Glycoproteome of elongating cotton fiber cells

Saravanan Kumar¹, Krishan Kumar¹, Pankaj Pandey¹, Vijayalakshmi Rajamani¹, Kethireddy Venkata Padmalatha², Gurusamy Dhandapani², Mogilicherla Kanakachari², Sadhu Leelavathi¹, Polumetla Ananda Kumar² & Vanga Siva Reddy¹*

¹Plant Transformation Group, International Centre for Genetic Engineering & Biotechnology (ICGEB), New Delhi, India

²National Research Centre on Plant Biotechnology (NRCPB), IARI, New Delhi, India

*To whom correspondence should be addressed: Vanga Siva Reddy, Plant Transformation Group, International Centre for Genetic Engineering and Biotechnology (ICGEB), Aruna Asaf Ali Marg, New Delhi 110067, India. Tel: +91 11 2674 1358; Email: vsreddy@icgeb.res.in
Glycopeptide capture

Tryptic peptides were suspended in buffer containing 10mM HEPES-NaOH (pH 7.5), 1mM CaCl₂, 1mM MnCl₂, 1mM MgCl₂ and subjected to lectin affinity chromatography in a manually packed spin column containing 0.2ml of concanavalin A (Con A) sepharose resin (Sigma). The resin was equilibrated with binding buffer (10mM HEPES-NaOH (pH 7.5), 1mM CaCl₂, 1mM MnCl₂, 1mM MgCl₂ and 1mM PMSF) for 10 column volumes followed by binding of peptide mixture at low flow rate. Unbound sample was washed with 5 column volumes of binding buffer followed by elution with buffer containing 10mM HEPES-NaOH (pH 7.5), 0.2 M methyl α-D mannopyranoside and 0.5 M methyl α-D mannopyranoside for 3 column volumes respectively. Eluted glycopeptides were pooled and immediately subjected to hydrophobic interaction chromatography (HIC) using 0.2 ml of Sepharose CL-4B resin pre-equilibrated with HIC buffer A containing water: ethanol: butanol (1:1:5, v/v), 1mM MgCl₂ and 1mM MnCl₂. Unbound sample was washed with HIC buffer A for 5 column volumes followed by elution with HIC buffer B containing water: ethanol (1:1, v/v) for 3 to 5 column volumes. Eluted samples were dried in vacuum and dissolved either in 50mM sodium phosphate buffer (pH 7.0) for PNGase F digestion or in 50mM citrate phosphate buffer (pH 5.0) for PNGase A digestion. Samples were incubated overnight at 37°C for deglycosylation.

Sample excision and gel phase digestion

Excised silver stained gel pieces were destained with solution containing 2 mg/ml potassium ferricyanide (K₃Fe(CN)₆) and 0.4 mg/ml sodium thiosulphate (Na₂S₂O₅·5H₂O) by washing them twice. Coomassie stained gel pieces were destained with solution containing 50% methanol, 10% acetic acid (v/v). The gel pieces were then rinsed twice with MQ H₂O and equilibrated with 100 mM ammonium bicarbonate (NH₄HCO₃) for 20 min at room temperature with gentle agitation. The gel pieces were washed with 1:1 (v/v) NH₄HCO₃ and acetonitrile (ACN) and dehydrated with 100% ACN for 20 min at room temperature. The ACN solution was then discarded and gel pieces were vacuum dried. The sample was subjected to reduction with 10 mM dithiothreitol (DTT) in 50 mM NH₄HCO₃ for 45 min at 56°C followed by alkylation with 50 mM iodoacetamide (IAA) in 50 mM NH₄HCO₃ for 30 min in dark at room temperature. The gel pieces were rinsed briefly with 1:1 (v/v) NH₄HCO₃ and ACN solution. Reduction and alkylation steps were skipped for 2D spot digestion. The gel pieces were dehydrated with 100% ACN for 20 min and were vacuum-dried. To the dried gel pieces, enzyme solution containing 15 ng/µl trypsin (Promega) in 25mM NH₄HCO₃, 1mM CaCl₂
was added and incubated for 16 to 18 h at 37 °C. Following overnight incubation, peptides were extracted twice with 50 µl of 1% trifluoroacetic acid (TFA) in 60% ACN and the extract was concentrated under vacuum.

