RESEARCH ARTICLE

Shotgun proteome profile of *Populus* developing xylem

Udaya C. Kalluri[1], *Gregory B. Hurst*[2], *Patricia K. Lankford*[3], *Priya Ranjan*[1] *and Dale A. Pelletier*[3]

[1] Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
[2] Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
[3] Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Understanding the molecular pathways of plant cell wall biosynthesis and remodeling is central to interpreting biological mechanisms underlying plant growth and adaptation as well as leveraging that knowledge towards development of improved bioenergy feedstocks. Here, we report the application of shotgun MS/MS profiling to the proteome of *Populus* developing xylem. Nearly 6000 different proteins were identified from the xylem proteome. To identify low-abundance DNA-regulatory proteins from the developing xylem, a selective nuclear proteome profiling method was developed. Several putative transcription factors and chromatin remodeling proteins were identified using this method, such as NAC domain, CtCP-like and CHB3-SWI/SNF-related proteins. Public databases were mined to obtain information in support of subcellular localization, transcript-level expression and functional categorization of identified proteins. In addition to finding protein-level evidence of candidate cell wall biosynthesis genes from xylem (wood) tissue such as cellulose synthase, sucrose synthase and polygalacturonase, several other potentially new candidate genes in the cell wall biosynthesis pathway were discovered. Further application of such proteomics methods will aid in plant systems biology modeling efforts by enhancing the understanding not only of cell wall biosynthesis but also of other plant developmental and physiological pathways.

Received: November 1, 2008
Revised: July 25, 2009
Accepted: July 27, 2009

Keywords:
LC-MS/MS / Plant proteomics / *Populus* / Xylem

1 Introduction

The release of the *Populus trichocarpa* (Torr. & Gray) genome sequence and the development of supportive reverse genetics, population genetics and transcript profiling capabilities have facilitated gene sequence–based research in an unprecedented manner [1]. Currently, widespread efforts are being undertaken by the research community to improve woody biomass–based biofuel production. Central to the improvement of feedstock properties is a comprehensive understanding of wood development and cell wall biosynthesis properties. Wood or dead secondary xylem in tree trunks is formed as a result of division, enlargement, differentiation, maturation and programmed cell death of xylem cells that are perennially produced by the vascular cambium. Moreover, plant cell walls, which are significant to development and morphology, water and solute transport mechanisms, disease resistance and strength of a plant as a whole, are predominantly of the secondary-walled type in xylem tissue [2]. Therefore, developing xylem cells serve as useful models to investigate secondary cell wall formation. Several studies have reported single-gene to transcriptome-level changes during xylem development or progression to secondary cell wall formation [3–7]. These studies are based on EST, microarray or RT-PCR transcript-level gene expression data, which are valuable indicators but are not entirely representative of gene product activity. Protein-level measurements are significant in deciphering the post-translational abundance, regulatory status and subcellular localization of the gene product and hence in suggesting its

Correspondence: Dr. Udaya C. Kalluri, Staff Scientist, Environmental Sciences Division, Oak Ridge National Laboratory, P. O. Box 2008, Oak Ridge, TN 37831, USA
E-mail: kalluriudayc@ornl.gov
Fax: +1-865-576-9939

Abbreviations: MudPIT, multidimensional protein identification technology; **NSAF,** normalized spectral abundance factor activity. However, protein-level information for *Populus* has thus far been primarily derived using 2-D PAGE approaches [8–11]. The ability to conduct proteome measurements in a high-throughput manner will be highly valuable in expediting research investigations as well as in making them more comprehensive.

HPLC interfaced with MS/MS offers a general and automated approach for identifying large numbers of individual protein components of the proteome. In the Multidimensional Protein Identification Technology (MudPIT) implementation of this approach [12, 13], the entire protein complement of a tissue, cell or subcellular compartment is enzymatically digested. The resulting peptides are analyzed using automated 2-D chromatographic separation (strong cation exchange followed by reverse phase) interfaced *via* electrospray with a mass spectrometer. MS/MS analysis of peptides eluting from the 2-D separation provides partial amino acid sequence information, which allows software-based identification of a peptide and further its assembly into protein identifications. The MudPIT approach has been previously applied to proteomics of rice [14] as well as to the study of ubiquitinated proteins in *Arabadopsis thaliana* [15]. To our knowledge, MudPIT has not been applied to proteomics studies of *Populus*.

