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

Running title:
Cotton fiber proteome dynamics: daily resolution

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[1,2,*]

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 epidermal trichoblasts emerge from the seed coat 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 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. Integration of proteomic and RNA-sequencing data revealed distinct modes of transcriptional and post-transcriptional control. Collectively these datasets provide large-scale gene function predictions that can serve as a toolkit to engineer cotton material properties.

Main

Cotton (*Gossypium spp*.) is the world's most important fiber crop and the foundation of a multi-billion dollar renewable textile industry [1,2]. Tens of thousands of unicellular seed coat trichoblasts synchronously differentiate and transform their length:width aspect ratios during development to ~3000:1 [3]. At maturity each fiber contains a helically patterned cellulose rich-wall that desiccates after programmed cell death to generate harvestable thin, strong, spinnable fibers [4-6]. Fiber size, spinnability, and strength are the output of a morphogenetic program of anisotropic diffuse growth [7,8] with an often underestimated degree of complexity (Fig. 1). Phenotyping daily reveals an early tapering phase that is completed by 2 DPA, during which the hemispherical fiber initial is converted to a thin elongated trichoblast [8,9]. After tapering the cell wall thickens as it enters a protracted phase of decelerating anisotropic diffuse growth [10,11] while simultaneously modulating plasmodesmata functions and turgor pressure [12]. Given progress in understanding the molecular and biomechanical control of trichoblast morphogenesis [8,13,14], there are emerging opportunities to use biotechnological interventions to engineer fiber traits. The goal of this research is to integrate high-temporal resolution quantitative proteomics, RNAseq, 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 (Supplementary Video 1 and Fig. 1a,b). Fiber development also includes reversible transitions between cell-autonomous growth and a tissue-morphogenesis during which the cells are tightly adhered [15]. Cotton has long served as a useful model for cell wall (CW) assembly and morphogenesis [16-18] and components of the molecular control continue to emerge [2]. A highly conserved cellulose patterning module underlies polarized diffuse growth in the plant kingdom [19], and the details of how interactions among cortical microtubules, patterned cellulose synthase (CESA) activity, and CW stress-sensing operate to control cell shape change are beginning to emerge [20,21]. Strong lateral interactions between microfibrils (Zhang et al., 2021) and interactions between fiber and matrix components generates domains of the CW that deform asymmetrically in response to turgor-generated mechanical stress in the CW [8,13,22]. Spatially patterned cortical microtubules influence microfibril deposition throughout development to maintain cell diameter and pattern a helical secondary CW (SCW) [4,8,11,16]. Important cotton microtubule-associated proteins are being discovered using forward and reverse genetic approaches [23-25]. 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 [2].

Proteomic profiling provides one 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 an altered abundances as a function of domestication [26-28], hormone or inhibitor responses [29,30], and developmental stages [31-33]. In general, the number of discovered proteins is low and the 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 [29,34,35]. 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 simple 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 encompassing primary cell wall synthesis and the transition into secondary wall synthesis. Clustering methods identified unique protein expression groups that reflected unique activities of the apoplast. Multiple groups had clear temporal transitions and broad memberships across the cell fractions. Correlation analysis with phenotypic variability across the same timeline provides a way to predict suites of genes that coordinate the unidirectional progression of fiber development. Integration of RNAseq and proteomic data can provide large scale predictions about general modes of gene expression control [36,37]. We used combinations of cross-correlation analysis and machine-learning methods to globally predict modes of transcriptional and post-transcriptional control. Collectively these data and analytical methods provide new knowledge about evolutionarily conserved protein and conserved cellular processes that control morphology in the plant kingdom. These data resources also provide a new way to link cellular processes with important phenotypes and predict groups of gene products that mediate key developmental transitions.

Results

Protein profiling of developing cotton fiber by proteomics

A central goal was to generate reliable time series data of protein abundance in multiple subcellular fractions from fiber tissue. Methods to isolate highly purified fiber samples and generate high quality cell fractions with limited cross contamination with broken cells were developed [38]. 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 include multiple locules from a single boll to increase coverage of the crude cytosolic (S200) fraction (Fig. 1c) and included three replicates daily from 5-24 DPA. The entire developmental progression from anthesis at 0 DPA to the show stopping event of boll opening at 48 DPA are shown in Fig. 1a,b and Supplementary Video 1. Protein yields in the different cell fractions are included in Supplementary Data 1. Our approach was to conduct quantitative shotgun proteomics of cytosol, microsome, and extracellular proteins obtained from dissected locules and purified fibers at daily intervals from 5 to 24 DPA (Fig. 1c). 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 secondary cell wall synthesis.

A crude apoplast (APO_T) fraction was generated after careful dissection and brief immersion of the intact locules in 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 with cytosol and microsome fractions [38]. After immersion, fibers were manually dissected, homogenized, and subjected to differential centrifugation to generate a crude microsomal (P200) and a soluble (S200) fraction (Fig. 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 coverages to date (Fig. 1d and Supplementary Data 3). 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 [35,39]. 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 (Supplementary Data 3). The P200 fraction consistently had the most diverse protein populations over the

time course, and, as observed previously [38], was more similar to the APO_T compared to S200 (Supplementary Data 3c). The APO_T fraction primarily consists of membrane-enclosed particles with a highly heterogeneous composition [38]. Widespread overlap among the different cell fractions can occur because many proteins are dually localized [40]. Our pipeline generated protein quantification across over 6 orders of magnitude (Fig. 1e). The protein types amenable to proteomic analyses are extremely diverse (Supplementary Data 3). The coverage of metabolism in this dataset is deep (Extended Data Fig. 1) and provides broad experimental evidence for the specific gene products that operate in central metabolism, phenylpropanoid biosynthesis [29,41], fatty acid synthesis [42], and cell wall assembly [43].