**Fractionation of peptides by SCX chromatography**

The peptide samples were fractionated using SCX column (Polysulfoethyl A, 4.6 mm x100 mm) on an Akta Purifier FPLC system (GE Healthcare). Peptides were eluted with a 100 min segmented three step gradient at 0.3 ml/min using a buffer system containing 0 M (buffer A) and 0.5 M KCl (buffer B) in 10 mM KH₂PO₄ (pH 2.75), 25% ACN. Briefly the following gradient was employed: 0-25% B for 5 min, 25-50% for 40 min, 50-100% B for 10 min followed by final washing with 100% B for 30 min. Fractionation was monitored at 215 nm. 30 fractions of 0.5 ml each was collected and pooled according to the sample complexity into a total of 23 fractions and desalted using C18 Sep Pak cartridges (Waters). Desalted peptides were dried in vacuum and frozen.

**Nano-LC based Reverse phase separation:**

Peptide pellets were re-dissolved in 20 µL of 0.1% TFA, and 15 µL of sample was bound onto a 100 µm i.d. x 20 mm EasyLC precolumn (Proxeon biosystems) at a flow rate of 5 µL/min. Reverse phase separation of the peptide mixture was performed in a nano-LC system (Proxeon biosystems) using a C-18 analytical column of 75 µm i.d. x 100 mm length at a flow rate of 300 nL/min with a solvent system comprising solvent A (0.1% TFA in 5% ACN, v/v), and solvent B (0.1% TFA in 90% ACN, v/v). The gradient conditions were 0% B for 5 min, followed by a linear gradient of 0% to 45% B for 65 min, 45% to 100% B for 1 min and 100% B for 10 min. The column eluants were directed to a Proteineer fc fraction collector (Bruker Daltonics) and fractions were spotted every 10 s onto Prespotted anchor chip 384/96 (PAC) target plates.

**Mass spectrometry based protein identification**

**MALDI TOF/TOF instrument parameters**

Spectral measurements were performed using the Ultraflex III MALDI TOF/TOF instrument (Bruker Daltonics, Germany) in positive ion reflector mode with the accelerating voltage of 25 kV (Ion source 1) and 21.85 kV (Ion source 2) respectively. The laser wavelength and frequency was 337 nm and 100 Hz. The final mass spectra were produced by averaging 1200 to 1500 laser shots.
For MS/MS based sequencing, the following parameters were used during acquisition. The precursor peptide ions were fragmented in positive mode using the LIFT.ift method (Bruker Daltonics, Germany). Briefly; the accelerating voltages were 8.00 kV, 7.15 kV for Ion source 1 and Ion source 2. The reflector 1 and 2 were set to 29.5 kV and 13.9 kV respectively with Lift 1 and 2 set at 19 kV and 2.8 kV. The detector range was set at 40-5000 da with a sampling rate of 2GS/sec.