Central to making sense of the large genome data sets and their predicted functions is the ability to put the proteome in a cell biology context. It is therefore highly desirable to be able to apply the shotgun profiling approaches to subcellular fractions. Here, we report the application of MudPIT to shotgun proteome profiling of *Populus*, a model bioenergy crop. Using developing xylem from *Populus* stems, we have applied subcellular fractionation techniques to obtain crude, pellet and nuclear protein fractions. LC-MS/MS analysis was followed with bioinformatics-based analysis of functional annotation, gene duplication, predicted subcellular localization and transcript-level expression support.

2 Materials and methods

2.1 Plant materials

Wild-type *Populus* plants, hybrid aspen clone 717 (*Populus tremula x alba*), were vegetatively propagated in soilless mix of perlite:peat (2:1) and grown for a minimum of 6 months under greenhouse conditions of ambient humidity and controlled light (high pressure sodium lamps, 16 h photoperiod), fertilization (weekly treatment of Jack's professional water soluble 20:10:20 fertilizer with micronutrients), temperature ($72\pm2°F$) and automated drip irrigation. Green stem tops representing younger internodes were excised and removed and the woody portion of the stem was used to generate developing xylem sample. Xylem tissue was collected by peeling off the bark, scraping the juicy (developing tissue) outer layers of exposed stems and flash-freezing the material in liquid nitrogen. Such a tissue sample is expected to be composed of cell types from various phases of secondary xylem development, including expanding as well as secondary cell wall forming xylem vessels and fibers (predominant tissue component) [4, 5]. Xylem tissue pooled from several plants was used in generating two batches of samples for proteome characterization.

2.2 Protein extraction and quantification

Xylem tissue was ground under liquid nitrogen using a mortar and pestle. A 3 g sample of the ground tissue was suspended in 15 mL lysis buffer containing 125 mM Tris (pH 8.5), 10% glycerol, 50 mM DTT and 1 mM EDTA [16].

The suspension was vortexed twice for 30 s each time, then sonicated (Branson 185 sonifier, power setting of 40) on ice for three rounds of 30 s each. Large debris was removed from the highly viscous sample by centrifugation for 6 min at 1200 × *g*. The supernatant was again centrifuged for 10 min at 12 000 × *g*, and the pellet discarded. A final centrifugation step at 100 000 × *g* for 1 h yielded a crude soluble protein fraction (crude/soluble fraction) and a pellet (pellet fraction). Protein determination using Lowry's method [17] indicated 135 mg total protein in the crude soluble fraction and 30 mg total protein in the pellet.

2.3 Nuclei isolation

Nuclei were extracted from xylem tissue using a CelLytic PN Isolation/Extraction Kit following the manufacturer's protocol (Sigma, St. Louis, MO, USA). Briefly, a 3.4 g sample of the ground xylem tissue was briefly suspended in 10 mL of 1 × nuclei isolation buffer containing 1 mM DTT; the sample was then gently agitated for 10 min at 4°C. The suspension was passed through filter mesh to remove debris into a 50 mL tube. The suspension was centrifuged at 1000 × *g* in tabletop centrifuge for 10 min. The supernatant was decanted. The resulting pellet was gently resuspended in 1 mL of nuclei isolation buffer and lysed by addition of 0.3% Triton X-100. About 2 mL of cell lysate was obtained. Aliquots of 600 uL of lysed cell material was applied to a 0.8 mL cushion of 1.5 M sucrose in a 1.5 mL centrifuge tube and centrifuged for 10 min at 12 000 × *g*. The upper green phase was aspirated leaving the nuclei pellets. The pellets were washed twice with ~0.5 mL of buffer, pelleted by centrifugation and resuspended in 150 μL of extraction buffer with 5 mM DTT and then vortexed at 4°C for 30 min. The sample was then centrifuged at 12 000 × *g* for 10 min yielding 150 μL of soluble extracted nuclei (nuclear fraction). Protein determination using Lowry's method [17] indicated 300 μg total protein.

