Published in final edited form as:

Electrophoresis. 2013 August ; 34(16): 2417–2431. doi:10.1002/elps.201200656.

A workflow for large-scale empirical identification of cell wall *N*-linked glycoproteins of tomato (*Solanum lycopersicum*) fruit by tandem mass spectrometry

Theodore W. Thannhauser^{1,*}, Miaoqing Shen¹, Robert Sherwood², Kevin Howe¹, Tara Fish¹, Yong Yang¹, Wei Chen², and Sheng Zhang^{2,*}

¹Robert W. Holley Center for Agriculture and Health, USDA-ARS, Cornell University, Ithaca, NY

²Institute for Biotechnology and Life Science Technologies, Cornell University, Ithaca, NY

Abstract

Glycosylation is a common post-translational modification of plant proteins that impacts a large number of important biological processes. Nevertheless, the impacts of differential site occupancy and the nature of specific glycoforms are obscure. Historically, characterization of glycoproteins has been difficult due to the distinct physicochemical properties of the peptidyl and glycan moieties, the variable and dynamic nature of the glycosylation process, their heterogeneous nature, and the low relative abundance of each glycoform. In this study, we explore a new pipeline developed for large-scale empirical identification of *N*-linked glycoproteins of tomato fruit as part of our ongoing efforts to characterize the tomato secretome. The workflow presented involves a combination of lectin affinity, tryptic digestion, ion-pairing HILIC and precursor ion-driven data dependent MS/MS analysis with a script to facilitate the identification and characterization of occupied *N*-linked glycosylation sites. A total of 212 glycoproteins were identified in this study, in which 26 glycopeptides from 24 glycoproteins were successfully characterized in just one HILIC fraction. Further precursor ion discovery (PID)-based MS/MS and deglycosylation followed by high accuracy and resolution MS analysis were used to confirm the glycosylation sites and determine site occupancy rates. The workflow reported is robust and capable of producing large amounts of empirical data involving *N*-linked glycosylation sites and their associated glycoforms.

Keywords

N-linked glycosylation; plant glycoproteomics; precursor ion scan; HILIC; lectin affinity chromatography

*To whom correspondence should be addressed: Dr. Theodore Thannhauser (tt34@cornell.edu; Tel: 607-255-8808; Fax: 607-254-6510), and Dr. Sheng Zhang (sz14@cornell.edu; Tel: 607-255-6802; Fax: 607-254-4847).

Publisher's Disclaimer: USDA-ARS disclaimer: Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. The USDA is an equal opportunity provider and employer.

All authors have declared no conflict of interest.

1 Introduction

Glycosylation represents an important post-translational modification of proteins, affecting many protein functions and cellular activities. It is estimated that 50% of all proteins are glycosylated based on the fact that two-thirds of entries in Swiss-Prot database were found to contain at least one *N*-glycosylation consensus sequon (NXS/T) with an expected site occupancy rate of 2/3, plus approximate 10% of solely *O*-glycosylated proteins [1, 2]. Glycoproteins can function as structural elements, as signaling molecules, as recognition markers for the assembly of protein complexes and can affect protein folding, solubility and stability. Thus, understanding the detailed structure of glycoproteins has the potential to provide valuable insights to their global functions.

In recent years, most research in glycobiology has focused on mammalian systems, particularly from the perspective of its biomedical applications [3, 4], which has driven a great deal of discovery. Unfortunately, far less is known about glycosylation in plant systems, although in general it has been determined that plant cells contain many of the same glycans as are found in animal cells. Furthermore, *N*-linked glycosylation is known to be a common feature of plant proteins, particularly of those that reside at the cell wall where they fulfill important roles in cell wall modification, sugar metabolism, signaling, and defense [5]. Knowledge of the identities and activities of the glycoproteins found in distinct subcellular compartments, such as the cell wall, coupled with information concerning the position, structure and relative proportions of the individual glycoforms is fundamental to developing an understanding of how glycosylation status modulates protein function and activity. However, various technical issues have inhibited the development of such an understanding.

Structural elucidation of glycans, glycopeptides and glycoproteins has been notoriously difficult as glycans are typically highly heterogeneous and chemically complex, containing a wide variety of different intersaccharide linkages, with a wide range of abundance between glycoforms and occupancy of glycosylated sites. Additionally, the hydrophilic nature of glycans and frequent multiple adduct formation has often caused poor retention by reverse phase (RP)-chromatography, reduced ionization efficiency and low MS signals. Furthermore, the labile nature of the glycan-peptide bond often causes the neutral loss of individual carbohydrates and few fragmented ions from the core peptide during CID, yielding little or no information for peptide identification, glycol-site determination and full glycan sequence. Finally, due to the high glycan heterogeneity, identification of the glycan structure through bioinformatics tool-based database searching is often difficult and limited, and requires time-consuming manual interpretation.

Comprehensive analysis of plant glycoproteins involves the isolation and identification of the glycoproteins, determination of the location of the glycosylation sites, identification of glycosylated peptides, and elucidation of the glycan structures including sugar composition, linkages and branches as well as their relative abundances [6–8], each of which presents a specific set of technical challenges. To overcome these analytical challenges, many new strategies have been developed and implemented involving the development of three major tools: (1) selective enrichment technologies for glycoproteins, glycopeptides and released

glycans; (2) enhanced MS technology and development of more effective analytical workflows; and (3) newly developed software packages and algorithms for facilitating interpretation of MS fragmentation spectra. The majority of these technologies have been initially developed to study mammalian glycoproteins, and to date there are only a few reported examples of their successful application to characterize plant glycosylation [6, 9, 10]. Glycosylation in plants is structurally distinct from that of mammalian or yeast proteins [7, 11] and these structural differences, such as the absence of non-reducing terminal sialic acid in plant *N*-glycans and the presence of α -1,3-fucose (Fuc) in the innermost GlcNAc residue of the *N*-glycans, requires tailoring of existing analytical workflows.

Advanced tandem MS technologies provide an effective means for glycopeptide amino acid sequencing and structural elucidation of oligosaccharides. It has recently become possible to combine advances in selective enrichment techniques, chromatography and mass spectrometry into robust workflows capable of overcoming some of these difficulties [10]. Improved protocols for the identification and characterization of glycoproteins and the global proteome have been reviewed recently in the context of both plant science [7, 11] and gene discovery for disease models [12].

Since *N*-glycosylation is a common post translational modification of the plant proteins that navigate the secretory pathway, Minic et al. [13] used an approach, coupling tissue homogenization with glycoprotein isolation by Concanavalin A (Con A) affinity chromatography and two-dimensional electrophoresis (2DE)-based fractionation prior to LC-MS to generate an enriched sample of cell wall proteins. A total of 102 presumably glycosylated proteins from *Arabidopsis thaliana* were identified in this effort [13]. In this study, we intend to utilize the tomato cell wall as a test case to explore a new pipeline we have developed and implemented for large-scale empirical identification of cell wall *N*-linked glycoproteins of tomato fruit by tandem mass spectrometry. Here, we present an approach that builds on our previous studies involving the cataloging of the cell wall proteome [14] and glycoproteins characterization [10]. Specifically, we employ a workflow combining high-speed centrifugation using sucrose cushions and gradients with salt-based extractions to produce an enriched sample of cell wall proteins followed by lectin affinity chromatography to isolate the cell wall glycoproteins from tomato fruits at the mature-green stage of development. Following tryptic digestion, we then employed a two dimensional fractionation involving HILIC and nano scale RPLC coupled with precursor ion (PI) driven data dependent MS/MS analysis along with a bioinformatic script to facilitate the identification and characterization of *N*-linked glycoforms in cell wall complex samples. Furthermore, precursor ion discovery (PID)-based MS/MS and deglycosylation of the HILIC fractions followed by high mass accuracy and resolution MS was used to confirm the glycosylation sites that are actually occupied and determine occupancy rates.

2 Materials and methods

2.1 Materials

Deionized water using Milli-Q ultrapure or Barnstead Nanopure water systems (Millipore, Bedford, MA or Thermo Fisher, Waltham, MA, respectively) was used and mass spectrometry grade acetonitrile and acetone were purchased from Fisher Scientific (Fair

Lawn, NJ). The protease inhibitor cocktail, ammonium biocarbonate, iodoacetamide (IAM), dithiothreitol (DTT), urea, sodium dodecyl sulfate (SDS) and methyl- α -D-mannopyranoside were obtained from Sigma-Aldrich (St. Louis, MO). Sequencing grade modified trypsin was obtained from Promega (Madison, WI). PNGase A was purchased from New England BioLabs (Ipswich, MA). Quick Start™ Bradford Protein Assay was from BioRad (Hercules, CA). All other reagents used in this study were reagent grade or better. The HiTrap® desalting columns (5 mL) and HiTrap® Con A 4B columns (1 mL) were purchased from GE Healthcare (Piscataway, NJ). Other chemical reagents, unless otherwise noted, were obtained from Sigma-Aldrich (St. Louis, MO).