**Protein identification from 2D gel spots**

Peptide pellets obtained from 2D gel spots were dissolved in 10 µL of 0.1% TFA, followed by desalting using C18 ziptips (Millipore). Peptide samples were mixed with equal volumes of matrix solution containing 10 mg/ml α-cyano hydroxy cinnamic acid in ACN/0.1%TFA (1:1, v/v) and allowed to air dry. Samples and calibration standards with the same matrix composition were spotted adjacent to each other on the ground steel target plate for optimal calibration. Spectra were externally calibrated with the calibration standard prior to peptide mass fingerprint (PMF) acquisition. For MS/MS based analysis fragmentation was performed using laser induced dissociation (LID). Fragmented peptides were annotated, smoothened and baseline subtracted through Flexanalysis software version 3.0 and was sent to database search using Biotools software version 3.2. The database search parameters were set as described: fragment masses were searched in NCBInr database through mascot search engine, taxonomy was set as Viridiplantae, enzyme was set as trypsin, fixed modifications included carbamidomethylation of cysteine, variable modifications included oxidation of methionine, protein mass was unrestricted, missed cleavage was set to 1, MS tolerance of +/-100 ppm and MS/MS tolerance of +/-0.75 da. Only peptides with an individual ion score of >45 (p<0.05) were considered for protein identification. Unidentified samples were further analysed through Nano-LC-MALDI TOF/TOF.

**Protein identification from Solution phase and 1D-Gel phase digested samples**

Peptide samples were injected into Nano-LC system, fractionated and spotted onto PAC (Pre Anchored Chip) targets. Sample plates were subjected to MADLI TOF/TOF based acquisition in automated mode through WARP-LC 1.2 (Workflow Administration by Result driven Processing-Liquid Chromatography, Bruker Daltonics) software tool. Briefly the following steps were employed for automatic acquisition: sample carrier plates were subjected to pre-teaching to ensure optimum and complete acquisition of the sample spots. Mass list calculation was done through WARP-LC
interface. A manually updated background list containing trypsin autolysis peaks, matrix peaks, keratin peaks was used during mass list generation (Supplemental Table 9). Spectral peaks (m/z) corresponding to background peaks were excluded for MS/MS measurement. Only those peaks (precursors) that have an S/N > 15 were included for measurement. Post acquisition processes including mass annotation, baseline subtraction, smoothening were performed using Flexanalysis software version 3.0 through WARP-LC. Protein identification was achieved using Biotools version 3.2 through in-house licensed Mascot server (version 2.3 - March 2010).

Database search strategy for protein identification

Two different databases were employed to achieve protein identification; they were (i) NCBI\textit{nr} database (06/03/2010) containing 10,551,781 sequences (total) including 290,173 sequences from green plants (\textit{Viridiplantae}) and (ii) 	extit{Gossypium raimondii} protein database containing 40,976 sequences downloaded from CottonGen website (ftp://ftp.bioinfo.wsu.edu/species/Gossypium_raimondii/CGP-BGI_G.raimondii_Dgenome/genes/). The following database search parameters were used for protein identification: (i) taxonomy was set to \textit{Viridiplantae} (Green plants), (ii) enzyme was set as semi trypsin/trypsin, (iii) fixed modifications: carbamidomethylation of cysteine, (iv) variable modifications: oxidation of methionine, carboxymethylation of lysine (K) and N-terminus of protein, (v) protein mass was unrestricted, (vi) missed cleavage was set to 1, (vii) MS tolerance of +/-100 ppm and (viii) MS/MS tolerance of +/-0.75 da. Search strategy involved first phase of identification using complete dataset against NCBI\textit{nr} database (protein database) followed by second phase of identification against the \textit{Gossypium raimondii} protein sequence database. The following criteria were used for identification: (i) significance threshold was set to achieve p<0.02; (ii) the expectancy cut off was set to 0.05; and (iii) individual ion score > 45 was only considered for identification. These parameters led to a FDR value < 1% in both the above mentioned database search strategies. FDR calculation was achieved through an inbuilt option available in Mascot search engine (Decoy strategy). The database search strategy for deglycosylated peptide identification is the same as mentioned above with minor additions: (i) variable modifications included deamidation of Asparagine, the peptide is considered as formerly glycosylated only if the deamidated asparagine (N) was followed by X-S/T (any amino acid except proline - serine/threonine).
Spectral counting based semi quantitative analysis

Spectral counting based semi quantitative analysis was carried out to calculate the protein abundance. Spectral counts (SpC) correspond to the number of mass spectra mapping to a unique peptide sequence belonging to unique protein member. Spectral counts were retrieved for all the unique protein entries from both gel based and gel free approaches independently. The spectral counts for each protein were summed to yield the total spectral counts and the same was normalized to the total number of spectra obtained within a single approach (‘percentage of the total spectral count’; Supplementary Table 3). Finally, the average of total spectral count % from both the gel based and gel free approaches were utilized in the current study.