2.4 Immunoblotting

Aliquots of total protein extracted from xylem cells and enriched nuclei were separated on a 4–20% Precision Protein acrylamide gel (Pierce, Rockville, IL, USA) and transferred to PVDF using the Invitrogen iBlot system (Invitrogen, Carlsbad, CA, USA). Membranes were washed overnight at 4°C in PBS containing 0.1% Tween 20 (PBS-Tween). Non-specific binding was blocked by incubation of the membrane with gentle rocking for 0.5 hour in 5% non-fat milk powder, 5% BSA in PBS-Tween ("Blotto"). The membrane was then incubated for 2 h with rocking at room temperature with rabbit anti-Histone H3 (Sigma). After extensive washing in PBS-Tween, the membrane was incubated with rocking for 1.5 h in HRP-conjugated goat anti-rabbit IgG (Bio-Rad, Hercules, CA, USA) at 1/1500 dilution in Blotto. After washing, color was developed by the addition of Immuno-Pure metal enhanced 3,3'-diaminobenzidine substrate (Pierce).

2.5 Protein digestion

The digestion protocol was adapted from similar methods previously applied in proteomics studies on a range of bacterial species [18, 19]. In preparation for MS analysis, samples were denatured with 6 M guanidine and 10 mM DTT for 1 h at 60°C. The reduced and denatured samples were diluted with 50 mM Tris-HCl, 10 mM $CaCl_2$ (pH 7.6) to bring the guanidine concentration to 1 M. Digestion was performed by adding 20 μg modified porcine trypsin (sequencing grade, Promega) to 3 mg protein (for crude and pellet fractions) or 3 μg trypsin to 300 μg protein (for nuclear fraction) at 37°C overnight, followed by a second addition of the same amount of trypsin and incubation for an additional 4 h at 37°C.

Crude and pellet samples were desalted using SepPak Plus C18 cartridges (Waters) following the manufacturer's protocol, with final elution using 100% ACN. Nuclear protein samples were similarly desalted using SepPak Lite cartridges (Waters.) A 500 μL portion of aqueous 0.1% formic acid was added to each desalted sample. ACN was removed using vacuum centrifugation (SpeedVac, Savant Instruments, Holbrook NY) to bring samples to a final volume of ~500 μL. Samples were passed through 0.45 μm Ultrafree-MC filters (Millipore, Bedford, MA, USA) to remove particulates.

2.6 LC-MS/MS

Proteins were identified from digests using MudPIT [12, 13, 20], with 2-D HPLC interfaced with MS/MS, as described previously [18, 19]. Each fraction (crude, pellet, nuclear proteins) was analyzed in duplicate.

Aliquots of 100 μL protein digest, containing ~500 μg protein from crude or pellet fractions or ~150 μg protein from nuclear protein fraction, were loaded *via* a pressure cell (New Objective) onto a "back column" constructed as follows. A 100 μm ID fused-silica capillary column contained a 3–4 cm length of C18 RP resin (Aqua, 5 μm particle, 200 Å pore size [Phenomenex]) upstream of a 3–4 cm length of strong cation exchange phase (SCX; Luna, 5 μm particle, 100 Å pore size [Phenomenex]). The back column was attached *via* a filter union (Upchurch) to the "front column," a 100 μm ID resolving column/nanospray tip packed with C18 RP resin (Jupiter, 5 μm particles, 300 Å pore size [Phenomenex]). The assembled columns were attached to the flow from an HPLC pump (Ultimate, LCPackings/Dionex, Sunnyvale CA, USA). A total flow of 150 μL/min from the pump was split to provide a flow through the column of ~300 nL/min.

