, Y., Xiang, J., Zhu, Q., Yang, Z., et al. (2023b). CottonMD: a multi-omics database for cotton biological study. Nucleic Acids Res **51**:D1446-D1456. 10.1093/nar/gkac863.

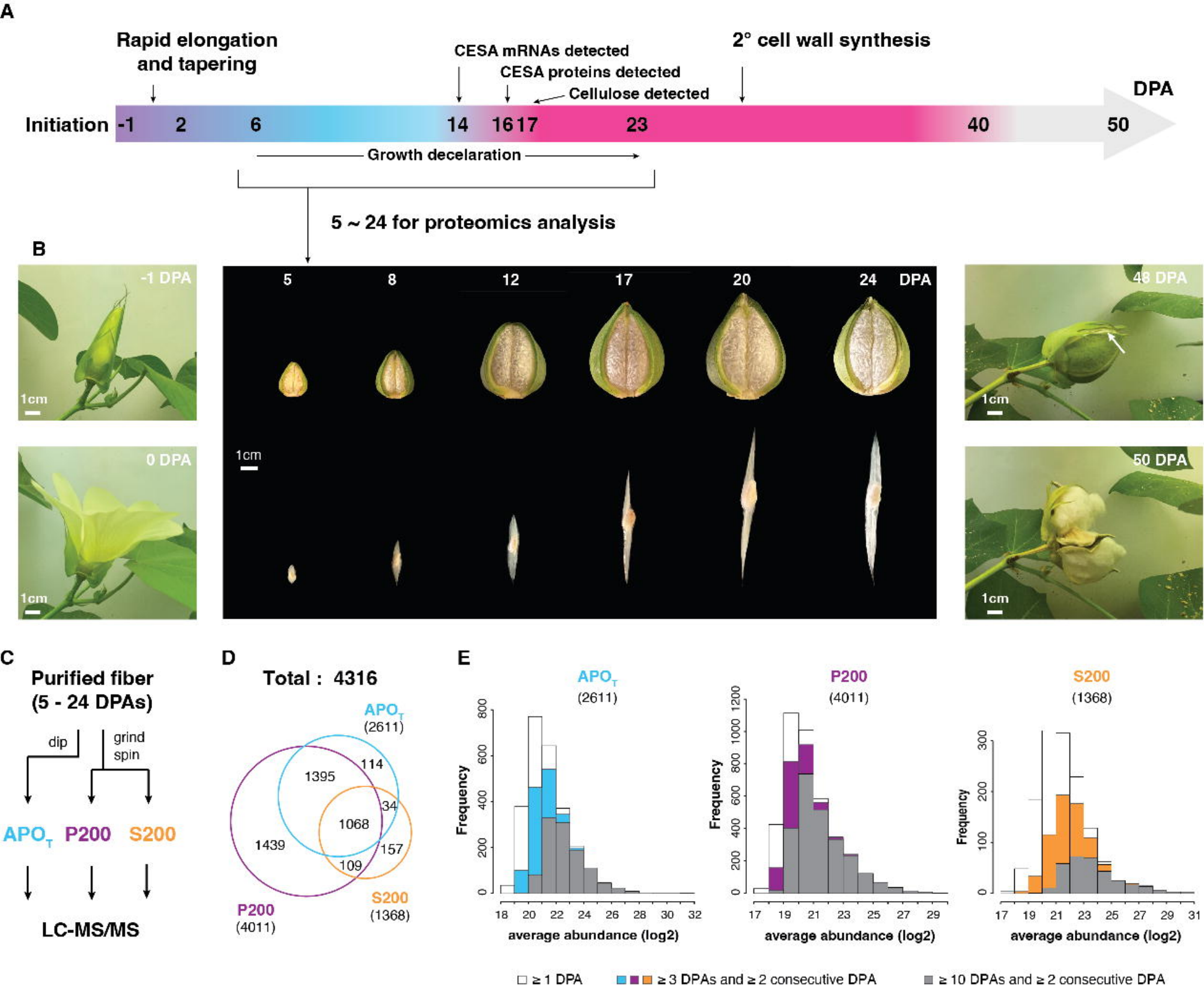
Yu, J., Jung, S., Cheng, C.H., Lee, T., Zheng, P., Buble, K., Crabb, J., Humann, J., Hough, H., Jones, D., et al. (2021). CottonGen: The Community Database for Cotton Genomics, Genetics, and Breeding Research. Plants (Basel) **10**10.3390/plants10122805.