2.2 Workflow

The main objective in this study was to develop a generally applicable MS-based workflow for glycosylation profiling of complex plant samples in combination with various glycoprotein/peptide enrichment steps as an important part of our ongoing study involving the characterization of tomato cell wall proteins. An overview of the experimental strategy for tomato cell wall *N*-linked glycosylation and workflow used in this study is shown in Figure 1. To our knowledge, there is no existing report presenting such an extensive workflow, allowing for the large-scale characterization of *N*-linked glycosylation sites of cell wall glycoproteins.

2.3 Plant material and isolation of cell wall proteins

Two independent sets of tomato (*Solanum lycopersicum* cv. Ailsa Craig) plants were grown in a greenhouse (14 h/10 h, 24 °C /16 °C day/night regime) until the fruit could be assigned to the ‘mature green’ (MG) stage of development by criteria previously described [15]. Crude cell wall preparations were obtained from the pericarp of this fruit by the following procedure: Pericarp tissue (30 g) was powdered in liquid nitrogen with a mortar and pestle, then homogenized in 40 mL of extraction buffer (17 mM Tris, pH 7.0, containing 5 mM β -mercaptoethanol and 10 μ L/g fresh weight of protease inhibitor cocktail) followed by centrifugation for 30 min at 15,000 $\times$ g (Beckman Avanti J-20 XP centrifuge, Fullerton, CA). After centrifugation, the supernatant (designated the soluble protein extract) was dialyzed using a Slide-A-Lyzer dialysis cassette (Pierce, Rockford, IL) with a MWCO of 3,500 against 4 L of 100 mM NaCl, 1 mM DTT, 10 mM sodium acetate (pH 6.0) and concentrated using Amicon centrifugal ultrafiltration devices (MWCO = 5,000), and stored at –80°C. The crude pellet was resuspended in 2 mL of buffer (17 mM Tris, pH 7.0, containing 5 mM β -mercaptoethanol and 20 μ L of the protease inhibitor cocktail), then overlaid on a 25 mL cushion of 60% sucrose, and centrifuged for 2.5 hrs at 50,000 $\times$ g (Beckman Optima XL-100k Ultracentrifuge). The pellet from the sucrose cushion was collected and resuspended in 2 mL of 17 mM Tris, pH 7.0, containing 5 mM β -mercaptoethanol and 20 μ L of protease inhibitor cocktail. This material was then loaded on a 20% – 60% sucrose gradient and centrifuged for 2 hrs at 85,000 $\times$ g. The resultant pellet containing the cell wall material was washed with 20 mL of a low-salt buffer (20 mM NaCl, 1 mM DTT, and 10 mM sodium acetate pH 6.0) and centrifuged at 15,000 $\times$ g for 20 min. The pellet was isolated, resuspended in a high-salt buffer, 6 mL of 1.7 M NaCl, 15 mM EDTA, 50 mM sodium citrate (pH 5.5) with gentle shaking for 1 hr at 4 °C, and centrifuged at 15,000 $\times$ g for 30 min. The supernatant containing cell wall associated proteins was

filtered through Miracloth and brought to 80% saturation with ammonium sulfate (5.2 g/10 mL). The pellet, obtained by centrifuging for 20 min at $15,000 \times g$, was dissolved in 5 mL of 100 mM NaCl, 1 mM DTT, 10 mM sodium acetate (pH 6.0) followed by dialysis (as above) against 4 L of 100 mM NaCl, 1 mM DTT, 10mM sodium acetate (pH 6.0) overnight at 4 °C.

2.4 Glycoprotein enrichment

To enrich for glycoproteins the high salt and low salt fractions described above were subjected to lectin affinity chromatography using 1.0 mL HiTrap® Concanavalin A (Con A) 4B columns on an AKTA Explorer liquid chromatography system (GEHealthcare, Piscataway, NJ) according to the manufacturers recommended protocol. The methyl- α -D-mannopyranoside was removed and exchanged to 50 mM ammonium bicarbonate using a 5.0 mL HiTrap® desalting column on an AKTA Explorer liquid chromatography system. Glycoprotein containing fractions were dried under vacuum using a Centrивap Concentrator (Labconco, Kansas City, MO) and stored at -80°C until used.

2.5 In-solution digestion of ConA enriched glycoproteins

The glycoproteins from the low salt and high salt washes of the cell wall fraction obtained from the sedimentation experiment described above were digested using the RapiGest protocol for recommended by the manufacturer (Waters, Milford, MA). Briefly, the glycoprotein pellets (150 μg) were dissolved in 100 μL of a 0.1% (v/v) RapiGest in 50 mM triethylammonium bicarbonate solution. Dithiothreitol was added to a final concentration of 5 mM and the solution was incubated at 56°C for 30 minutes. The sample was then cooled to room temperature and iodoacetamide was added to a final concentration of 15 mM. The sample was placed in the dark for 30 minutes, at room temperature, after which trypsin was added to a final substrate-enzyme ratio of 50 to 1 (w/w). The digest solution was incubated at 30°C overnight, followed by the addition of TFA to stop reaction. The RapiGest was hydrolyzed by incubating the solution for 45 minutes at 37°C . The sample was then centrifuged at $16,000 \times g$ for 10 minutes to pellet the RapiGest hydrolysis products. The tryptic peptides were desalted via reverse phase SPE on Sep-Pak® C₁₈ cartridges (0.1 g from Waters, Milford, MA), dried in a speedvac concentrator and stored at -80°C until analyzed.

2.6 Enrichment of glycopeptides by HILIC and deglycosylation

The tryptic peptides of glycoproteins were subjected to HILIC for partitioning glycopeptides and non-glycopeptides. HILIC was carried out using a Dionex UltiMate 3000 high performance LC system with a built-in microfraction collection option in its autosampler and UV monitoring (Thermo-Dionex, Sunnyvale, CA). The tryptic peptides of high-salt fraction were reconstituted in 80% ACN containing 0.25% TFA for ion pair normal phase separation [16, 17] and loaded onto a Polyhydroxyethyl A™ column (5 μm , $2.1 \times 200\text{ mm}$, 200 Å, PolyLC, MD) with 10% ACN as eluent A and 90% ACN as eluent B. The LC was performed using a gradient from 90 to 40% eluent B in 30 min at a flow rate of 200 $\mu\text{L}/\text{min}$. Forty four fractions were collected at 1 min intervals, pooled into 31 fractions based on the UV absorbance at 214 nm. The resulting fractions were dried and reconstituted in 100 μL of 0.5% formic acid (FA) for screening glycopeptide-containing fractions by nanoLC-MS/MS on a 4000 Q trap operating in the precursor ion (PI) scan-triggered, information dependent

acquisition (IDA) mode. A quarter (25 μ L) of the reconstituted fractions containing glycopeptides were further treated with 50 μ U of PNGase A at 37 °C for 16 hrs in 100 mM sodium citrate/sodium phosphate monobasic pH = 5.0. The PNGase A treated samples were cleaned up using Omix C18 tips, and reconstituted in 25 μ L of 0.2% FA prior to high resolution MS and MS/MS analysis on a LTQ Orbitrap Velos. Another quarter of each of the 31 fractions (25 μ L) were dried and reconstituted to 20 μ L of 0.2% FA for protein identifications using Synapt HDMS.

2.7 NanoLC-MS/MS analyses

To identify the glycoproteins in the 31 pooled HILIC fractions described above from both salt fractions, nanoLC-MS/MS was performed as described previously [18, 19]. Briefly, nano-LC separation of tryptic peptides was performed with a nanoACQUITY system (Waters), equipped with a Symmetry C18, 5 μ m, 20 mm $\times$ 180 μ m trapping column and a UPLC BEH C18 1.7 μ m, 15 cm $\times$ 75 μ m analytical column (Waters). The samples (5 μ L) were loaded to the trapping column at a flow rate of 7 μ L/min for 3 min. Trapped peptides were separated with a 30-min gradient of 2% to 40% ACN in 0.1% FA at a flow rate of 300 nL/min. Glu-fibrinopeptide B (100 fmol/ μ L) in 25% ACN with 0.1% FA was used as the lock mass compound and was delivered via the auxiliary pump of the LC system at a flow rate of 300 nL/min to the reference sprayer of the NanoLockSpray source of the Synapt HDMS. The eluted peptides were detected by the Synapt HDMS equipment through a nano ion source containing a 10 μ m analyte emitter (NewObjective, Woburn, MA). The Synapt HDMS was operated in the data-dependent acquisition (DDA) mode and the parameters were set as reported previously [18]. A real time dynamic exclusion window of 40 s was applied to each precursor ion selected for fragmentation. The acquisition mode was switched from MS to MS/MS when the abundance of an individual ion exceeded 25 counts per second (cps), and returned to MS mode when the total ion current for the MS/MS acquisition exceeded 10,000 cps or after 3 MS/MS scans on multiple charge states (2+,3+ and 4+) had been completed. The mass range for each MS survey scans was set to m/z 300–1500, and for MS/MS scans at m/z 50–2000. All data were acquired by MassLynx v4.1 (Waters).