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. Proteins that met this detection frequency threshold tended to have slightly higher signal intensities in each cell fraction compared to those that did not meet the threshold (Extended Data Fig. 5). However, the protein population composed of those detected in at least 2 of 3 consecutive DPA and the subset detected in 10 or more DPA and had similar intensity distributions (Fig. 1e and Extended Data Fig. 5), 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 Supplementary Data 3 and Extended Data Figs. 2–4.

Our deep proteome profiling provides a glimpse into compartmentalized protein behaviors during development, identifying diverse groups of proteins that coordinate fiber development. However, the widespread occurrence of missing values is a widely recognized shortcoming of label-free quantitative proteomics. We developed interpolation and data filtering methods to identify informative expression profiles in the three subcellular locales. To identify expression groups, potentially informative profiles were analyzed using a machine learning-based clustering method termed Self-Organizing Map (SOM) analysis [44,45]. Each SOM group has a "code" that reflects a central expression pattern of the members (Fig. 2a and Supplementary Data 4). 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 DPA (Supplementary Data 3d); or 3) the pattern was defined by string(s) 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 44 potentially informative groups were filtered further to remove unreliable members with discordant features in the expression profile (Fig. 2b, Extended Data Figs. 2–5). In the end, these 44 expression groups provide large scale functional data on gene sets and cellular activities (Fig. 2c, Supplementary Data 5). These data 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 (Fig. 2c). 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 (Supplementary Data 6). 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 (Supplementary Data 5). These "constitutive increase/decrease" groups may reflect broad reprogramming of gene expression relative to

the transition to SCW synthesis [46] 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. The cell fraction profiles were merged to generate a protein expression model that included 11 merged groups with distinct non-uniform profiles (Fig. 3a). The mean normalized intensities of the merged group members were used to estimate the behavior of the group (Supplementary Data 8). Aggregation of 7 developmentally variable fiber phenotypes from the same or similar developmental windows [11,47] provided a way to test for significant associations between protein expression groups path to test for associations using correlations analyses.

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 (Fig. 3c, Supplementary Data 8b). M16 contained proteins present in all three subcellular compartments and included many proteins involved in CW biosynthesis, as well as many previously uncharacterized proteins (Fig. 3d and Supplementary Data 8 and 9). M16 contained *G. hirsutum* orthologs of each the SCW-associated CESAs [46,48] that were induced at ~16 DPA to similar levels (Fig. 4d and Supplementary data 7c,). M19 was negatively correlated with SCW transition and may contain proteins that are specifically involved in the fiber elongation and remodeling the primary CW (Fig. 3a,c and Supplementary 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 [38], we find M14 members were either specific to the particulate fraction or dual-localized in the soluble and particulate fractions of the apoplast (Supplementary data 7d). The particulate fraction was enriched in spherical stain-excluding compartments that could be visualized using negative staining and transmission electron microscopy (Fig. 3b). M14 contained more than half of the total APO_T proteome. M9 and M18 were also APO_T-specific expression groups with similar profiles (Fig. 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 (Fig. 3c). M20 may contain groups of proteins associated with a fiber state with faster relative growth rate. M11 and M17 were anti-correlated with relative growth rate and auxin-response factors, ABA receptors, and vesicle trafficking machineries (Supplementary Data 7c, column T, M11) that enable persistent axial growth and/or early components that prime the cell for the transition to tissue-based morphogenesis [15] or SCW synthesis [43]. 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 [46]. 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 (Fig 4a and Extended Data Figs. 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 (Fig. 4b, Supplementary Data 11). About ~ 48 % 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 a reduced abundance after ~20 DPA (Fig. 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, which is usually associated with non-cell-autonomous cell proliferation [49,50], was down regulated at ~14 DPA and correlated with transcript levels (Supplemental Data 7).

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 (Fig. 4c and Supplementary Data 12). The full clustering result is shown graphically in Extended Data Fig. 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 (Fig. 4c), of which 48 of 98 were captured in the PCC analysis described in Fig. 4b. Among the large population of proteins with a more uniform expression, there were 79 genes in which mRNA levels fell at least 8-fold over the time course (Fig. 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 proteins levels were stable (Fig. 3c) indicating that translation of the mRNA may be regulated or the steady state of the protein may be the results of physical interactions with a stabilizing factor. The SOM weighted cross correlation analyses provides 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 temporal transcriptional control with those known to be involved in cell wall assembly or the transition to SCW synthesis (Supplementary Data 13). 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 (Fig. 4d). GhCESA4-B, 7-A/B, and 8-B proteins were first detected at with ~2 day lag compared to when the 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 (Supplementary Data 13). These data 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

This paper leverages 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 [11,38,46,47]. 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 4300 fiber proteins, and reliable abundance profiles over a wide window of development for more than 2900. This dataset provides useful time-series data on the localization and abundance of 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 (Extended Data Fig.1, Supplementary Data 5,6). The roles of these small molecules in fiber morphogenesis merit further analyses. All of the proteomic and associated RNAseq data are publicly available in a simple search interface (link to be provided).