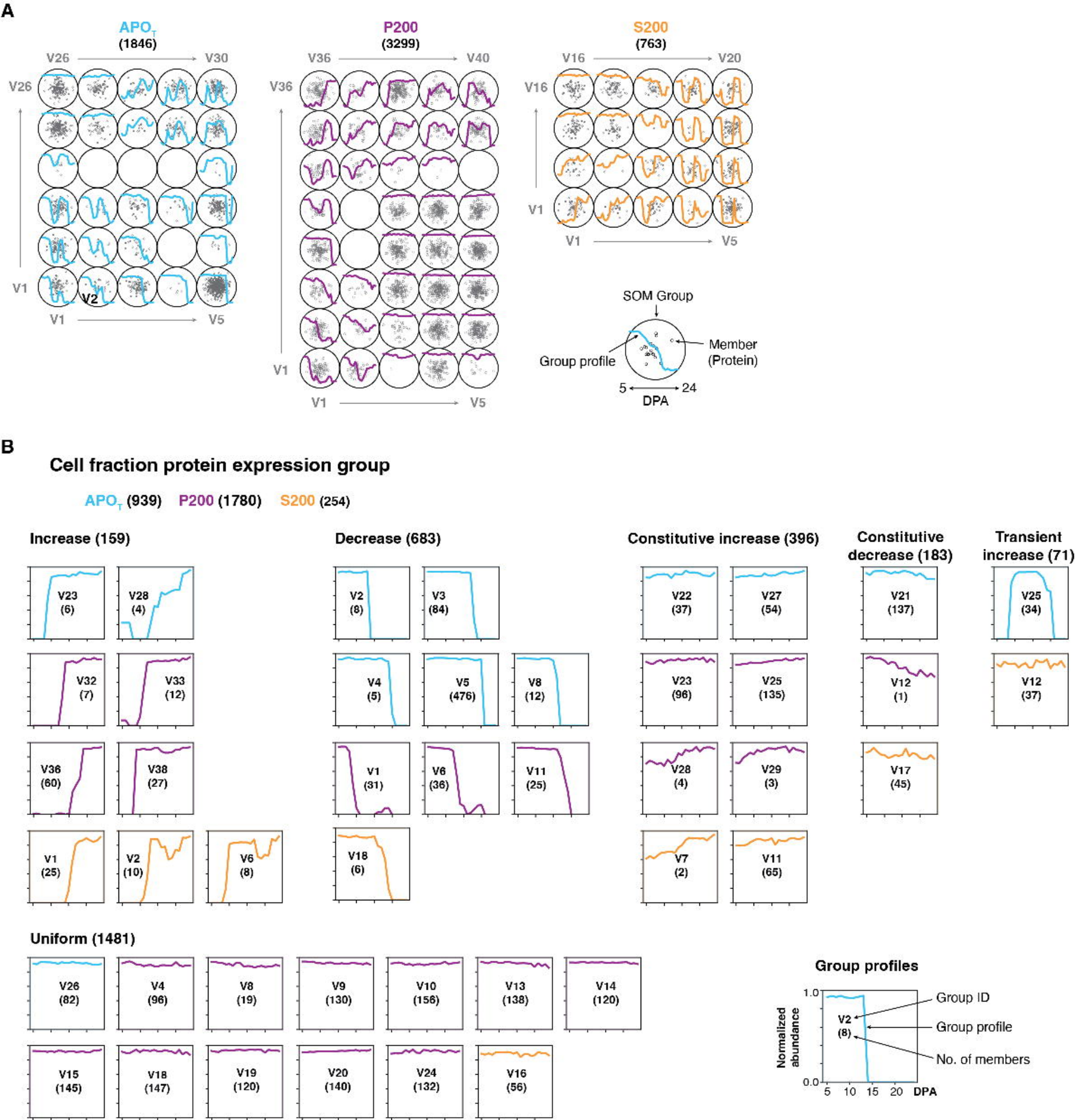
Yu, Y., Wu, S., Nowak, J., Wang, G., Han, L., Feng, Z., Mendrinna, A., Ma, Y., Wang, H., Zhang, X., et al. (2019). Live-cell imaging of the cytoskeleton in elongating cotton fibres. Nat Plants **5**:498-504.