The nanoLC-MS/MS analysis for characterization of glycosylation sites was performed on an UltiMate3000 nanoLC (Dionex, Sunnyvale, CA) coupled with a hybrid triple quadrupole linear ion trap mass spectrometer, the 4000 Q Trap (AB SCIEX, Framingham, MA). The tryptic peptides in each HILIC fraction (5 μ L) from high-salt sample were injected with an autosampler onto a PepMap C18 trap column (5 μ m, 300 μ m $\times$ 5 mm, Dionex) with 0.1% FA at 20 μ L/min for 1 min, then separated on a PepMap C18 RP nano column (3 μ m, 75 μ m $\times$ 15 cm, Dionex) using a 60-minute gradient of 10% to 35% ACN in 0.1% FA at 300 nL/min, followed by a 3-min ramp to 95% ACN-0.1% FA and a 5-min hold at 95% ACN-0.1% FA. MS data acquisition was performed using Analyst 1.4.2 software (Applied Biosystems) for PI scan-triggered IDA analysis. The precursor ion scan of the oxonium ion (HexNAc⁺ at m/z 204.08) was monitored using a step size of 0.2 Da across a mass range of m/z 500 to 1600 and the parameters were set as reported previously [10]. For IDA analysis, each precursor ion scan was followed by one enhanced resolution scan and the two highest

intensity ions with multiple charge states were selected for MS/MS using rolling collision energy that was set based on the charge state and m/z value of each ions.

Additionally, one HILIC fraction at RT 19min from high-salt sample was further analyzed in parallel using Precursor Ion Discovery (PID) product mode [20–22] on NanoAcquity/Synapt HDMS platform with similar nanoLC conditions as described above. The Synapt HDMS was operated in V mode with a spray voltage of 2.6 kV and a cone voltage of 45 V. The Synapt HDMS, when operating in the survey MS mode alternates between: (1) the low (collision) energy mode (for which the gas cell collision voltage was set to 10 V), and (2) the high (collision) energy mode (with the collision voltage set to 30 V). After detection of a candidate precursor ion the instrument switches to MS/MS mode with the collision energy applied based on observed charge states and masses of precursor ion. MS to MS/MS switch criteria was dependent on the target ion intensity (10 counts/s) with 3 MS/MS scans per duty cycle using m/z 204.08 and m/z 366.13 as target ions.

To confirm the *N*-linked glycosite identification and determine the glycosylation site occupancy, high resolution nanoLC-MS/MS analysis of PNGase A treated HILIC fractions was carried out using an UltiMate3000 RSLCnano (Dionex, Sunnyvale, CA) coupled to a LTQ-Orbitrap Velos (Thermo-Fisher Scientific, San Jose, CA) mass spectrometer as described previously [10]. The same nanoLC conditions were used as mentioned above except the gradient time was reduced from 60 min to 30 min. The Orbitrap Velos was operated in parallel DDA mode using the FT mass analyzer for one survey MS scan on precursor ions followed by MS/MS scans of the 7 highest intensity multiply charged ions with a threshold ion count > 7,500 in both the LTQ mass analyzer and the high energy collision dissociation (HCD)-based FT mass analyzer at a resolution of 7,500 (fwhm at m/z 400). The instrument parameters were set exactly as reported previously [10]. All data were acquired using Xcalibur 2.1 operating software (Thermo-Fisher Scientific).

2.8 Data analysis and interpretation

All of the DDA raw data from the Synapt were output as PKL files by the ProteinLynx Global Server 2.4 (PLGS, Waters), while the raw data files from the Orbitrap were converted into MGF files using Proteome Discoverer 1.2 (PD1.2, Thermo). The subsequent searches were carried out by Mascot Daemon (version 2.3, Matrix Science, Boston, MA) against the recently completed Tomato database [23] with the official protein annotations of the SL2.40 genome build by the International Tomato Annotation Group (ITAG). The database (ITAG Release 2.3) with 34,727 entries was downloaded on 09/16/2011 from the following website: http://solgenomics.net/organism/Solanum_lycopersicum/genome. The default search settings used for the Mascot analysis were: one missed cleavage site by trypsin allowed with fixed carbamidomethyl modification of cysteine and variable modifications of methionine oxidation, deamidation of Asn and Gln residues. The peptide mass tolerance was set to 25 ppm and MS/MS mass tolerance was set to 0.1 Da for the data from the Synapt HDMS. For the Orbitrap DDA data, the peptide tolerance was set to 10 ppm and the MS/MS tolerance was set to 0.6 Da (for MS/MS spectra acquired by LTQ) and 0.05 Da (MS/MS spectra acquired by the Orbitrap FT mass analyzer). Only scores for the peptides defined by Mascot probability analysis (www.matrixscience.com/help/

[scoring_help.html#PBM](#)) greater than “identity” were considered significant for peptide identification and modification site determinations. All MS/MS spectra for the identified peptides with Asn deamidation were manually inspected and validated using both Proteome Discoverer 1.2 and Xcalibur 2.1 software (Thermo-Fisher Scientific, San Jose, CA). All acquired MS/MS spectra triggered by PI scan on m/z 204 were manually inspected and interpreted with Analyst 1.4.2 and BioAnalysis 1.4 software (Applied Biosystems) for the identification of the *N*-linked glycosylation sites and glycan structure. An in-house written script serving as an *in silico* tool for outputting the theoretical molecular weights of proteolytic peptides with additional *N*-linked consensus motif via automatically imported protein identification lists or designated database, along with the GlycoMod Tool software [24] were used to provide predictions of core peptide masses and potential glycan compositions for the glycan masses determined from MS/MS analysis. The extracted ion chromatograms (XICs) from Orbitrap DDA files for both the amide and deamidated form of all target core peptides containing *N*-linked sites were generated with Xcalibur 2.1 software with a mass tolerance at 5 ppm and a mass precision at the 4th decimal point. The ratio of XIC peak areas for deamidated form to amide form was used to determine the glycosylation site occupancy.

2.9 Prediction of subcellular location and protein function

The presence of an N-terminal ER-import signal peptide (SP) was predicted using the program SignalP 4.0 (www.cbs.dtu.dk/services/SignalP). NetGlyc (<http://www.cbs.dtu.dk/services/NetNGly/>) was used to predict glycosylation consensus sequences NX(T/S). Subcellular localization was predicted with WoLF-pSORT (<http://www.psort.org>) and Blast2GO (<http://www.blast2go.com/b2ghome>) was used to predict the functions and molecular processes in which the identified glycoproteins participate.

3 Results and discussion

3.1 Experimental design

This analytical pipeline starts with the isolation of a fraction enriched in tomato cell wall proteins using sucrose gradient centrifugation and salt fractionation (Figure 1). This is followed by lectin-based affinity chromatography to obtain a sample enriched in cell wall glycoproteins and simultaneously reduce the sample complexity. Next, a step for robust glycopeptide enrichment using ion-pairing HILIC [16, 17] was effectively deployed. We found this step to be extremely critical because the enrichment and concentration of glycopeptides helps to alleviate two of the most difficult challenges associated with large-scale investigation of glycoproteome. These are the low stoichiometry of the glycopeptides relative to non-glycosylated native peptides and the comparatively poor ionization efficiency of glycopeptides that results in glycopeptide ion suppression in MS [25, 26]. Furthermore, the HILIC fractionation represents an orthogonal separation technique to the subsequent nanoRPLC-MS analysis. It can be viewed as a conventional pre-fractionation approach, as is used in most shotgun proteomics studies to reduce the complexity of the sample entering the mass spectrometer and enhancing both the dynamic range and proteome coverage. More importantly, in this study the HILIC fractions were analyzed using a shotgun strategy to generate the protein ID list which served as the protein database for the *in silico* enzymatic

digest that facilitated the identification of glycosylation sites (as described below). Next we incorporated PI scan-driven IDA (4000 Q trap equipment) to selectively identify and structurally characterize the glycopeptides, glycosylation sites and glycoforms with the assistance of the *in silico* tool. The identified glycopeptides by PI-IDA analysis were also further validated and complemented by PID product MS/MS analysis on a high resolution Q-TOF instrument (Synapt HDMS). The percent occupancy of each glycosylation site was also determined after treatment of tryptic peptides in HILIC fractions with PNGase A and subsequent MS/MS analysis on a high resolution instrument (LTQ Orbitrap Velos) capable of distinguishing between peptides incorporating an aspartic acid (glycoform) or an asparagine (native amide). We demonstrate that this newly developed workflow can be used for large-scale empirical identification of cell wall *N*-linked glycoproteins of tomato fruit as well as other plant glycoproteomics studies.