Twelve HPLC cycles were performed *per* sample. The first cycle consisted of an RP gradient from 100% solvent A (95% H_2O, 5% ACN, 0.1% formic acid) to 50% solvent B (30% H_20, 70% ACN, 0.1% formic acid) over 45 min followed by a ramp to 100% solvent B over 10 min. In cycles 2–11, 100% solvent A was applied for 5 min, followed by a 2 min salt step gradient of 400 mM ammonium acetate (10, 15, 20, 25, 30, 35, 40, 45, 50 and 60%, respectively, in solvent A), followed by 3 min of 100% solvent A, then an RP gradient (100% solvent A to 50% solvent B over 110 min). In cycle 12, 100% solvent A was applied for 5 min, followed by a 10 min salt step gradient of 400 mM ammonium acetate (100%), followed by 9 min of 100% solvent A, then an RP gradient (100% solvent A to 100% solvent B over 75 min). The back column was removed, and the front column subjected to a final wash. The equilibration step was performed by ramping from 100% solvent A to 100% solvent B over 10 min, holding 2 min at 100% solvent B, ramping to 100% solvent A over 5 min, and holding at 100% solvent A for 10 min. A single front column was used for several experiments, while a new back column was prepared for each LC-MS/MS analysis.

The LC eluent was interfaced *via* a nanospray source with the mass spectrometer (LTQ, ThermoFinnigan, San Jose CA, USA), controlled by XCalibur software. Acquisition of tandem mass spectra was triggered in a data-dependent mode provided by the XCalibur software, with collision-activated dissociation of five parent ions selected from the most intense ions in each full scan mass spectrum. Parent ions selected for MS/MS analysis more than once (*i.e.* repeat count = 1) within 1 min were placed on an exclusion list for 3 min, during which time they were not subjected to collision-activated dissociation. For both full scan mass spectra and tandem mass spectra, two microscans were averaged.

2.7 Protein identification

Peptides were identified using version 27 of the software program Sequest [21] to compare experimental tandem mass spectra with predicted fragmentation patterns of tryptic peptides generated from the protein database for *P. trichocarpa* (version 1.1, available at http://genome.jgi-psf.org/cgi-bin/searchGM?db = Poptr1_1 on July 20, 2007, file contained 45 555 proteins in total) plus common contaminant proteins. A decoy database, containing amino acid sequence-reversed analogs of each protein, was appended to allow estimates of false discovery rates [22, 23]. Sequest searches were carried out with parent ion tolerance of 3.0 m/z units, fragment mass tolerance of 0.5 m/z units. The trypsin cleavage rule was applied, with up to four internal missed cleavage sites allowed *per* peptide. Peptide identifications from Sequest were filtered and organized into protein identifications using DTASelect version 1.9 [24]. Ambiguous peptide identifications were removed using the *–a false* option in DTASelect. Peptide identifications were retained for Sequest results of $\text{XCorr} \geq 1.8$ ($z = 1$), $\text{XCorr} \geq 2.5$ ($z = 2$) or $\text{XCorr} \geq 3.5$ ($z = 3$), and $\text{DeltaCN} \geq 0.08$. These values yielded peptide false discovery rates <0.5%. Identification of a protein required the identification of two or more peptides from that protein or the identification of a single peptide in at least two charge states [25]. False discovery rates at the protein level were < = 1.2%; each known false protein identification was a 2-peptide hit, with one exception which was a 3-peptide hit. Lists of peptide sequences identified from six different runs of xylem protein extracts are available at http://compbio.ornl.gov/populus_tremula_x_alba_proteome/. Estimates of protein quantities were based on comparisons of summed spectrum counts from conserved and/or unique peptides. For a given protein, spectrum count is the number of tandem mass spectra that are assigned to peptides from that protein and provides an approximate indicator of protein abundance [26]. While quantitative estimates discussed in this paper were based on spectrum counts, we have also provided the normalized spectral abundance factor (NSAF) for every protein [27], which can be found in Supporting Information Tables 5–10 at http://compbio.ornl.gov/populus_tremula_x_alba_proteome/. NSAF value is the spectrum count for a protein divided by its length in amino acids, divided by the sum across all proteins of that same quantity. Since only unique peptides are considered, the NSAF is equal to zero for any protein for which no unique peptides were identified.