10.1038/s41477-019-0418-8.

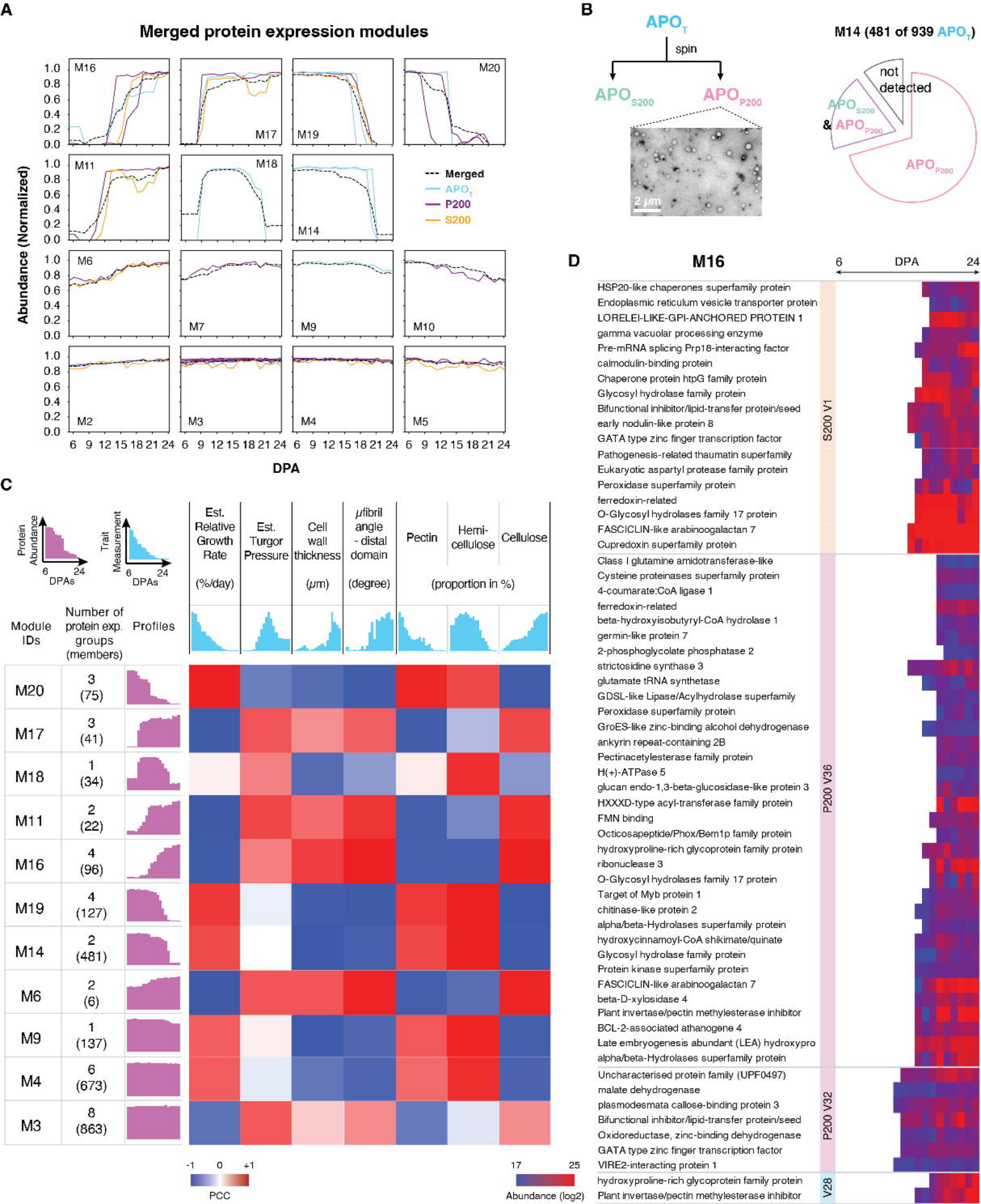
Zang, Y., Hu, Y., Xu, C., Wu, S., Wang, Y., Ning, Z., Han, Z., Si, Z., Shen, W., Zhang, Y., et al. (2021). GhUBX controlling helical growth results in production of stronger cotton fiber. iScience **24**:102930. 10.1016/j.isci.2021.102930.