3.2 Protein identification for cell wall low-salt and high-salt fractions

The cell wall pellet was isolated by a sucrose gradient centrifugation and the resulting proteins were fractionated with low-salt and high-salt wash. Both fractions were subjected to lectin affinity chromatography using ConA for the enrichment of glycoproteins. Figure 2A shows the UV trace of Con A chromatography. The bound fraction containing glycoproteins were collected and transferred to a more convenient buffer for digestion and lyophilization followed by gel filtration chromatography using a HiTrap desalting column (Figure 2B). The glycopeptides from both the high and low salt fractions were then digested with trypsin and the tryptic peptides isolated by SPE and dried at reduced pressure. They were then re-dissolved and subjected to HILIC to selectively isolate the glycopeptides. Our results indicate that the ion-pairing HILIC step not only effectively separates the glycopeptides from non-glycopeptides but also provides an effective prefractionation of all peptides. Initial PI-IDA analysis in 4000 Q Trap revealed that no or few glycopeptides were detected in the first 11 fractions, while the following 20 fractions were rich in glycopeptides (data not shown). Consistently, relatively higher UV absorbance was detected for the first 11 fractions than the lately eluted 20 fractions as shown in Figure 2C. This result is in excellent agreement with a previous report [16], demonstrating that 0.25% TFA can efficiently retain the hydrophilic glycopeptides in the ion-pairing HILIC.

Nevertheless, all 31 of the HILIC fractions were directly analyzed by regular MS-based DDA in Synapt HDMS equipment to obtain the initial glycoprotein identification list. A total of 192 and 69 proteins were identified from the high-salt and low-salt fractions respectively and summarized in Table 1. A total of 49 proteins were identified in both fractions with 20 distinct proteins being identified in the low-salt fraction. The detailed protein IDs are provided in Supplemental Table S1 for high-salt fraction and Table S2 for low-salt fraction. The identified proteins include polygalacturonase-inhibiting protein, expansin and pectinesterase, which are well-known cell wall proteins as reported previously [27–30] and play important functions in cell wall biology. More interestingly, further analysis of the identified proteins suggested that more than 70% of them contain secretion signal peptides, while almost 95% contain the consensus *N*-linked glycosylation sequence motif (NXS/T). Thus, the remaining ~5% of proteins likely belong to the category *O*-linked glycoproteins. These results fit well with the expected features of cell wall glycoproteins,

suggesting the isolation of cell wall proteins and lectin affinity enrichment in this study worked quite well.

A comparison of the results reported here with others from the literature is difficult to carry out because we know of no other that involve such a comprehensive analytical approach or that involve the cell wall proteome of tomato fruit as their subject of study. Perhaps the most closely related study is that of Catala et al. [14] that utilized a similar strategy for the identification of cell wall glycoproteins of tomato by coupling Con A enrichment with two-dimensional LC. Using this approach, nearly 80% of the proteins identified were predicted to be associated with the cell wall. However, despite the apparent high level of enrichment achieved, only a modest number of glycoproteins (133) were identified in the study and no information was given concerning the identity or degree of occupancy of the glycosylation sites, or the composition, structure or configuration of the glycoforms. Although the previous study used the same cultivar of tomato [14], their focus was on different developmental stage of fruit development (red ripe as opposed to mature green). Furthermore, the prefractionation strategy they utilized was significantly different than was employed here. These differences are certain to minimize the overlap in proteins identified between the two studies as well as obscure the biological implications that one might be tempted to draw from any similarities or differences observed. Nevertheless, of the 133 proteins identified in the earlier study 55 (41%) were also reported here.

More recently there has been a report of the identification of 323 *N*-linked glycosylation sites that are associated with the response to the infection of a pathogen (*Lobesia botrana*) infection in grape vine (*Vitis vinifera*) [31]. In this study the authors used a similar approach that involved a ZIC-HILIC-based enrichment of glycopeptides followed by deglycosylation analysis using PNGase A with subsequent identification of deamidated residues by LC-MS/MS analysis. This work is based on a previous report by Mann and colleagues [32]. However, there is no information concerning the structure and distribution of glycoforms or the rate of site occupancy associated with either of these studies.

It should be noted that in this study the protein ID lists were used to create a small target database (instead of the entire tomato database) facilitating the prediction of the masses of all core peptides containing the *N*-linked motif through our in-house developed script. Thus the script served as an *in silico* bioinformatic tool to significantly reduce the number of candidate peptides for discovery screening.

3.3 Identification of *N*-linked glycosylation sites and glycoforms

In this study, we used the hybrid triple quadrupole/linear ion trap feature of the 4000 Q Trap, operating in the triple quadrupole mode to perform precursor ion scans for the characteristic oxonium ion of *N*-acetylhexosamine (HexNAc) at m/z 204.1, to selectively detect glycan-containing peptide ions. MS/MS was carried out on the detected glycan containing precursor ions by rapidly switching to the linear ion trap mode for enhancing sensitivity [10, 33]. In the PI scan triggered IDA mode, all ions of significant abundance identified in the PI scan as peptidyl-glycans were sequentially selected for MS/MS fragmentation, which allows us to determine glycan composition and structural information, peptide sequence and the glycosylation site. It is often difficult to obtain reliable information

from direct MS analysis of glycopeptides due to their fragmentation behavior, even with relatively enriched glycopeptides, due to the labile nature of the glycan–peptide bond and the relatively fragile internal glycan bonds. These often cause a high level of fragmentation from glycan moieties and few fragment ions from the core peptide during traditional CID, resulting in little or ambiguous information concerning peptide sequence which exacerbates the already difficult process of determining the glycosylation site and glycan sequence. However, we took advantage of the previous finding of the formation of a distinct Y_1 ion (with the innermost GlcNAc residue attached to the peptide), which is often one of the most abundant Y type product ions observed in the glycopeptide fragmentation [33–35]. Beyond the Y_1 ion, all MS/MS spectra of glycopeptides tended to show predominately high m/z Y-type fragment ions allowing us to readily determine the charge state of the Y_1 ion acquired even with the relatively low resolution Q Trap instrument. As a result, we are able to determine the m/z of the predominant Y_1 ion along with potential Y_2 or Y_3 ions (for determining the charge state of Y_1 ion), and therefore the mass of the core peptide. Then the observed mass of peptide was searched against the predicted masses of all peptides containing the N-linked glycosylation consensus motif from the previously identified tomato cell wall glycoprotein list using our *in silico* script. The specifically matched candidate peptide sequence and its predicted y or b ion series were then compared to the existing MS/MS spectrum for confirmation of the correct assignment of peptide sequence and glycosylation site. Finally, once a peptide sequence was assigned and confirmed, the mass of the glycan can be determined. The initial glycan composition was determined by using the web-based GlycoMod Tool software [24]. The glycan sequence was then determined by manual analysis of the remaining MS/MS fragment ions. We have demonstrated this software-assisted manual interpretation approach by successfully identifying 26 distinct glycosylation sites from 24 different glycoproteins in a single HILIC fraction at retention time of 19 min (Figure 2C). The identification of the glycopeptides with their protein accession numbers, glycan sequences for each of the glycosylation sites and variable glycoforms were shown in Table 2. As expected, an overwhelming majority of identified glycopeptides revealed a typical complex type N-glycans representing a pentose (β 1–2 xylose) and/or deoxyhexose (α 1–3 fucose) linked to the core $\text{Man}_3\text{GlcNAc}_2$ or $\text{Man}_2\text{GlcNAc}_2$. Only one (IFGSLPPGLKDVPLQFFNVSYNR) out of 26 glycopeptides contained a high-mannose type N-glycan with core $\text{Man}_3\text{GlcNAc}_2$ substituted by two or three mannose residues with glycoforms: $\text{Man}_5\text{GlcNAc}_2$ and $\text{Man}_6\text{GlcNAc}_2$ (Table 2). Interestingly, in our PI-IDA analysis for the 19 min HILIC fraction, two doubly-charged precursor ions at m/z 848.48 and m/z 902.47 were also found to contain an identical complex type glycan of $\text{Xyl}_1\text{Man}_2\text{GlcNAc}_2\text{Fuc}_1$. But their MS/MS spectra matched to the peptide sequences: AVLGA $\underline{T}$ R and LTGMA $\underline{F}$ R, respectively, whose sequences have only a single Thr residue without Asn residue. Thus both glycopeptides appear to be O-linked glycosylation (Table 2).