2.8 Bioinformatics analysis

The annotation information for gene models including locus ID, gene description, protein sequence, conserved domain information, gene ontology (GO annotation), EuKaryotic Orthologous Groups (KOG) information, Enzyme Commission (EC) number annotations and Kyoto Encyclopedia of Genes and Genomes (KEEG) pathway information was obtained from the Joint Genome Institute (JGI) website for *Populus* genome (http://genome.jgi-psf.org/Poptr1_1/). The sub-cellular localization of proteins was predicted using a locally downloaded version of WoLF PSORT (http://wolfpsort.org/) [28], which is an extension of the PSORT II program. WoLF PSORT predicts proteins localizing to major subcellular sites such as nuclear, cytosol, mitochondrial and extracellular regions based on sorting signals, amino acid composition and functional motifs such as DNA-binding motifs. The presence or absence of membrane spanning regions for gene models was predicted using a locally downloaded version of TMpred software (http://www.ch.embnet.org/software/TMPRED_form.html) [29]. A PERL script was written to automate the whole process and assemble the information. The presence or absence of EST support for a given gene model was inferred from NCBI poplar EST and the PopulusDB EST data sets. Nucleotide sequences of gene models were queried against these ESTs using a local BLAST database with an *e*-value cut-off of 1E-10.

The best EST category, which refers to tissue library in which ESTs corresponding to the gene model is best represented, was identified based on EST library distribution information at PopulusDB [30]. A list of highly similar paralogous gene pairs occurring within large conserved syntenic blocks that resulted from the salicoid duplication event in the *Populus* genome has been previously published [31]. This information was used to correlate the extent to which our proteome profiling method could uniquely identify or differentiate between protein products of duplicate gene pairs.

3 Results and discussion

3.1 Identification and functional classification of proteins in *Populus* xylem proteome

The complete xylem proteome data set consisted of proteins representing 5847 distinct *Populus* gene models, which were identified based on matches of tandem mass spectra to the *Populus* genome sequence-based-peptide database (Supporting Information Table 1). The proteins were obtained from developing xylem extracts of wild-type *Populus* plants and further fractionated into crude, pellet and nuclear fractions prior to LC-MS/MS analysis. While the xylem proteome set contained some proteins that were identified from more than one fraction, 1124 proteins were identified exclusively in the crude fraction, 907 were identified in only the pellet fraction and 775 were identified in only the nuclei (Fig. 1).

Sequest searches of the LC-MS/MS data identified 26 757 distinct tryptic peptide sequences. The vast majority of peptides (23 889) contained zero or one missed trypsin cleavage sites, while only 61 peptides contained four missed cleavage sites, indicating the completeness of the digestion. Most of the identified peptides (17 728 or 66%) occur in only a single locus in the annotated *Populus* genome, while 6572

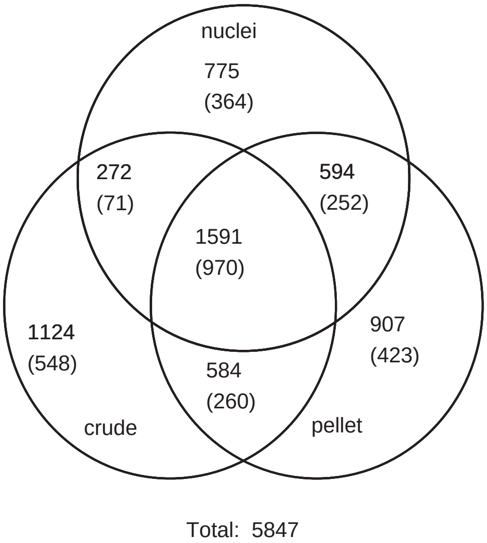

Figure 1. Summary of numbers of identified proteins in various fractions. Numbers in parentheses represent proteins detected in both LC-MS/MS experiments of relevant fractions.

(25%) occur in two loci, 1190 (4%) in three loci, 669 (3%) in four loci and the remainder (598 or 2%) in five or more loci. In total, 48% (12 748) of the peptides were identified by three or more MS/MS spectra.