The SOM clustering method and cross-correlation with time-resolved phenotypes was used to generate models of how protein groups in multiple cellular compartments are linked to specific developmental transitions (Fig. 3). The merged profiles capture constitutively expressed proteins with uniform (M2,M3,M4,M5), increasing (M6,M7), and decreasing (M8,M9) 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 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 that are conserved across species and cell types with a thickened SCW [51]. Our dataset contains orthologs of each of the three CESA clades [52] and COBRA-like [53] proteins that are known to influence secondary cell wall synthesis in cotton, and many of them appear to be transcriptionally regulated (Figure 3D and 4D). SCW patterning is a complex process that includes a radical reorganization of the cytoskeletal, vesicle trafficking, and cell wall assembly systems [54]. 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 [15]. 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 (Fig. 3c). A prior publication on the apoplast proteome [38] and the time-series analyses here point to the existence of heterogeneously sized membrane-enclosed compartments in the extracellular space (Fig. 3b). The merged expression group profiles M18 and M14 reflect time-points of transition to and from tissue-based morphogenesis (Fig. 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 RNAseq data can reveal distinct modes of gene expression control. These data define 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 [25,52,53] as well as a broader set of unknown genes that have been annotated to have cell-wall related functions (Fig. 4d). These gene sets provide potential toolkits 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 controls mechanisms and will require further research to develop methods to alter their activities.

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 (Fig. 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 cell wall material properties and cell-wall geometry/force -sensing mechanisms underlie morphogenesis [8,21]. 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 [55]. These improvements increase the prospects for crop enhancement via cotton fiber engineering. We anticipate growing proteomic datasets and additional high-dimensionality transcriptome and metabolomic data [41] will be fully aggregated in public databases [56,57] to fuel machine learning and AI approaches that broadly predict relationships among genetic maps, molecular features, cellular functions, and important fiber traits.

Figures

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

(A) Key developmental stages of cotton fibers [11,46,47]. Flower opening day is defined as 0 days post anthesis (DPA).

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

(C) Cell fractionation of triplicate samples for quantitative proteomics analysis [38].

(D) Proteome coverages for 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.

Fig. 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.

(B) Data filtering to identify the subset of reliable members of an expression group. APO_T V2 was used as an example. Outliers were removed using pair-wise similarity with group representing profiles (Cosine similarity < 0.8). Unreliable profiles, which refer to isolated non-zeros, were moved from the group. Then, lower-abundance proteins were filtered out to create protein group representing profiles.

(C) 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.

Fig. 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.

(B) M14 protein expression group reflects developmentally regulated extracellular vesicles. The proteomics pipeline used to identify APO_{P200} and APO_{S200} proteins [38]. TEM images of vesicles present in APO_{P200}. Overlap of APO_T proteins in M14 with APO_{P200} proteins (Supplementary Data 7d).

(C) Integration analyses of fiber phenotypes and protein network modules. Pearson correlation coefficient (PCC) between protein patterns and phenotypes [11,47] were 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 Supplementary Data 9a.

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

Fig. 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 [46]. Maximum mRNA and protein abundances across fiber development were visualized. The filtered transcriptome dataset is available in Supplementary Data 10.

(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 Fig. 2c.

(C) SOM-based weighted cross correlation of paired mRNA and proteins that have flat profiles across development. Protein patterns were defined in Fig. 2c, and mRNA patterns were defined using fold change (8-times abundance change). The entire SOM result can be found in Supplementary Data 12 and Extended Data Fig. 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 Fig. 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 Supplementary Data 12. At and Dt indicate homoeologs for the A and D genomes.

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

Online Methods

Cotton growth conditions

Gossypium hirsutum cv. TM1 were cultivated in a Conviron® E15 growth chamber (Conviron, Pembina, ND, USA) at the Plant Growth Center in Purdue University as previously described [38]. 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 [38]. 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 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 [38]. Purified fresh fiber tissues were homogenized under the cold MIB (sample to MIB ratio was 1:4) using Polytron homogenizer (Brinkman 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 Supplementary Data 1.

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 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).

LC-MS/MS sample preparation

For LC-MSMS analysis, proteins were digested using trypsin as described previously [38,58]. 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 additional ammonium bicarbonate before trypsin (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 [59]. 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. MS/MS data were collected at a resolution of 7,500.

Peptide Identification and Quantification

Andromeda search engine on MaxQuant (version 1.6.14.0) was used for relative protein abundance quantification and protein identification [60]. The search was conducted as described [61]. 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 [60,62] across all timepoints (Supplementary Data 2). 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: the match

between runs function was set with a maximum matching time window of 0.7 min as default; proteins identified by a single unique peptide were selected; Label-free quantification was selected within each parameter group; all other parameters were set as default. The cotton reference built by [46] to facilitate transcriptome and proteome data integration. 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 Supplementary Data 2.

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 [38]. 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 Supplementary Data 3a and 3b, respectively.

Profile interpolation

Up to 2 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 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 Supplementary Data 3c.

Self-organizing maps (SOM) for protein expression patterns

Self-organizing maps (SOM), a robust unsupervised machine learning method [45], were used for analyzing protein expression patterns. We applied a SOM algorithm [44], based on an inner product distance metric, to cluster the interpolated protein profiles of each cell fraction. The SOM clusters defined the corresponding protein expression profile patterns (Fig. 2a). The SOM clustering results for the protein profiles in each group are available in Supplementary Data 4; Extended Data Fig 2–4. Cluster profiles were flagged as unreliable if they reflected the overall distribution pattern of protein coverage (Supplementary Data 3d) 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 (Supplementary Data 4).