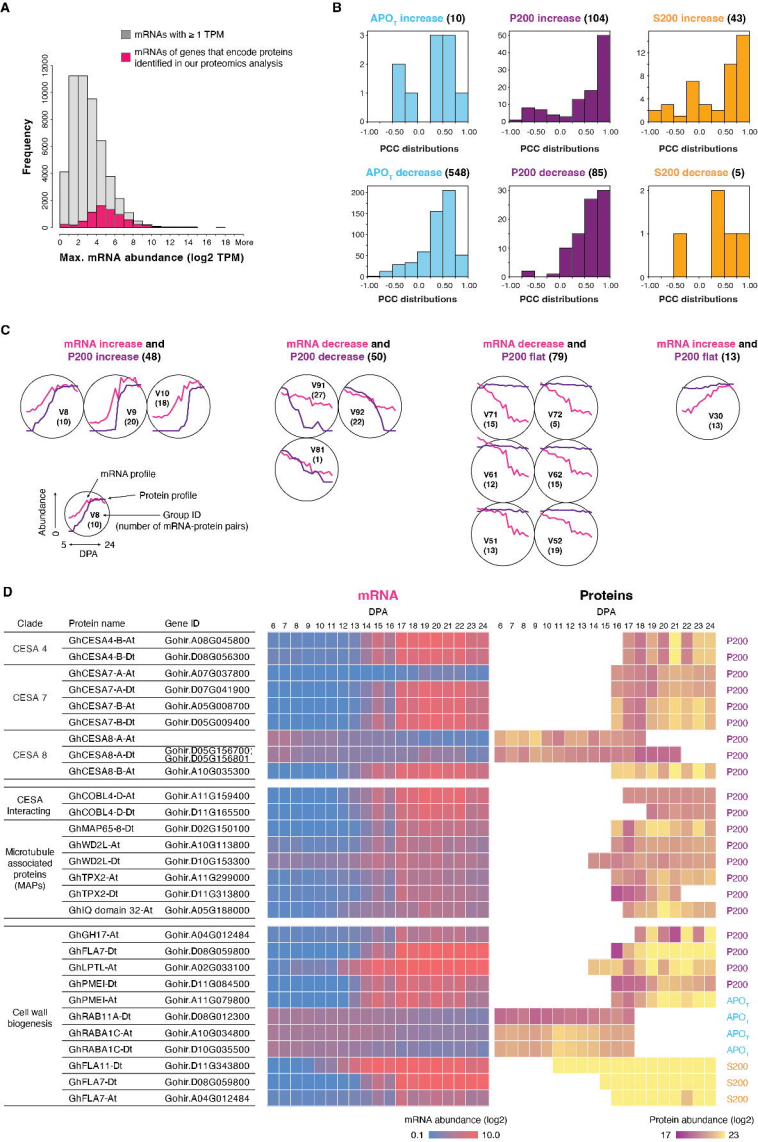
Zhang, Y., Yu, J., Wang, X., Durachko, D.M., Zhang, S., and Cosgrove, D.J. (2021). Molecular insights into the complex mechanics of plant epidermal cell walls. Science **372**:706-711.
10.1126/science.abf2824.

Zhou, X., Hu, W., Li, B., Yang, Y., Zhang, Y., Thow, K., Fan, L., and Qu, Y. (2019). Proteomic profiling of cotton fiber developmental transition from cell elongation to secondary wall deposition. Acta Biochim Biophys Sin (Shanghai) **51**:1168-1177. 10.1093/abbs/gmz111.