Based on the annotation of the *Populus* genome sequence, these peptides were assembled into protein identifications. The nomenclature of Yang *et al.* (2004) was followed to classify identified proteins from each LC-MS/MS experiment (parsimony columns in Supporting Information Table 1) and the classification codes are summarized in (Supporting Information Table 2) [32]. *Distinct* proteins contain only peptides that are uniquely found in that locus. *Differentiable* proteins contain one or more peptides that are unique to that locus, as well as one or more peptides that also occur elsewhere in the annotated proteome. The remaining classes contain proteins that are characterized by zero identified peptides that are unique to the locus. *Indistinguishable* proteins have a set of identified peptides that is identical to the set of identified peptides for one or more other proteins in the data set. The indistinguishable proteins can be combined into groups that share sets of peptides. All identified peptides in a *subset* protein also occur in another protein, which contains additional identified peptides. Note that overlap is possible between indistinguishable and subset proteins. Finally, the identified peptides in *subsumable* proteins can be found in two or more proteins encoded by other loci in the annotated genome. Out of the 5847 total protein identifications in the present study, 4283 proteins were uniquely identified (classified as *Distinct* or *Differentiable*).

The distribution of isoelectric points predicted from the amino acid sequences of the experimentally identified proteins (Supporting Information Fig. 1, upper left) is similar to the distribution for all proteins from the predicted proteome (Supporting Information Fig. 1, lower left). The distribution of predicted molecular masses of identified proteins (Supporting Information Fig. 1, upper right) is skewed slightly to higher masses than that of all predicted proteins (Supporting Information Fig. 1, lower right), suggesting a slight bias against smaller proteins in our measurement. A similar experimental bias was noted in a large-scale proteomics study of another plant model system, *Arabidopsis* [33], which employed a significantly different workflow that combined gel electrophoretic separation, in-gel digestion and RP LC-MS/MS analysis.

The molecular clock is estimated to be ticking slowly in *Populus* relative to rice, *Arabidopsis* and human genomes owing to its perennial, long-lived and vegetatively propagating nature [31]. Due to this fact, duplicate genes arising out of the salicoid duplication event, a recent 60 mya genome-wide duplication event that has resulted in nearly two-thirds of genes in the *Populus* genome, are often very similar at the protein sequence level (~90% or higher identity). Our analysis reveals that such duplicate pairs can be differentiated at appreciable levels. For example, 47% of the proteins encoded by duplicated genes that were detected in the crude fraction were differentiable based on identification of unique representative peptides (Supporting Information Table 3). The observation that 74% of all crude fraction proteins were identified based on distinct or differentiable peptide sequences is encouraging with respect to the role proteome profiling can play in poplar biology. It is also significant to note that, among the proteins that were supported by conserved peptides (Indistinguishable peptide groups), about 64% represent proteins from duplicate gene pairs but 36% do not represent duplicates; instead, they were found to represent other closely related protein family members in the same fraction.

The xylem proteome data set could be broadly classified into 23 functional categories (Fig. 2). It was found, as expected, that proteins with housekeeping functions such as histone-fold/TFIID-TAF and histone H4 genes had very high spectrum counts (~400) and certain transcription factors such as MYB1 DNA-binding protein [eugene3.01450016], bZIP family transcription factor [eugene3.01630045], and Homeobox DNA-binding protein [fgenesh4_pg.C_LG_I002015] were found to have low (<10) spectrum counts. A genome-scale proteomics study conducted using various *Arabidopsis* plant organs also found the aforementioned distinction in the categories of proteins under- and over-represented in proteome data sets [33].

It is significant to note that BLAST search of the gene models against locally downloaded poplar ESTs revealed that our shotgun profiling effort uncovered proteins for 99 representative *Populus* gene models that previously lacked experimental validation ("EST support" column in Supporting Information Table 1).