Quality control for reliable protein expression patterns

For each SOM cluster, outliers and unreliable protein profiles were removed. 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 (Step 1, Fig. 2b). Additionally, unreliable profiles, defined as isolated non-zeros, were moved from the group (Step 2). This

includes profiles with either a single non-zero value between two zeros or three consecutive zeros in the middle. Low abundance, unreliable profiles were identified based on signal distributions (Extended Data Fig. 5). When the averaged abundances across DPAs fall below 20 for APO_T, 19 for P200, and 20 for S200, these proteins tend to exhibit more zeros in their profiles. The cutoffs for low abundance proteins were determined using the median abundance values of proteins removed during earlier filtering steps (Step 3). reliable profiles are expected to have an averaged protein abundance of at least 20 (log2-transformed). In this dataset a 10 % increase or decrease in abundance in the log2 scale corresponds to a 4-fold change. To calculate average abundance profiles for protein groups, member intensities were first normalized to a 0–1 range (based on its own maximum value within the group) and then averaged, ensuring all profiles share the same scale. This set of 44 cell fraction protein expression group patterns (12 groups in APO_T, 23 groups in P200, and 9 groups in S200) is provided in Supplementary Data 5.

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 nodes to optimize node number in the SOM analysis (Supplementary Data 7a; Extended Data Figs. 6–9). As the number of nodes increased, the SOM produced more empty clusters or clusters contained only a single group pattern, typically with flat profile patterns. Therefore, we selected the 20-cluster SOM method, which defined 15 distinct clusters as merged protein expression modules (Fig. 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 Supplementary Data 7b,c.

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 [46]. 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 a high-quality, reliable transcriptome information. Supplementary Data 10 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 [60], so ambiguously identified homoeologous or orthologous proteins within a protein group can limit to link proteins and their 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 the 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 Supplementary Data), we integrated the protein profiles (in protein expression groups

shown in Fig. 2c) with their mRNA profiles and calculated PCCs between mRNA-protein pairs in each of the cell fractions (Extended Data Fig. 10–12; Supplementary Data 11).

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 nodes (10 rows and 10 columns) were used (Extended Data Fig. 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 Supplementary Data 12.

Statistic tests and data analysis

Statistical analysis was performed using R version 4.2.0 [63] on RStudio 2022.07.1 [64]. The Flexible Self-Organizing Maps in Kohonen 3.0 package for R [44]. Microsoft Excel on Office 365 for Mac was used to organized and display the analyzed data.

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository [65] 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.

Ethics declarations:

Competing interests

The authors declare no competing interests.

Author information

Authors and Affiliations
Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN, 47909, USA
Youngwoo Lee & Daniel B Szymanski

Center for Plant Biology, Purdue University, West Lafayette, IN, 47909, USA
Youngwoo Lee & Daniel B Szymanski

Department of Statistics, Purdue University, West Lafayette, IN, 47909, USA
Pengcheng Yang, Gideon Miller & Jun Xie

Cereal Crops Research Unit, USDA-ARS, Madison, WI, 53726, USA
Heena Rani

Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA, 50011, USA
Corrinne E Grover & Jonathan F Wendel

Roy J Carver Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA, 50011, USA
Sivakumar Swaminathan & Olga A Zabotina

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

Corresponding author
Correspondence to Daniel B Szymanski.

Extended data

Extended Data Fig. 1. Enzyme coverages of metabolic pathways.

Extended Data Fig. 2. SOM grouping and profile filtering (APO_T).

Extended Data Fig. 3. SOM grouping and profile filtering (P200).

Extended Data Fig. 4. SOM grouping and profile filtering (S200).

Extended Data Fig. 5. Analysis of proteins removed at each data filtering step during identifying the subset of reliable members.

Extended Data Fig. 6. Merged SOM protein expression modules (9 cluster model)

Extended Data Fig. 7. Merged SOM protein expression modules (20 cluster model)

Extended Data Fig. 8. Merged SOM protein expression modules (30 cluster model)

Extended Data Fig. 9. Merged SOM protein expression modules (40 cluster model)

Extended Data Fig. 10. Profiles of paired mRNA and proteins (APO_T)

Extended Data Fig. 11. Profiles of paired mRNA and proteins (P200)

Extended Data Fig. 12. Profiles of paired mRNA and proteins (S200)

Extended Data Fig. 13. Weighted cross correlation SOM of paired mRNA and proteins (P200)

Supplementary information

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

Supplementary Data 1. Fiber protein yields in the APO_T, P200, and S200 fractions.

- Protein amount yielded in each DPA from three different cellular fractions were measure by BCA assay.

Supplementary Data 2. Raw abundances of peptides and proteins (APO_T, P200, and S200).

- Raw peptide and protein abundance information obtained from a MaxQuant search were reported.

Supplementary Data 3. 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**
- Locus ID converter from Gorai to Gohir ID system
- Numbers of proteins at each filtering step, i.e. raw (unfiltered), MS/MS, and LFQ.

Supplementary Data 4. 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.**

Supplementary Data 5. 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 the Fig. 2C.**

Supplementary Data 6. Gene ontology enrichment analysis of P200 and S200 protein expression groups with increase patterns.

Supplementary Data 7. 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 the Supplementary Data 4 were pooled, so localization and cell fraction group information can be found in column O.
- M14 APO_T analysis

Supplementary Data 8. 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 Fig. 3c.

Supplementary Data 9. Gene ontology enrichment analysis of the M16 protein expression module.
- This was visualized in Fig. 3d.

Supplementary Data 10. Abundance profiles of transcripts in cotton fiber across development [46].
- Filtered mRNA dataset downloaded from Grover et al, 2024 paper.

Supplementary Data 11. Analyses of Pearson correlation coefficient (PCC) between mRNAs and proteins present in cell fraction protein expression groups with trend-up or -down patterns.

Supplementary Data 12. 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 Fig. 4c.
- mRNA and protein profiles were assigned in each SOM group

Supplementary Data 13. Integrative analysis of mRNAs and proteins annotated as cell wall 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 cell wall related gene families with their mRNA and protein profiles were reported.