Graphical tools

The venn diagrams presented in the current study were created using Venny tool (http://bioinfogp.cnb.csic.es/tools/venny) (1) and Venn Diagram Plotter tool (http://omics.pnl.gov/software/VennDiagramPlotter.php).

Plant materials for Quantitative Real time PCR analysis

Cotton plants (G.hirsutum cv. Coker 310) were gown in climate controlled green house. Leaves were collected from young plant. Flowers were tagged on the day of anthesis and considered as 0 dpa (days post anthesis). Cotton bolls were collected at 0, 5, 10, 15, 20 and 25 dpa stages. Roots were collected from 15 days old seedling cultured in growth chamber. Harvested cotton bolls, leaves and roots were immediately frozen in liquid nitrogen and stored at -70°C until used for total RNA extraction.

RNA isolation

Fibres were stripped from the ovules of 10, 15, 20 and 25 dpa. Total RNA was isolated from leaves, roots, fibre-bearing ovules at 0 and 5 dpa and fibres at 10, 15, 20 and 25 dpa. RNA was isolated using Spectrum™ Plant Total RNA kit (Sigma, USA) according to the manufacturer’s protocol.
Quantitative real-time PCR (qRT-PCR)

Quantitative real-time PCR analysis was performed on randomly selected genes that encode glycoproteins. First strand cDNA was synthesized using 1 µg of total RNA using Affinity Script qRT-PCR cDNA synthesis kit (Stratagene, Agilent Technologies, USA) according to the manufacturer’s instructions. cDNA was diluted to 10 times in nuclease free water and 1 µl of diluted cDNA was used for qRT-PCR. Gene specific qRT-PCR primers were designed using Primer Quest software (www.eu.idtdna.com) and cross checked with Becon Designer (www.premierbiosoft.com/qOligo/Sequence.jsp?PID=1) to evaluate parameters like cross dimer, primer-dimer and hairpin structures. The list of primers is presented in Supplemental Table 8. The qRT-PCR was performed in triplicates using the Brilliant-III Ultra Fast SYBR Green QPCR master mix in Stratagene MX 3005P (Agilent Technologies, USA) detection system. The GhPP2A1 gene (Accession No: DT545658) from G. hirsutum was used as reference gene to normalize the expression values (2). The thermal cycling program was set as follows; 5 min at 95°C, followed by 40 cycles of amplification with 30 s of denaturation at 95°C, 30 s of annealing at 55°C or 60°C and 30 s of extension at 72°C. Amplicons were subjected to the melt curve analysis to check the specificity of the amplified products. The expression level in leaf tissue was selected as calibrator. The fold change value (log₂ scale) for mRNA expression level compared/relative to expression in leaf tissue was calculated using comparative ∆∆Ct method. In this method the fold change = $2^{-\Delta\Delta C_t}$ where $\Delta\Delta C_t = (C_t \text{ (gene of interest)} - C_t \text{ (GhPP2A1)}) \text{ leaf} - (C_t \text{ (gene of interest)} - C_t \text{ (GhPP2A1)}) \text{ other stage}$. Heat map with hierarchical clustering using log₂ fold change value was generated with Multi Experiment Viewer software (3).
Supplemental References

**Supplemental Figure legends**

Supplemental Figure 1: SCX chromatography elution profile of tryptic peptides from Con A bound fiber glycoproteins. X-axis represents retention time in minutes (min) and Y-axis represents absorbance at 215 nm (mAu).