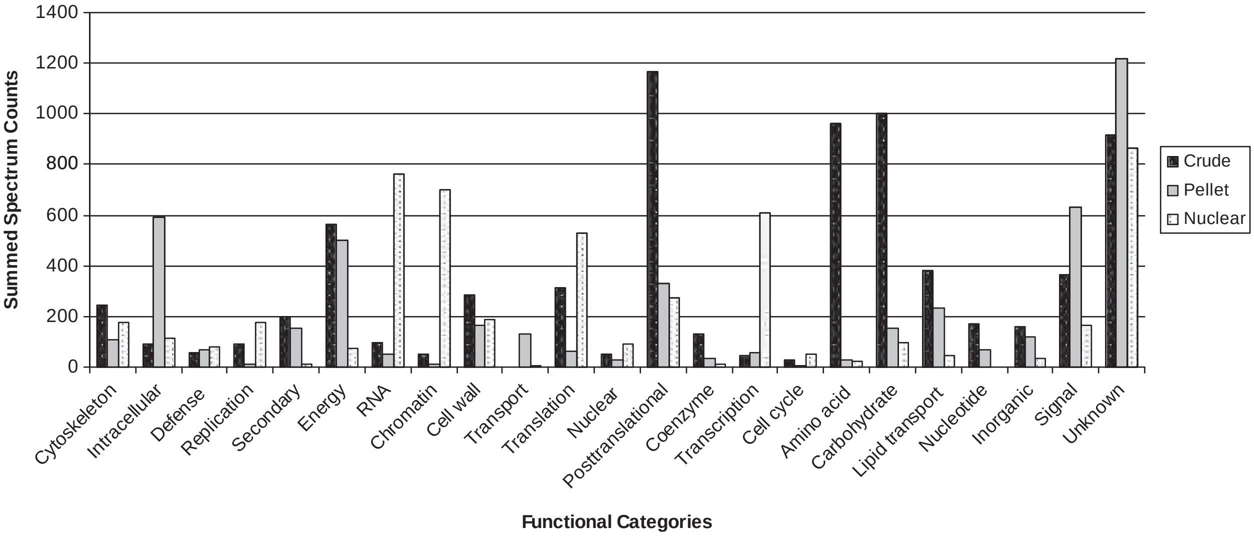

Figure 2. Functional categorization of xylem crude, pellet and nuclear proteomes. Quantitative distribution of proteins identified exclusively from either crude, pellet or nuclear xylem proteomes. For a given protein, spectrum count is the number of tandem mass spectra that are assigned to peptides from that protein, and provides an approximate indicator of protein abundance. The *x*-axis labels represent various predicted functional categories (see Section 2.8); cytoskeleton: cytoskeleton, Intracellular: intracellular trafficking, secretion, and vesicular transport, Defense: defense mechanisms, Replication: replication, recombination and repair, Secondary: secondary metabolites biosynthesis, transport and catabolism, Energy: energy production and conversion, RNA: RNA processing and modification, Chromatin: chromatin structure and dynamics, Cell wall: cell wall/membrane/envelope biogenesis, Transport: transport, Translation: translation, ribosomal structure and biogenesis, Nuclear: nuclear structure, intracellular trafficking, secretion, and vesicular transport, Posttranslational: posttranslational modification, protein turnover, chaperones, Coenzyme: coenzyme transport and metabolism, Transcription: transcription, Cell cycle: cell cycle control, cell division, chromosome partitioning, Amino acid: amino acid transport and metabolism, Carbohydrate: carbohydrate transport and metabolism, Lipid: lipid transport and metabolism, Nucleotide: nucleotide transport and metabolism, Inorganic ion: inorganic ion transport and metabolism, Unknown: unknown, Signal: signal transduction mechanisms. The *y*-axis represents the summed spectrum count of proteins in each functional category.