A
Rapid elongation and tapering
CESA mRNAs detected
CESA proteins detected
Cellulose detected
2° cell wall synthesis
DPA
Initiation
-1
2
6
14
16
17
23
40
50
Growth decelaration
5 ~ 24 for proteomics analysis
B
-1 DPA
0 DPA
1cm
5
8
12
17
20
24
DPA
48 DPA
50 DPA
C
Purified fiber (5 - 24 DPAs)
dip
grind spin
APO_T
P200
S200
LC-MS/MS
D
Total : 4316
APO_T (2611)
P200 (4011)
S200 (1368)
1395
114
34
1068
1439
157
109
E
Frequency
average abundance (log2)
≥ 1 DPA
≥ 3 DPAs and ≥ 2 consecutive DPA
≥ 10 DPAs and ≥ 2 consecutive DPA


A
APO$_T$
(1846)
P200
(3299)
S200
(763)
V26
V30
V36
V40
V16
V20
V1
V5
V2
SOM Group
Group profile
Member
(Protein)
5
24
DPA
B
Cell fraction protein expression group
APO$_T$ (939) P200 (1780) S200 (254)
Increase (159)
Decrease (683)
Constitutive increase (396)
Constitutive decrease (183)
Transient increase (71)
V23 (6)
V28 (4)
V32 (7)
V33 (12)
V36 (60)
V38 (27)
V1 (25)
V2 (10)
V6 (8)
V2 (8)
V3 (84)
V4 (5)
V5 (476)
V8 (12)
V1 (31)
V6 (36)
V11 (25)
V18 (6)
V22 (37)
V27 (54)
V23 (96)
V25 (135)
V28 (4)
V29 (3)
V7 (2)
V11 (65)
V21 (137)
V12 (1)
V17 (45)
V25 (34)
V12 (37)
Uniform (1481)
V26 (82)
V4 (96)
V8 (19)
V9 (130)
V10 (156)
V13 (138)
V14 (120)
V15 (145)
V18 (147)
V19 (120)
V20 (140)
V24 (132)
V16 (56)
Group profiles
Normalized abundance
1.0
0.0
5
10
15
20
DPA
Group ID
Group profile
No. of members


A
Merged protein expression modules
Abundance (Normalized)
M16
M17
M19
M20
M11
M18
M14
M6
M7
M9
M10
M2
M3
M4
M5
Merged
APO$_T$
P200
S200
DPA
B
APO$_T$
spin
APO$_{S200}$
APO$_{P200}$
2 μm
M14 (481 of 939 APO$_T$)
not detected
APO$_{S200}$ & APO$_{P200}$
APO$_{P200}$
C
Protein Abundance
Trait Measurement
DPAs
Est. Relative Growth Rate (%/day)
Est. Turgor Pressure
Cell wall thickness (μm)
μfibril angle - distal domain (degree)
Pectin
Hemi-cellulose
Cellulose
(proportion in %)
Module IDs
Number of protein exp. groups (members)
Profiles
M20 3 (75)
M17 3 (41)
M18 1 (34)
M11 2 (22)
M16 4 (96)
M19 4 (127)
M14 2 (481)
M6 2 (6)
M9 1 (137)
M4 6 (673)
M3 8 (863)
PCC
Abundance (log2)
D
M16
DPA
S200 V1
HSP20-like chaperones superfamily protein
Endoplasmic reticulum vesicle transporter protein
LORELEI-LIKE-GPI-ANCHORED PROTEIN 1
gamma vacuolar processing enzyme
Pre-mRNA splicing Prp18-interacting factor
calmodulin-binding protein
Chaperone protein htpG family protein
Glycosyl hydrolase family protein
Bifunctional inhibitor/lipid-transfer protein/seed
early nodulin-like protein 8
GATA type zinc finger transcription factor
Pathogenesis-related thaumatin superfamily
Eukaryotic aspartyl protease family protein
Peroxidase superfamily protein
ferredoxin-related
O-Glycosyl hydrolases family 17 protein
FASCICLIN-like arabinoogalactan 7
Cupredoxin superfamily protein
P200 V36
Class I glutamine amidotransferase-like
Cysteine proteinases superfamily protein
4-coumarate:CoA ligase 1
ferredoxin-related
beta-hydroxyisobutyryl-CoA hydrolase 1