Supplemental Figure 2: Sequence alignment and domain organisation of identified AGP members in the current study. Amino acids highlighted in red correspond to the identified protein sequence regions (MS/MS). Regions showing amino acid differences are shown within box. Identified N-linked glycosylation sites are highlighted in Red. Signal/Secretory peptide is highlighted Turquoise. AP, PA, SP, TP - stretches are highlighted in yellow (AGP like modules). Fasciclin domain is highlighted in green. Potential sites for GPI anchors are highlighted in Blue (Supplemental Table 3, 5)

Supplemental Figure 3: Sequence alignment and domain organisation of identified FLA members in the current study. Amino acids highlighted in red correspond to the identified protein sequence regions. Regions showing amino acid differences are shown within box. Identified N-linked glycosylation sites are highlighted in Red. Signal/Secretory peptide is highlighted Turquoise. AP, PA, SP, TP - stretches are highlighted in yellow (AGP like modules). Fasciclin domain is highlighted in green. Potential sites for GPI anchors are highlighted in Blue (Supplemental Table 3, 5)

Supplemental Figure 4: Annotated line spectrum corresponding to the unknown/novel peptide sequences homologous to fructokinase (a) and benzoquinone reductase (c) identified in the current study. Annotated line spectrum corresponding to the peptide sequences homologous to the novel sequences of fructokinase (b) and benzoquinone reductase (d).
Supplemental Figure 5: Diagrammatic representation of representative N-linked glycosylation sites identified in the current study harbouring two different core N-linked glycans as suggested by independent deglycosylation reactions. N-linked glycosylation sites of FLA 6 (a), Mono copper oxidase like protein (b), vacuolar invertase 1(c), Polygalacturonase inhibiting protein (d)

Supplemental Figure 6: Quantitative Real Time -PCR analysis during fiber development stages (0, 5, 10, 15, 20 and 25 dpa) and root tissue relative to leaf. Y-axis represents the log$_2$ fold change values for mRNA expression level at various developmental stages as relative to expression in leaf tissue.
Supplemental Table legends

**Supplemental Table 1:** List of proteins identified from 2D-PAGE gels (pI range: 4-7 and 3-10 NL).

**Supplemental Table 2:** List of glycosylation sites identified using glycopeptide capture followed by deglycosylation using PNGase F and A.

**Supplemental Table 3:** List of unique proteins identified from four independent proteomic approaches using two different databases (NCBIinr and *G. raimondii* protein database). Protein sequence coverage, number of unique peptides, spectral counts corresponding to unique peptides and percentage of spectral counts are shown. Also, *in silico* based prediction of N-linked glycosylation site, transmembrane domain regions, signal peptide cleavage site, GPI-anchor site, protein family and domain classification are shown.

**Supplemental Table 4:** Percentage of unique proteins classified under various protein families and percentage of total spectra corresponding to protein families.

**Supplemental Table 5:** List of AGPs and FLAs identified in the present study. Molecular weight and pI distribution of the identified AGPs and FLAs in 1D and 2D PAGE gels. Signal peptide cleavage site, GPI anchor site, PAST% and amino acid position of FAS domain are shown.

**Supplemental Table 6:** Percentage of unique proteins classified under various protein families from plant glycoproteome and cotton fiber glycoproteome datasets.

**Supplemental Table 7:** List of proteins classified under protein families from cotton fiber and other plant glycoproteome datasets.

**Supplemental Table 8:** List of primers used for qRT-PCR analysis of the selected glycoprotein encoding genes.

**Supplemental Table 9:** List of background masses (m/z) from trypsin, keratin and MALDI matrix that were not considered during the protein identification.
S Kumar et al Supplemental Figure 1
S Kumar et al Supplemental Figure 3
Novel/unknown amino acid regions determined by MS/MS are highlighted in red color.
S Kumar et al Supplemental Figure 6
S Kumar et al Supplemental Figure 6