A base peak chromatogram of a typical LC-MS precursor ion scan-triggered IDA for HILIC fraction at RT 19 min is shown in Fig. 3A, where the most intense peaks (eluting at 12.7 min, 17.5min, 20.0 min, 26.4 min and 27.8 min) were found to contain multiple glycoforms associated with 6 core peptides, respectively. Two representative MS spectra of the PI scan peak at 17.0 min and 17.8 min for HILIC fraction “RT 19 min” showed the multiple charge

states of the same glycopeptides (top panel of Figure 3B) and three different glycoforms of the same glycopeptide: YSSNISK (bottom panel of Figure 3B). These results indicate that multiple glycoforms were often detected from a single core peptide that produced the characteristic oxonium product ion. CID of the doubly-charged ions demonstrated the presence of at least three different complex type glycoforms on the YSSNISK glycopeptide belonging to an unknown protein. Fig. 4A shows an example MS/MS spectrum for the doubly-charged precursor (m/z 1114.29), which was identified to be the Xyl₁Man₃GlcNAc₂Fuc₁ glycoform at the Asn residue of the tryptic peptide FTLNGTLYK from the glycoprotein Aldose 1-epimerase. The MS/MS spectrum provides direct evidence for the complex type *N*-linked glycosylation on the FTLNGTLYK peptide.

In this study we also assessed the PID method run on a high resolution Q-TOF instrument (Synapt HDMS) to identify *N*-linked glycosylation sites for one HILIC fraction (RT 19 min) which was manually interpreted from the PI-IDA analysis in 4000 Q Trap. The PID product ion method was set up to detect the characteristic fragment ions at m/z 204.08 and m/z 366.13 from a series of survey scans. In the first survey scan with low energy applied, only precursor peptide ions were shown. When ramped to high energy in the second survey scan, the fragment ions were generated from the ions observed in the first survey scan spectrum. The high-energy survey scan spectra were examined for the presence of the predefined product ions. The MS/MS spectra were only acquired when the MS survey scan between low and high collision energies switch indicating the presence of the correct product ions. Thus, the PID mode allows for selective identification of all glycan isoforms associated with each peptide. Using this mode, we were able to validate 10 out of 26 glycopeptides identified by PI-IDA analysis for the HILIC fraction (Table 2). An example of MS/MS spectrum was shown in Figure 4B for the doubly-charged precursor (m/z 1114.47), which was identified to be the Xyl₁Man₃GlcNAc₂Fuc₁ glycoform of the tryptic peptide FTLNGTLYK, which was similar to findings in PI-IDA analysis (Figure 4A). Our results suggest that PID mode on the Q-TOF instrument (Synapt HDMS) can also selectively detect the glycopeptides in relatively complex digests. But PI-IDA scan mode in 4000 Q Trap clearly offers better sensitivity and high selectivity while PID mode appears to yield high quality spectra with better mass accuracy and high resolution for facilitating interpretation and unambiguous determination of glycan and amino acid sequences.

In the process of manual interpretation of the acquired MS/MS spectra for assigning glycopeptides, we were surprised to find several examples where an additional 57 Da was added to the N-terminus of the core peptide while both peptide forms showed exactly the same glycan patterns and similar peptide fragment patterns. Two representative MS spectra of the PI scan peak at 26.3 min and 27.1 min for the HILIC fraction at RT 19-min showed the multiply charged states of the core IFGSLPPGLKDVPLQFFNVSYNR glycopeptides (top panel of supplemental Figure S1A) and the same charge states and same glycoform on the modified peptide form: (57)IFGSLPPGLKDVPLQFFNVSYNR as shown in the bottom panel of supplemental Figure S1A. MS/MS spectra of the both native (supplemental Figure S1B) and modified forms (supplemental Figure S1C) carrying the same high mannose type glycan: Man₅GlcNAc₂ showed the exact same glycan fragmentation patterns and y series ions, but the b ion series being shifted by 57 Da for the modified form. The same

phenomena were also observed in the MS/MS spectra of the nonglycosylated equivalent peptide as shown in supplemental Figure S2A and S2B. Further investigation of our sample preparation protocol indicated that the addition of 57 Da was carboxyamidomethylation of the N-terminus due to the use of iodoacetamide as the blocking agent and the lack of a quenching step right after alkylation in the Waters' RapiGest protocol. Subsequent work has determined that the addition of a quenching step after blocking can significantly reduce the N-terminal alkylation and the use of a more specific thiol blocking agent such as methylmethanethiosulfonate can completely eliminate it. Although this unexpected modification was artificially introduced and did complicate our initial analysis, it does provide additional data points to validate our identification of glycosylation for at least three glycopeptides as shown in Table 2.

3.4 Verification of the glycosylation sites and determination of site occupancy

To further verify the glycosylation sites identified in the HILIC fraction by the PI scan-IDA analysis and to determine the glycosylation site occupancy, a portion of HILIC fraction RT 19-min was treated with PNGase A to remove glycans from the target glycopeptides. During the deglycosylation reaction, the target Asn residue was deamidated to Asp, leading to a mass increase of 0.984 Da in the deglycosylated peptide compared to its native Asn-containing peptide. The mass spectral data was acquired by Orbitrap instrument at high resolution (60,000) and accuracy (< 2 ppm) making it possible to distinguish between the amide and deaminated forms. After complete deglycosylation, all degenerate glycoforms of a particular core peptide will collapse into the acidic form significantly enhancing sensitivity. As a result, the identification of the peptides with deaminated Asn located at *N*-linked consensus motifs has been often used as an empirical evidence confirming the identity of *N*-linked glycosylation sites [10, 36]. Another benefit of the deglycosylation treatment prior to nanoLC-MS/MS analysis is that the relative quantitation of the deamidated form versus the amide form for the target peptide can be obtained through the XIC of its precursor *m/z* at < 5 ppm mass tolerance. Although the difference of Asp and Asn may cause slight variation in ionization efficiency and retention times, these differences are usually very small and inconsequential in practice. Therefore, the quantitation of the amide form (Asn) and deamidated form (Asp) can be used to determine the glycosylation site occupancy as $[\text{Asp}] / \{[\text{Asp}] + [\text{Asn}]\}$.

Our deglycosylation results showed that 12 out of 26 glycosylation sites identified in the HILIC fraction RT 19min by PI-IDA analysis using the 4000 Q Trap were confirmed with the deamidated form (Asp) present in the equivalent core peptides. Furthermore, 5 additional tryptic peptides with deamidated form, which fit the *N*-linked sequence motif were also identified (Table 2). Interestingly, further analysis of the glycopeptides which failed to be identified in this deglycosylation experiment by PNGase A revealed that almost all those glycopeptides were either too short or too long. For those very short peptides, they will not be detected given the fact that their precursor *m/z* at multiply-charged states are likely to be under the *m/z* 400 threshold to be selected for the regular DDA analysis. In our experiment, we failed to detect any long peptides containing >15 residues. Our results confirmed previous report that PNGase A is only active on and releases glycans from relatively short glycopeptides [11]. Glycosylation site occupancy was also determined for all 17 identified

peptides (Table 2). A majority of those peptides showed a high occupancy rate with more than 95% occupancy. A few sites were found as low as 35% occupancy. It is surprising that 4 out of 5 additionally identified glycopeptides revealed by deamidation analysis showed more than a 90% occupancy rate, but failed to be identified by the PI-IDA analysis (4000 Q Trap). However, given the fact that the digestion with PNGase A was performed at 37 °C for 16 hrs in 100 mM sodium citrate/sodium phosphate monobasic pH = 5.0, it is possible that there might have been some contributions to the apparent high occupancy rate from spontaneous nonspecific deamidation at the target Asn residues in some of those peptides.