References

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

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

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

4 Seagull, R. W. 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 (1992).

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

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

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

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

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

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

11 Wilson, M. C. *et al.* Developmental variability in cotton fiber cell wall properties linked to important agronomic traits. *bioRxiv*, 2024.2008.2019.607249 (2024). https://doi.org/10.1101/2024.08.19.607249

12 Ruan, Y., Llewellyn, D. & Furbank, R. 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 (2001).

13 Yanagisawa, M. *et al.* Patterning mechanisms of cytoskeletal and cell wall systems during leaf trichome morphogenesis. *Nature Plant* **1**, 15014 (2015). https://doi.org/doi:10.1038/nplants.2015.14

14 Yu, Y. *et al.* Live-cell imaging of the cytoskeleton in elongating cotton fibres. *Nat Plants* **5**, 498-504 (2019). https://doi.org/10.1038/s41477-019-0418-8

15 Singh, B. *et al.* A Specialized Outer Layer of the Primary Cell Wall Joins Elongating Cotton Fibers Into Tissue-like Bundles. *Plant Physiol* (2009).

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

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

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

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

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

21 Li, J., Szymanski, D. B. & Kim, T. Probing stress-regulated ordering of the plant cortical microtubule array via a computational approach. *BMC Plant Biol* **23**, 308 (2023). https://doi.org/10.1186/s12870-023-04252-5
22 Zhang, Y. *et al.* Molecular insights into the complex mechanics of plant epidermal cell walls. *Science* **372**, 706-711 (2021). https://doi.org/10.1126/science.abf2824
23 Rajangam, A. S. *et al.* 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 (2008). https://doi.org/pp.108.121913 [pii] 10.1104/pp.108.121913 [doi]
24 Zang, Y. *et al.* GhUBX controlling helical growth results in production of stronger cotton fiber. *iScience* **24**, 102930 (2021). https://doi.org/10.1016/j.isci.2021.102930
25 Chen, B. *et al.* GhWDL3 is involved in the formation and development of fiber cell morphology in upland cotton (Gossypium hirsutum L.). *Journal of Cotton Research* **7**, 6 (2024). https://doi.org/10.1186/s42397-024-00167-0
26 Bao, Y. *et al.* 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 (2011).
27 Hu, G. *et al.* Proteomic profiling of developing cotton fibers from wild and domesticated *Gossypium barbadense*. *New Phytol* **200**, 570-582 (2013). https://doi.org/10.1111/nph.12381
28 Qin, Y. *et al.* Proteomic Analysis of Differences in Fiber Development between Wild and Cultivated Gossypium hirsutum L. *Journal of Proteome Research* **16**, 2811-2824 (2017). https://doi.org/10.1021/acs.jproteome.7b00122
29 Yang, Y. *et al.* 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 (2023). https://doi.org/10.1038/s41598-023-28084-3
30 Wang, L., Liu, C., Liu, Y. & Luo, M. Fumonisin B1-Induced Changes in Cotton Fiber Elongation Revealed by Sphingolipidomics and Proteomics. *Biomolecules* **10** (2020). https://doi.org/10.3390/biom10091258
31 Mujahid, H. *et al.* Comparative Proteomic Analysis of Cotton Fiber Development and Protein Extraction Method Comparison in Late Stage Fibers. *Proteomes* **4** (2016). https://doi.org/10.3390/proteomes4010007
32 Zhou, X. *et al.* Proteomic profiling of cotton fiber developmental transition from cell elongation to secondary wall deposition. *Acta Biochim Biophys Sin (Shanghai)* **51**, 1168-1177 (2019). https://doi.org/10.1093/abbs/gmz111
33 Jiang, J., Shi, Z., Ma, F. & Liu, K. Identification of key proteins related to high-quality fiber in Upland cotton via proteomics analysis. *Plant Cell Rep* **41**, 893-904 (2022). https://doi.org/10.1007/s00299-021-02825-y
34 Hu, G., Koh, J., Yoo, M. J., Chen, S. & Wendel, J. F. Gene-Expression Novelty in Allopolyploid Cotton: A Proteomic Perspective. *Genetics* (2015). https://doi.org/10.1534/genetics.115.174367
35 Lee, Y. & Szymanski, D. B. Multimerization variants as potential drivers of neofunctionalization. *Sci Adv* **7** (2021). https://doi.org/10.1126/sciadv.abf0984
36 Nie, L., Wu, G. & Zhang, W. 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 (2006). https://doi.org/10.1534/genetics.106.065862
37 Manzoni, C. *et al.* Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief Bioinform* **19**, 286-302 (2018). https://doi.org/10.1093/bib/bbw114
38 Lee, Y., Rani, H., Mallery, E. L. & Szymanski, D. B. A cell fractionation and quantitative proteomics pipeline to enable functional analyses of cotton fiber development. *bioRxiv*, 2024.2008.2008.603566 (2024). https://doi.org/10.1101/2024.08.08.603566

39 Hu, G., Koh, J., Yoo, M.-J., Chen, S. & Wendel, J. F. Gene-Expression Novelty in Allopolyploid Cotton: A Proteomic Perspective. *Genetics* **200**, 91 (2015). https://doi.org/10.1534/genetics.115.174367

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

41 Tuttle, J. R. *et al.* Metabolomic and transcriptomic insights into how cotton fiber transitions to secondary wall synthesis, represses lignification, and prolongs elongation. *Bmc Genomics* **16** (2015). https://doi.org/ARTN 477
10.1186/s12864-015-1708-9

42 Qin, Y. M. *et al.* Saturated very-long-chain fatty acids promote cotton fiber and Arabidopsis cell elongation by activating ethylene biosynthesis. *Plant Cell* **19**, 3692-3704 (2007). https://doi.org/10.1105/tpc.107.054437