germin-like protein 7
2-phosphoglycolate phosphatase 2
strictosidine synthase 3
glutamate tRNA synthetase
GDSL-like Lipase/Acylhydrolase superfamily
Peroxidase superfamily protein
GroES-like zinc-binding alcohol dehydrogenase
ankyrin repeat-containing 2B
Pectinacetylesterase family protein
H(+)-ATPase 5
glucan endo-1,3-beta-glucosidase-like protein 3
HXXXD-type acyl-transferase family protein
FMN binding
Octicosapeptide/Phox/Bem1p family protein
hydroxyproline-rich glycoprotein family protein
ribonuclease 3
O-Glycosyl hydrolases family 17 protein
Target of Myb protein 1
chitinase-like protein 2
alpha/beta-Hydrolases superfamily protein
hydroxycinnamoyl-CoA shikimate/quinate
Glycosyl hydrolase family protein
Protein kinase superfamily protein
FASCICLIN-like arabinoogalactan 7
beta-D-xylosidase 4
Plant invertase/pectin methylesterase inhibitor
BCL-2-associated athanogene 4
Late embryogenesis abundant (LEA) hydroxypro
alpha/beta-Hydrolases superfamily protein
P200 V32
Uncharacterised protein family (UPF0497)
malate dehydrogenase
plasmodesmata callose-binding protein 3
Bifunctional inhibitor/lipid-transfer protein/seed
Oxidoreductase, zinc-binding dehydrogenase
GATA type zinc finger transcription factor
VIRE2-interacting protein 1
V28
hydroxyproline-rich glycoprotein family protein
Plant invertase/pectin methylesterase inhibitor


A
mRNAs with ≥ 1 TPM
mRNAs of genes that encode proteins identified in our proteomics analysis
Frequency
Max. mRNA abundance (log2 TPM)
B
APO1 increase (10)
P200 increase (104)
S200 increase (43)
APO1 decrease (548)
P200 decrease (85)
S200 decrease (5)
PCC distributions
C
mRNA increase and P200 increase (48)
mRNA decrease and P200 decrease (50)
mRNA decrease and P200 flat (79)
mRNA increase and P200 flat (13)
mRNA profile
Protein profile
Group ID (number of mRNA-protein pairs)
Abundance
DPA
D
mRNA
Proteins
DPA
Clade
Protein name
Gene ID
CESA 4
GhCESA4-B-At Gohir.A08G045800
GhCESA4-B-Dt Gohir.D08G056300
CESA 7
GhCESA7-A-At Gohir.A07G037800
GhCESA7-A-Dt Gohir.D07G041900
GhCESA7-B-At Gohir.A05G008700
GhCESA7-B-Dt Gohir.D05G009400
CESA 8
GhCESA8-A-At
GhCESA8-A-Dt Gohir.D05G156700 Gohir.D05G156801
GhCESA8-B-At Gohir.A10G035300
CESA Interacting
GhCOBL4-D-At Gohir.A11G159400
GhCOBL4-D-Dt Gohir.D11G165500
Microtubule associated proteins (MAPs)
GhMAP65-8-Dt Gohir.D02G150100
GhWD2L-At Gohir.A10G113800
GhWD2L-Dt Gohir.D10G153300
GhTPX2-At Gohir.A11G299000
GhTPX2-Dt Gohir.D11G313800
GhIQ domain 32-At Gohir.A05G188000
Cell wall biogenesis
GhGH17-At Gohir.A04G012484
GhFLA7-Dt Gohir.D08G059800
GhLPTL-At Gohir.A02G033100
GhPMEI-Dt Gohir.D11G084500
GhPMEI-At Gohir.A11G079800
GhRAB11A-Dt Gohir.D08G012300
GhRABA1C-At Gohir.A10G034800
GhRABA1C-Dt Gohir.D10G035500
GhFLA11-Dt Gohir.D11G343800
GhFLA7-Dt Gohir.D08G059800
GhFLA7-At Gohir.A04G012484
P200
APO1
S200
mRNA abundance (log2)
Protein abundance (log2)