3.5 Functional classification of cell wall glycoproteins

The subcellular locations, functional classifications and biological processes that involve the glycoproteins identified are displayed in Figure 5 A–C. The subcellular location of each identified glycoprotein was predicted by WoLF-pSORT and is shown in Fig. 5A. The locations predicted are fully consistent with the information provided by SignalP, as the predicted locations suggest that $\approx 70\%$ of the proteins contain secretion signals. However, proteins that contain signal peptides and that enter the secretory pathway are not necessarily targeted to the cell wall, but can be retained within, or redirected to various compartments of the endomembrane system such as the ER or Golgi, or to other organelles including the vacuole and chloroplast. WoLF-pSORT uses amino acid sequence homology-based location prediction algorithms [37] to correctly predict the subcellular location of proteins. As with most location prediction algorithms, it does not incorporate information concerning glycosylation into its prediction strategy. Thus, to the extent that glycosylation represents a “protein sorting” signal, or is otherwise involved in the sorting process, all such prediction schemes are blind to the information that glycosylation can provide. It has long been hypothesized that *N*-glycans act as protein sorting signals [38, 39], however the hypothesis remains controversial primarily because the mechanisms through which the sorting information is “programmed” into the glycans and “understood” by the cell’s sorting machinery has yet to be described. Furthermore, *N*-glycans are also important for various processes that occur before or after the sorting event, such as protein folding, ER-associated degradation, ER-to-Golgi trafficking, receptor recognition etc. Thus, simply investigating the effect of changing the number of glycosylation sites or the nature of *N*-linked glycans at a particular site on the abundance of the glycoprotein at its target location does not allow one to distinguish between effects on these other processes and the “protein sorting”. However, the putative connection between glycosylation state and protein sorting remains a viable hypothesis and there is a growing body of experimental evidence to support it [38–40].

Figures 5B and 5C show the number of GO Term identified that are predicted to be involved with particular biological functions (B) and molecular processes (C) by the Blast2Go program [41, 42]. These results are consistent with the activities associated with the subcellular locations identified above and suggest that many of the proteins identified are likely involved in multiple biological functions and processes. While the multi-functionality of proteins is a well-accepted concept it does complicate our interpretation of proteomic data as the generality of the information generated is such that it is often useless in trying to assign biological significance to an increase or decrease in a particular protein’s abundance.

As with the localization prediction schemes discussed above, Blast2Go relies largely on amino acid sequence homology to proteins of known function to predict the activity of an unknown protein and does not consider the occupation of putative glycosylation sites or the nature of the glycans involved. Glycosylation is known to affect receptor recognition, solubility, structure and activity of proteins and thus programs that do not involve information concerning glycosylation can only return a family of activities for a particular amino acid sequence; whereas, to develop an understanding of a biological function or process requires the assignment of specific activities to a defined molecular species. Knowledge of the glycosylation status of a protein can provide the code to resolve this functional degeneracy such that we can assign a specific activity/function (or a limited set of activities/functions) to a chemically defined molecule. Sadly, since the data available concerning the glycosylation status of most proteins is so limited (nonexistent in most cases) this goal will remain unachieved for some time. Nevertheless, the workflow and associated procedures presented herein provide an opportunity to begin the hard work of generating a library of glycoforms that can be integrated into the annotation of genomic and protein databases that will one day be used to improve predictions of protein localization as well as those concerning biological process and function.

4 Concluding remarks

In this study, we reported a workflow combining selective enrichment of glycoproteins/peptides by lectin affinity, tryptic digestion and ion-pairing HILIC, and precursor ion-triggered data dependent MS/MS analysis with a bioinformatic script to facilitate the identification and characterization of *N*-linked glycosylation sites of secreted proteins from tomato fruit. The workflow also covers precursor ion discovery (PID)-based MS/MS and PNGase A digestion of HILIC fractions followed by high accuracy and resolution MS to confirm the glycosylation sites and determine occupancy rates. Using the tomato cell wall protein fractions as a test case, we have demonstrated that this newly developed workflow can be used for large-scale empirical identification of *N*-linked glycoproteins of tomato fruit secreted proteins. Using the workflow described above it is possible to identify and characterize large numbers of glycopeptides in terms of both the primary structure of their peptidyl and glycan moieties in other plant glycoproteomics studies. This represents a necessary (although insufficient) first step in the full characterization of glycoproteins which will improve predictions involving protein localization and biological activities and will allow for a more specific and useful interpretation of changes in protein abundance. Ultimately, the information generated using the procedures described herein will have to be integrated at the protein level because the glycoprotein represents the biologically active species and they often contain multiple glycosylation sites which vary independently with respect to their degree of occupancy and the nature of the glycans (glycoforms) present at each site. Modification to any or all of these sites can have important consequences for the protein in terms of activity, site of action, protein stability, and solubility. Thus, future extensions of these efforts will require the adoption of a “top down” as opposed to the “bottom up” approach used here. However, it will be extremely difficult to adapt traditional top down approaches into high throughput strategies to fully characterize the glycoproteome of a particular organism in real time. Fortunately Kelleher and colleagues have recently

introduced a high throughput top down strategy [43] that could be adapted to this challenge, although improvements in both sensitivity and compatible mass range will be required for such an approach to be practical. Nevertheless, the bottom up approach reported here can provide information involving glycopeptide and glycan structures that can ultimately be used to select glycoproteins for targeted characterization or facilitate the interpretation of the high throughput top down work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank Professor Jocelyn Rose and Dr. Michelle Cilia for their thoughtful insight and useful discussions concerning this work and Mr. Simon Hucko for his expert technical assistance. This work was partially supported by NIH SIG grant 1S10RR025449-01, ARS CRIS projects 1907-21000-033/034-00D and NSF Plant Genome Grant DBI-0606596.

Abbreviations

PI	precursor ion
PID	precursor ion discovery
Con A	Concanavalin A
IDA	information dependent acquisition
DDA	data-dependent acquisition
HCD	high energy collision dissociation
XIC	extracted ion chromatogram
HexNAc	<i>N</i> -acetylhexosamine.

References

1. Apweiler R, Hermjakob H, Sharon N. *Biochim Biophys Acta*. 1999; 1473:4–8. [PubMed: 10580125]
2. Budnik BA, Lee RS, Steen JA. *Biochim Biophys Acta*. 2006; 1764:1870–1880. [PubMed: 17118724]
3. An HJ, Kronewitter SR, de Leoz ML, Lebrilla CB. *Curr Opin Chem Biol*. 2009; 13:601–607. [PubMed: 19775929]
4. Pan S, Chen R, Aebersold R, Brentnall TA. *Molecular & Cellular Proteomics : MCP*. 2011; 10:1–14. R110 003251.
5. Rose JK, Lee SJ. *Plant Physiol*. 2010; 153:433–436. [PubMed: 20237018]
6. Fitchette AC, Dinh OT, Faye L, Bardor M. *Methods Mol Biol*. 2007; 355:317–342. [PubMed: 17093320]
7. Ruiz-May E, Thannhauser TW, Zhang S, Rose JK. *Front Plant Sci*. 2012; 3:150. [PubMed: 22783270]
8. Song W, Henquet MG, Mentink RA, van Dijk AJ, Cordewener JH, Bosch D, America AH, van der Krol AR. *J Proteomics*. 2011; 74:1463–1474. [PubMed: 21605711]