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

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

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

46 Grover, C. E. *et al.* A high-resolution model of gene expression during Gossypium hirsutum (cotton) fiber development. *bioRxiv*, 2024.2007.2020.604417 (2024). https://doi.org/10.1101/2024.07.20.604417

47 Swaminathan, S. *et al.* Daily glycome and transcriptome profiling reveals polysaccharide structures and correlated glycosyltransferases critical for cotton fiber growth. *The Plant Journal* **120**, 1857-1879 (2024). https://doi.org/https://doi.org/10.1111/tpj.17084

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

49 Cui, H. C. *et al.* SPINDLY, ERECTA, and Its Ligand STOMAGEN Have a Role in Redox-Mediated Cortex Proliferation in the
Root. *Mol Plant* **7**, 1727-1739 (2014). https://doi.org/10.1093/mp/ssu106

50 Sugano, S. S. *et al.* Stomagen positively regulates stomatal density in Arabidopsis. *Nature* **463**, 241-244 (2010). https://doi.org/10.1038/nature08682

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

52 Wen, X. *et al.* Molecular studies of cellulose synthase supercomplex from cotton fiber reveal its unique biochemical properties. *Sci China Life Sci* **65**, 1776-1793 (2022). https://doi.org/10.1007/s11427-022-2083-9

53 He, Q. *et al.* COBRA-LIKE 9 modulates cotton cell wall development via regulating cellulose deposition. *Plant Physiol* **197** (2024). https://doi.org/10.1093/plphys/kiae675

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

55 Kangben, F. *et al.* Phylogenetic and functional analysis of tiller angle control homeologs in allotetraploid cotton. *Frontiers in Plant Science* **14** (2024). https://doi.org/ARTN 1320638
10.3389/fpls.2023.1320638

56 Yu, J. *et al.* CottonGen: The Community Database for Cotton Genomics, Genetics, and Breeding Research. *Plants (Basel)* **10** (2021). https://doi.org/10.3390/plants10122805

57 Yang, Z. *et al.* CottonMD: a multi-omics database for cotton biological study. *Nucleic Acids Res* **51**, D1446-D1456 (2023). https://doi.org/10.1093/nar/gkac863

58 McBride, Z. *et al.* A label-free mass spectrometry method to predict endogenous protein complex composition. *Molecular & Cellular Proteomics* **18**, 1588 (2019). https://doi.org/10.1074/mcp.RA119.001400

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

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

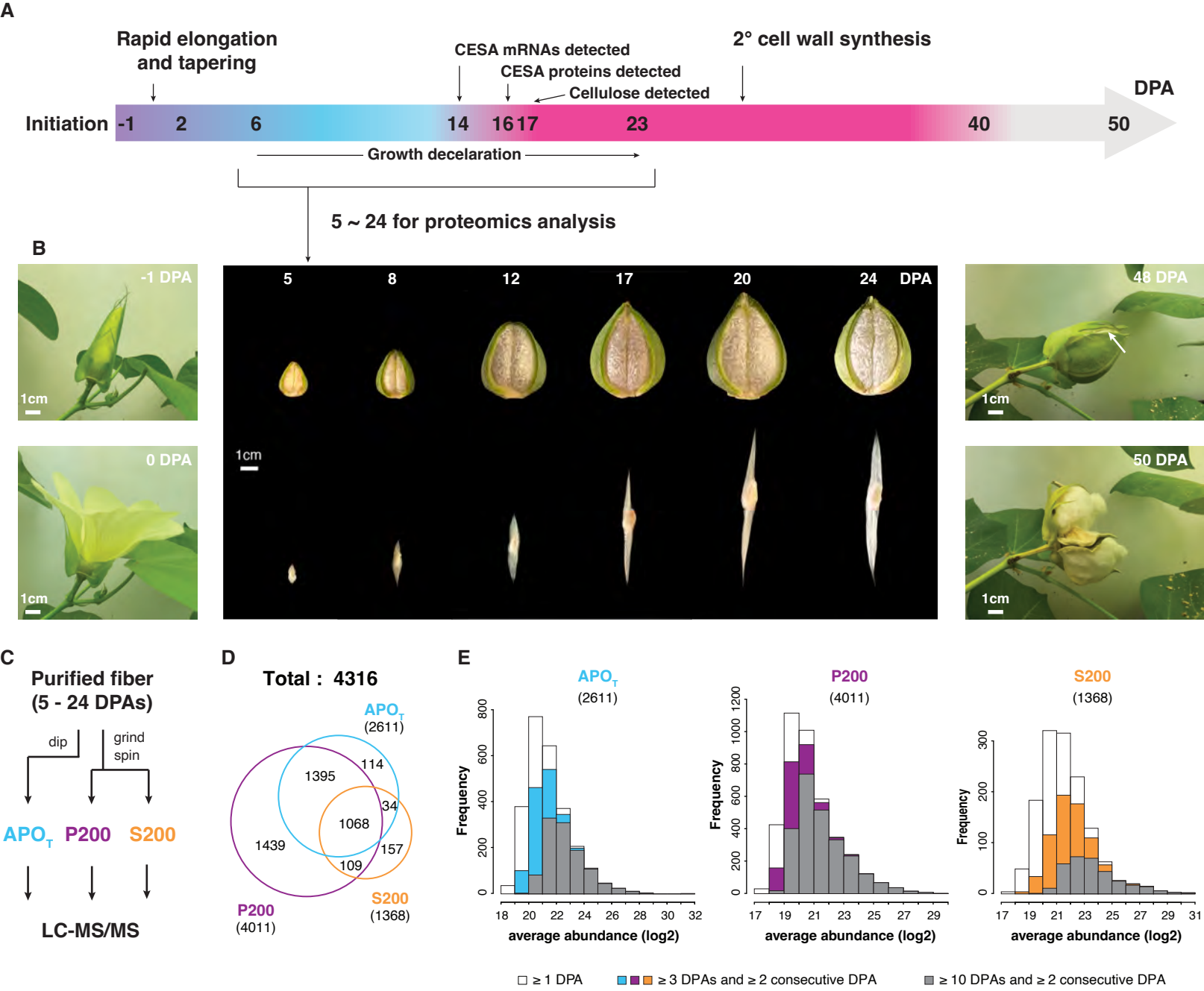
61 Lee, Y., Okita, T. W. & Szymanski, D. B. A co-fractionation mass spectrometry-based prediction of protein complex assemblies in the developing rice aleurone-subaleurone. *The Plant Cell* **33**, 2965-2980 (2021). https://doi.org/10.1093/plcell/koab182