9. Bardor M, Cabrera G, Stadlmann J, Lerouge P, Cremata JA, Gomord V, Fitchette AC. *Methods Mol Biol.* 2009; 483:239–264. [PubMed: 19183903]
10. Zhang S, Sherwood RW, Yang Y, Fish T, Chen W, McCardle JA, Jones RM, Yusibov V, May ER, Rose JK, Thannhauser TW. *Proteomics.* 2012; 12:1269–1288. [PubMed: 22577028]
11. Ytterberg AJ, Jensen ON. *J Proteomics.* 2010; 73:2249–2266. [PubMed: 20541636]
12. Zhang, S.; Yang, Y.; Thannhauser, TW. *Gene Discovery for Disease Models.* Gu, WaWY., editor. Hoboken, NJ: John Wiley and Sons, Inc.; 2011. p. 471-530.
13. Minic Z, Jamet E, Negroni L, Arsene der Garabedian P, Zivy M, Jouanin L. *J Exp Bot.* 2007; 58:2503–2512. [PubMed: 17526915]
14. Catala C, Howe KJ, Hucko S, Rose JK, Thannhauser TW. *Proteomics.* 2011; 11:1530–1544. [PubMed: 21381198]
15. Gonzalez-Bosch C, Brummell DA, Bennett AB. *Plant Physiol.* 1996; 111:1313–1319. [PubMed: 12226364]
16. Ding W, Nothhaft H, Szymanski CM, Kelly J. *Molecular & Cellular Proteomics : MCP.* 2009; 8:2170–2185. [PubMed: 19525481]
17. Mysling S, Palmisano G, Hojrup P, Thaysen-Andersen M. *Anal Chem.* 2010; 82:5598–5609. [PubMed: 20536156]
18. Cilia M, Howe K, Fish T, Smith D, Mahoney J, Tamborindéguy C, Burd J, Thannhauser TW, Gray S. *Proteomics.* 2011; 11:2440–2458. [PubMed: 21648087]
19. Yang Y, Qiang X, Owsiany K, Zhang S, Thannhauser TW, Li L. *J Proteome Res.* 2011; 10:4647–4660. [PubMed: 21842911]
20. Chambery A, Di Maro A, Parente A. *Phytochemistry.* 2008; 69:1973–1982. [PubMed: 18514239]
21. Chambery A, Severino V, D'Aniello A, Parente A. *Analytical biochemistry.* 2008; 374:335–345. [PubMed: 18068660]
22. Zhou Y, Huang SX, Song JZ, Qiao CF, Li SL, Han QB, Xu HX. *J Am Soc Mass Spectrom.* 2009; 20:1846–1850. [PubMed: 19643632]
23. Consortium TTG. *Nature.* 2012; 485:635–641. [PubMed: 22660326]
24. Cooper CA, Gasteiger E, Packer NH. *Proteomics.* 2001; 1:340–349. [PubMed: 11680880]
25. Dell A, Morris HR. *Science.* 2001; 291:2351–2356. [PubMed: 11269315]
26. Peterman SM, Mulholland JJ. *J Am Soc Mass Spectrom.* 2006; 17:168–179. [PubMed: 16406561]
27. De-la-Pena C, Badri DV, Vivanco JM. *Biochim Biophys Acta.* 2008; 1780:773–783. [PubMed: 18222123]
28. Kirsch R, Wielsch N, Vogel H, Svato A, Heckel DG, Pauchet Y. *BMC Genomics.* 2012; 13:587. [PubMed: 23116131]
29. Maulik A, Sarkar AI, Devi S, Basu S. *Plant Biol (Stuttg).* 2011; 14(Suppl 1):22–30. [PubMed: 22039764]
30. Vannerum K, Huysman MJ, De Rycke R, Vuylsteke M, Leliaert F, Pollier J, Lutz-Meindl U, Gillard J, De Veylder L, Goossens A, Inze D, Vyverman W. *BMC Plant Biol.* 2011; 11:128. [PubMed: 21943227]
31. Melo-Braga MN, Verano-Braga T, Leon IR, Antonacci D, Nogueira FC, Thelen JJ, Larsen MR, Palmisano G. *Molecular & Cellular Proteomics : MCP.* 2012; 11:945–956. [PubMed: 22778145]
32. Zielinska DF, Gnäd F, Wisniewski JR, Mann M. *Cell.* 2010; 141:897–907. [PubMed: 20510933]
33. Zhang S, Williamson BL. *J Biomol Tech.* 2005; 16:209–219. [PubMed: 16461944]
34. Ritchie MA, Gill AC, Deery MJ, Lilley K. *J Am Soc Mass Spectrom.* 2002; 13:1065–1077. [PubMed: 12322954]
35. Segu ZM, Mechref Y. *Rapid Commun Mass Spectrom.* 2010; 24:1217–1225. [PubMed: 20391591]
36. Zhao J, Liu YH, Reichert P, Pflanz S, Pramanik B. *J Mass Spectrom.* 2010; 45:1416–1425. [PubMed: 21053369]
37. Horton P, Park KJ, Obayashi T, Nakai Kenta. *Proceedings of Asian Pacific Bioinformatics Conference.* 2006; APBC06:39–48.
38. Fiedler K, Simons K. *Cell.* 1995; 81:309–312. [PubMed: 7736583]

39. Gut A, Kappeler F, Hyka N, Balda MS, Hauri HP, Matter K. EMBO J. 1998; 17:1919–1929. [PubMed: 9524115]
40. Vagin O, Kraut JA, Sachs G. Am J Physiol Renal Physiol. 2009; 296:F459–F469. [PubMed: 18971212]
41. Conesa A, Gotz S. International journal of plant genomics. 2008; 2008:619832. [PubMed: 18483572]
42. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A. Nucleic acids research. 2008; 36:3420–3435. [PubMed: 18445632]
43. Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, Tipton JD, Vellaichamy A, Kellie JF, Li M, Wu C, Sweet SM, Early BP, Siuti N, LeDuc RD, Compton PD, Thomas PM, Kelleher NL. Nature. 2011; 480:254–258. [PubMed: 22037311]

Figure 1.

Experimental design and schematic diagram of workflow used in this study. Three nanoLC-MS/MS techniques were used to obtain N-linked glycosylation profiles of the tomato cell wall glycoproteins. These techniques include PI scan-triggered MS/MS analysis in 4000 Q Trap, PID in Synapt HDMS and high mass resolution data dependent MS/MS analysis in Orbitrap Velos for verification of the identified glycosylation sites and determination of glycosylation site occupancy. High-salt sample was used as a test case for the above workflow.

A

B

C

Figure 2.

Enrichment of glycoproteins and glycopeptides. Tomato cell wall fractions were initially enriched by the lectin Con A for enrichment of glycoproteins (A), followed by HiTrap gel filtration desalting (B), trypsin digestion and finally the glycopeptides were enriched by HILIC analysis (C). All HILIC fractions starting from RT 2 min to RT 41 min were analyzed by PI-IDA analysis. No or few glycopeptides were detected in the first 11 fractions indicated by rightwards arrows, while rich glycopeptides were found in the following 20 fractions indicated by left right arrows. The MS/MS spectra for characterization of glycopeptides and glycans in fraction RT 19 min was interpreted manually and reported in this study.

A

MS Spectra of NanoLC-PI Scan (+204) at 17.0 and 17.8-min Elution Time

B

Figure 3.

Precursor ion scan triggered information dependent MS/MS analysis (PI-IDA) was used for characterizing *N*-linked glycosylation of HILIC fractions. A representative base peak chromatogram of a LC-MS precursor ion scan (on *m/z* 204.08) triggered IDA for HILIC fraction RT 19 min (A) where a few of *N*-linked glycosylation sites identified in this study are labeled. Two representatives of MS spectrum acquired from a peak identified by the LC-MS precursor ion scan (+204.08) triggered IDA analysis eluted at 17.0 min and 17.8 min (B). Several different *N*-linked glycosylation sites along with various types of glycoforms associated with glycosylation site were detected and characterized by the subsequent MS/MS on each of the glycoform ions.

B

Figure 4.

MS/MS spectra used for identifying *N*-linked glycopeptides, defining the glycosylation sites and elucidation of the glycan structure by the selective PI scan triggered IDA and PID product MS/MS analysis. An MS/MS Spectrum of m/z 1114.29²⁺ ion identifying the glycopeptide (FTLNGTLYK) from a protein called Aldose 1-epimerase by the selective PI scan triggered IDA approach using a 4000 Q Trap (A), and by PID product MS/MS in Synapt HDMS (B) as the glycoform of (Pent₁Hex₃GlcNAc₂Fuc₁). An MS/MS spectrum of m/z 529.28²⁺ acquired using an LTQ Orbitrap for the HILIC fraction treated with PNGase A confirming the presence of the expected deaminated peptide: FTLDGTTYK (C).

A

B

C

Figure 5.

Predicted subcellular locations and functional classification of proteins identified from both high-salt and low-salt fractions. Subcellular location predictions were determined by WoLF-pSORT (A), Chlo: Chloroplast; Cysk: cytoskeleton; Cyto: cytoplasm; ER: endoplasmic reticulum; NP: no prediction; Nuc: nuclear; Plas: plasma membrane; Vac: vacuole. The protein functions were assigned using literature references or predicted by homology to other proteins using Blast2GO program. The functional categories of all identified proteins were shown as Level 2 biological function (A) and biological process (B).

Table 1 Summary of protein identifications^a

Fractions	Number of protein IDs	SignalP	N-Glycosylation motifs
Low-salt wash	69	77% (53/69)	94% (65/69)
High-salt wash	192	75% (145/192)	91% (175/192)

^a see Tables S1–2 for details.

Table 2
Summary of *N*-linked sites and their glycoforms identified in HILIC fraction RT19 min

RT (min)	m/z	Z	M+H	Pep+H	Mass of glycan	Glycan composition	Pep Sequence ^c	Accession Number	ID & % Occupancy in PNGase A digests
11.45	787.42	2	1573.83	549.24	1024.59	[(Hex)3 (HexNAc)2 (Pent)1]	SSGGNK(t)	Solyc07g017520.2.1	ND
11.76	921.50	2	1841.99	833.40	1008.59	[(Hex)2(NAcHex)2(Deoxyhexose)1(Pent)1]	IFP^NGTGK^a	Solyc03g119080.2.1	35%
12.21	855.46	2	1709.92	833.28	876.64	[(Hex)2 (HexNAc)2 (Pent)1]			
12.05	848.48	2	1695.95	687.36	1008.59	[(Hex)2 (HexNAc)2 (Deoxyhexose)1 (Pent)1]	AVLGAT^R ^b	Solyc02g064940.1.1	ND
12.39	828.43	2	1655.86	647.16	1008.70	[(Hex)2 (HexNAc)2 (Deoxyhexose)1 (Pent)1]	INWSK	Solyc02g091920.2.1	ND
12.68	912.65	2	1824.30	815.4	1008.90	[(Hex)2 (HexNAc)2 (Deoxyhexose)1 (Pent)1]	QILNVTK^a	Solyc08g067030.2.1	>95%
12.68	941.10	2	1881.30	872.4	1008.90	[(Hex)2 (HexNAc)2 (Deoxyhexose)1 (Pent)1]	[57]QILNVTK^a		
12.86	902.00	2	1803.01	794.4	1008.61	[(Hex)2 (HexNAc)2 (Deoxyhexose)1 (Pent)1]	TFANLTK^a	Solyc10g052880.1.1	ND
13.80	899.95	2	1798.91	774.36	1024.55	[(Hex)3 (HexNAc)2 (Pent)1]	VTWVNR(t)	Solyc09g018790.2.1	ND
13.91	994.01	2	1987.03	978.24	1008.79	[(Hex)2 (HexNAc)2 (Deoxyhexose)1 (Pent)1]	SPNFSGTLR^a	Solyc01g096450.2.1	>95%
13.96	793.08	3	2377.24	1368.48	1008.76	[(Hex)2 (HexNAc)2 (Deoxyhexose)1 (Pent)1]	VPSHTISNPFNR(t)^a	Solyc01g074030.2.1	>95%
14.54	1002.03	2	2003.06	994.2	1008.86	[(Hex)2 (HexNAc)2 (Deoxyhexose)1 (Pent)1]	QSYNLTIR	Solyc01g087800.2.1	34%
16.20	1158.14	2	2315.28	1144.44	1170.84	[(Hex)3 (HexNAc)2 (Deoxyhexose)1 (Pent)1]	YNTTLTYLR	Solyc07g062480.1.1	>95%
17.25	1205.71	2	2410.41	1239.48	1170.93	[(Hex)3 (HexNAc)2 (Deoxyhexose)1 (Pent)1]	NLTGPIPTIVK	Solyc09g014480.1.1	ND
17.35	1114.28	2	2227.35	1056.6	1170.75	[(Hex)3 (HexNAc)2 (Deoxyhexose)1 (Pent)1]	FTLNGTLYK^a	Solyc02g087790.1.1	>95%
17.64	903.96	2	1806.92	798.48	1008.44	[(Hex)1 (HexNAc)2 (Deoxyhexose)1 (Pent)1]	SLHNLSK^a	Solyc08g048120.1.1	>95%

RT (min)	m/z	Z	M+H	Pep+H	Mass of glycan	Glycan compisition	Pep Sequence ^c	Accession Number	ID & % Occupancy in PNGase A digests
17.90	822.96	2	1644.93	798.36	846.57	[(Hex)2 (HexNAc)2 (Deoxyhexose)1 (Pent)1]1 [(Hex)2 (HexNAc)2 (Deoxyhexose)1 [(Hex)5 (HexNAc)2]			
17.90	756.90	2	1512.80	798.40	714.40				
18.21	1066.08	2	2131.16	914.4	1216.76		INILNLSK	Solyc11g056680.1.1	>95%
19.15	1015.23	3	3043.68	1670.06	1373.62	[(HexNAc)(Hex)3 (HexNAc)2 (Deoxyhexose)1 (Pent)1]1	LTAFNHSSLLFEYK	Solyc08g083250.2.1	>95%
19.31	885.22	3	2653.65	1482.6	1171.05	[(Hex)3 (HexNAc)2 (Deoxyhexose)1 (Pent)1]1	IGGAQFTLNGTLYK	Solyc02g087780.2.1	>95%
19.49	953.18	3	2857.53	1686.44	1171.09	[(Hex)3 (HexNAc)2 (Deoxyhexose)1 (Pent)1]1	AALNESHLGIFNSWK	Solyc11g056680.1.1	ND
19.60	902.47	2	1803.95	795.36	1008.59	[(Hex)2 (HexNAc)2 (Deoxyhexose)1 (Pent)1]1	LIGMAFR a,b	Solyc10g005510.2.1	ND
20.07	875.56	3	2624.67	1615.58	1009.09	[(Hex)2 (HexNAc)2 (Deoxyhexose)1 (Pent)1]1	SLTNLSSLMoxHLDLR a	Solyc11g056680.1.1	>95%
20.38	894.49	3	2681.48	1672.56	1008.92	[(Hex)2 (HexNAc)2 (Deoxyhexose)1 (Pent)1]1	(57)SLTNLSSLMoxHLDLR		
21.56	1021.45	3	3062.34	1687.44	1374.90	[(HexNAc)(Hex)3 (HexNAc)2 (Deoxyhexose)1 (Pent)1]1	FGWTNSVNTNWAK	Solyc02g088390.2.1	ND
21.58	1089.11	3	3265.33	1687.2	1578.13	[(HexNAc)2(Hex)3 (HexNAc)2 (Deoxyhexose)1 (Pent)1]1			
22.50	985.53	3	2954.58	1945.8	1008.78	[(Hex)2 (HexNAc)2 (Deoxyhexose)1 (Pent)1]1	APINSTISISQTLVFEK	Solyc10g084320.1.1	ND
22.92	1117.32	3	3349.97	2179.22	1170.75	[(Hex)3 (HexNAc)2 (Deoxyhexose)1 (Pent)1]1	HFGLLYPNGTNIYGIDLSGK	Solyc09g057630.2.1	ND
26.34	997.81	4	3988.24	2608.76	1379.48	(Hex) ₆ (HexNAc) ₂	IFGSLPPGLKDVPLQFFNVSYNR	Solyc07g065090.1.1	ND
26.63	1329.80	3	3987.40	2608.58	1378.82	(Hex) ₆ (HexNAc) ₂			
26.73	957.26	4	3826.02	2608.51	1217.51	(Hex) ₅ (HexNAc) ₂			
26.76	1276.11	3	3826.34	2608.32	1218.02	(Hex) ₅ (HexNAc) ₂			
27.02	971.52	4	3883.08	2665.58	1217.50	(Hex) ₅ (HexNAc) ₂	(57)IFGSLPPGLKDVPLQFFNVSYNR		
27.05	1294.76	3	3882.28	2665.58	1216.70	(Hex) ₅ (HexNAc) ₂			

RT (min)	m/z	Z	M+H	Pep+H	Mass of glycan	Glycan compisition	Pep Sequence ^c	Accession Number	ID & % Occupancy in PNGase A digests
27.49	1189.99	3	3567.98	2380.58	1187.40	[(Hex)4 (HexNAc)2 (Pent)1]	VFVSTPLTVLTCIPIDAT <u>NR</u> (t)	Solyc07g006570.2.1	ND
27.55	1184.61	3	3551.82	2380.58	1171.24	[(Hex)3 (HexNAc)2 (Deoxyhexose)1 (Pent)1]			
27.71	889.01	4	3553.04	2380.58	1172.46	[(Hex)3 (HexNAc)2 (Deoxyhexose)1 (Pent)1]			
27.81	1140.39	3	3419.17	2380.58	1038.59	(Hex) ₃ (HexNAc) ₂ (Deoxyhexose) ₁			
27.87	1146.09	3	3436.27	2380.58	1055.69	(Hex) ₄ (HexNAc) ₂			
28.13	855.66	4	3419.64	2380.58	1039.06	(Hex) ₃ (HexNAc) ₂ (Deoxyhexose) ₁			
31.26	1194.42	3	3581.25	2410.46	1170.79	[(Hex)3 (HexNAc)2 (Deoxyhexose)1 (Pent)1]	YDSISLPALL <u>NT</u> ASAPVNLPLK	Solyc08g080190.2.1	ND
Additional deamidation sites found in PNGase A digest of HIL-IC fraction RT19 min									
	610.30	2		1219.59			NQMoxNVLE <u>LN</u> K(t)	Solyc02g088390.2.1	90%
	710.85	2		1420.70			IVFT <u>NG</u> SQDPWR	Solyc03g033620.2.1	84%
	530.80	2		1060.61			LLEQ <u>IN</u> TTK	Solyc02g078950.2.1	>95%
	789.36	2		1577.72			TVCGGAG <u>IN</u> ACSNPAK	Solyc12g017460.1.1	>95%
	510.28	2		1019.54			IFPSA <u>NI</u> TR	Solyc01g087610.2.1	34%

^a glycopeptides were identified and confirmed by PID-MS/MS analysis as well;

^b peptides with potential *O*-linked glycosylation were italic;

^c glycosylated residues were underlined.