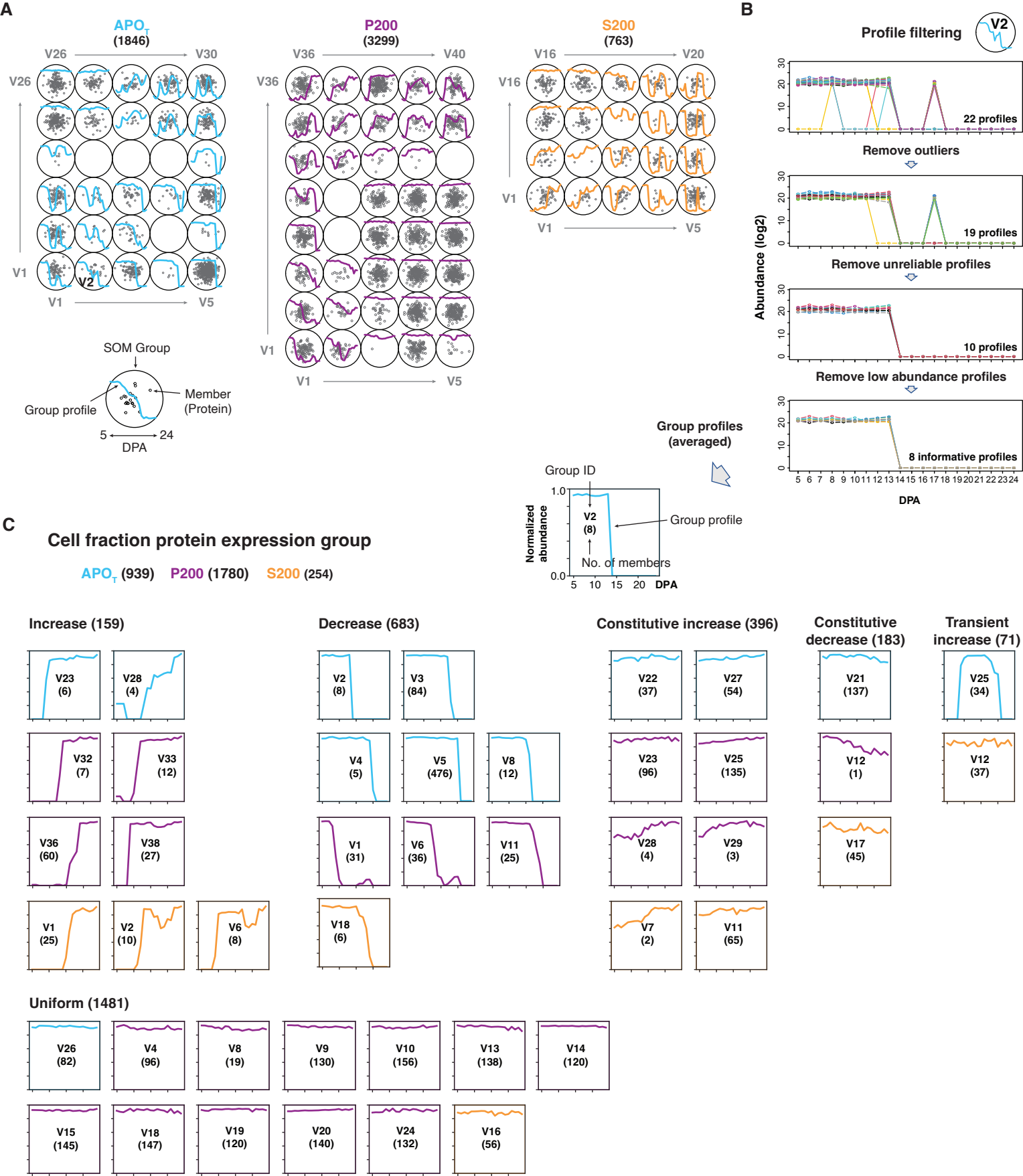
62 Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* **13**, 2513-2526 (2014). https://doi.org/10.1074/mcp.M113.031591

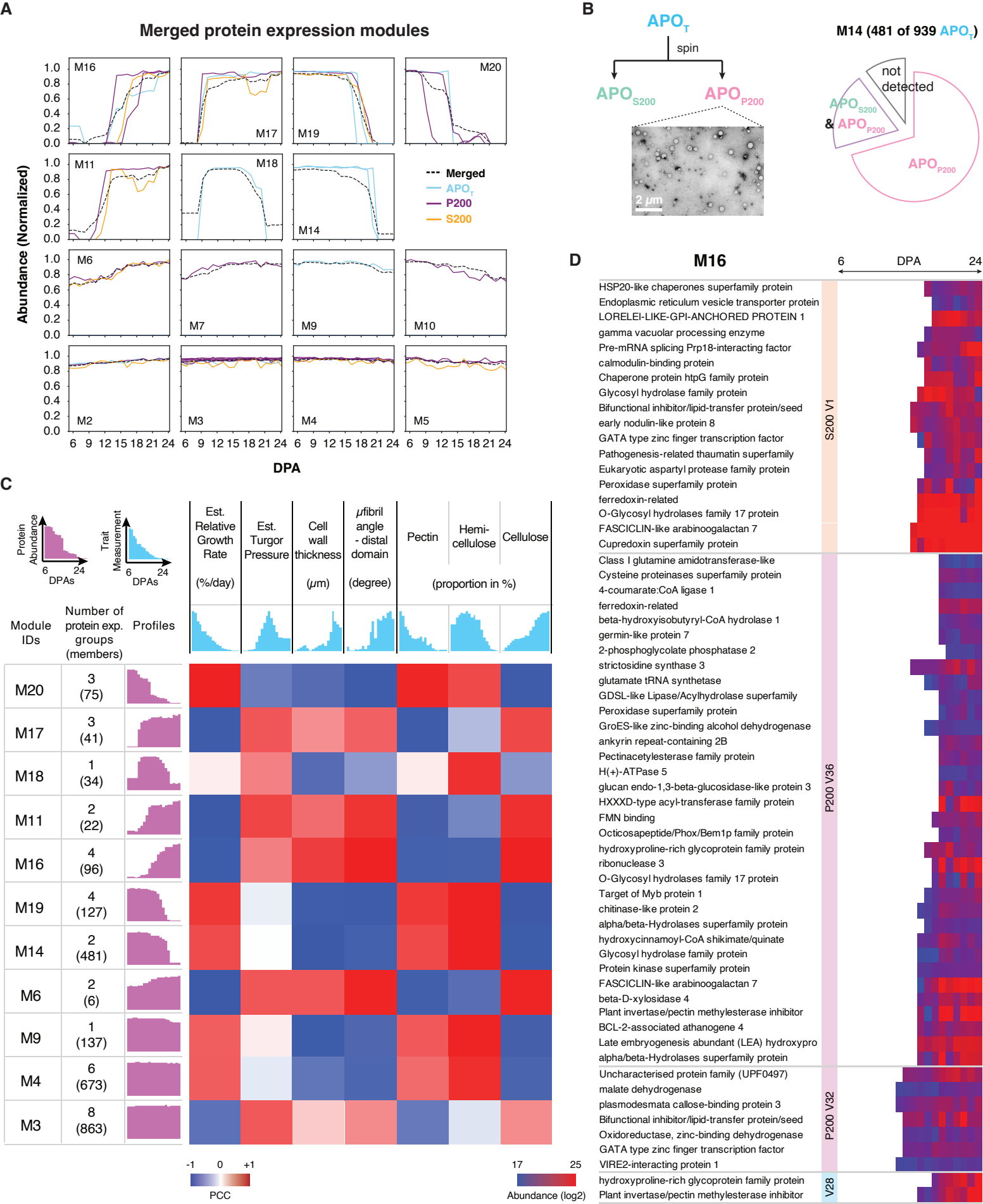
63 R: A language and environment for statistical computing v. 3.5.1 (R Foundation for Statistical Computing, Vienna, Austria, 2018).

64 RStudio: Integrated Development Environment for R v. 1.1.463 (RStudio, Inc., Boston, MA, 2018).

65 Perez-Riverol, Y. *et al.* The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research* **50**, D543-D552 (2021). https://doi.org/10.1093/nar/gkab1038


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
1cm
0 DPA
1cm
5
8
12
17
20
24
DPA
1cm
48 DPA
1cm
50 DPA
1cm
C
Purified fiber (5 - 24 DPAs)
dip
grind
spin
APO$_T$
P200
S200
LC-MS/MS
D
Total : 4316
APO$_T$ (2611)
1395
114
34
1068
1439
157
109
P200 (4011)
S200 (1368)
E
APO$_T$ (2611)
P200 (4011)
S200 (1368)
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
Profile filtering
22 profiles
Remove outliers
19 profiles
Remove unreliable profiles
10 profiles
Remove low abundance profiles
8 informative profiles
Abundance (log2)
DPA
Group profiles
(averaged)
Group ID
Normalized
abundance
Group profile
No. of members
C
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)
Uniform (1481)


A
Merged protein expression modules
Abundance (Normalized)
DPA
M16
M17
M19
M20
M11
M18
M14
M6
M7
M9
M10
M2
M3
M4
M5
Merged
APO$_T$
P200
S200
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
DPAs
Trait Measurement
DPAs
Module IDs
Number of protein exp. groups (members)
Profiles
Est. Relative Growth Rate (%/day)
Est. Turgor Pressure
Cell wall thickness (μm)
μfibril angle - distal domain (degree)
Pectin
Hemi-cellulose
Cellulose
(proportion in %)
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
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
S200 V1
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 V36
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
P200 V32
hydroxyproline-rich glycoprotein family protein
Plant invertase/pectin methylesterase inhibitor
V28

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 (50)
mRNA decrease and P200 flat (79)
mRNA increase and P200 flat (13)
V8 (10)
V9 (20)
V10 (18)
V91 (27)
V92 (22)
V81 (1)
V71 (15)
V72 (5)
V61 (12)
V62 (15)
V51 (13)
V52 (19)
V30 (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
mRNA abundance (log2)
0.1
10.0
Protein abundance (log2)
17
23