3.2 Functional categorization of pellet fraction proteins

The protocol applied in collecting the pellet fraction of xylem protein extracts is expected to provide enrichment of membrane proteins, but can also contain significant impurities. In this study, ~3800 proteins were experimentally identified from the pellet fraction of xylem extracts, out of which, 907 were identified exclusively from the pellet fraction. The top functional groups into which proteins found in pellet fractions are categorized are intracellular trafficking, secretion, and vesicular transport; energy production and conversion; cell wall/membrane/envelope biogenesis; posttranslational modification, protein turnover, chaperones; lipid transport and metabolism; signal transduction mechanisms and unknown (Fig. 2). Of the proteins detected in our study that were predicted by TMpred to have at least three transmembrane domains (“TmPred” column in Supporting Information Table 1), 78% were experimentally detected in the pellet fraction. Conversely, only 19% of proteins that were predicted to have at least one transmembrane domain were found exclusively from the crude fraction. Significantly higher spectrum counts were detected in the pellet fractions for the integral membrane protein, cellulose synthase, as compared with crude and nuclear fractions (Fig. 3). The observed differences in spectrum count across fractions for cellulose synthase were considerably greater than that observed for trypsin, which was present in similar amounts in each LC-MS/MS analysis (6 μg for crude and pellet; 3 μg for nuclei). Spectrum counts for autolysis products of trypsin averaged 96.5 in the crude fraction, 70 in the pellet fraction and 79 in the nuclear fraction. This measurable difference in cellulose synthase is supportive of membrane enrichment in the pellet fraction.

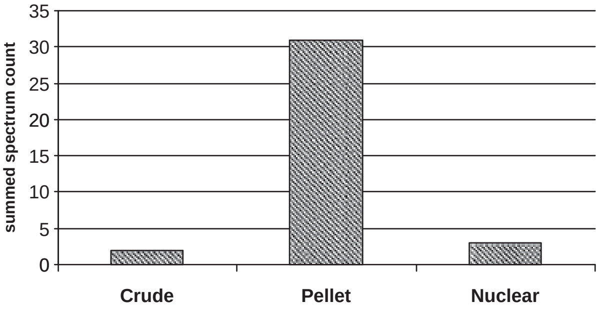

Figure 3. Quantitative estimate of cellulose synthase proteins found exclusively from crude, pellet or nuclear fractions.

It is known that there are several cell wall remodeling or biosynthesis genes that associate or integrate with cellular membranes. Our study shows that the proteins identified exclusively from the pellet fraction have significant over-representation in the cell wall biogenesis sub-category within the broader cell wall/membrane/envelope biogenesis functional category (Supporting Information Table 4). Such proteins include cellulose synthase (CesA) (Fig. 3), sucrose synthase (SuSy), pectinacetylesterase-pectinesterase, rhamnogalacturonatelyase, glycosyl transferases and glycosyl hydrolases (polygalacturonase) [34]. Over-representation of such proteins is further supportive of membrane enrichment in the pellet fraction.

3.3 Identification of known xylem- or cell wall development-associated proteins

Supporting Information Table 4 presents proteomics results for genes found to be enriched in xylem development or wood formation-related PopulusDB EST libraries (tension wood, shoot meristem, cambial zone, active cambium, bark, wood cell death, roots, petioles and apical shoot [30]). A sub-data set of the genes that were classified into the cell wall functional category and had strong EST support from tension wood (characterized by high cellulose content in xylem cell walls as well as higher xylem cell count) libraries contained such known cellulose biosynthesis pathway proteins as the secondary cell-wall-associated CesAs [eugene3.00040363 and gw1.XI.3218.1] and several α-TUBULIN and β-TUBULIN proteins [*e.g.* gw1.IX.2621, *Differentiable* protein] [6]. Though peptides representing several different SuSy proteins were discovered from the xylem proteome set, just two SuSy isoforms [estExt_fgenesh4_pm.C_LG_XVIII0009 and estExt_fgenesh4_pg.C_280066] were present in the sub-data set generated based on criteria of EST expression in tension wood library [6, 35–37]. Based on the high spectrum count relative to the average count of proteins in the cell wall category, both SuSy isoforms appear to be highly expressed at the protein level. Interestingly, only these two SuSy isoforms were uniquely identified (DS or DF), from the pellet fraction. It is known that these isoforms express at a higher level during enhanced cellulose biosynthesis (as occurs in xylem and under tension stress) relative to other tissue types, and it is believed that SuSy participates in primary metabolism when localized in cytosol and in secondary metabolism (cell wall synthesis), when membrane-associated [37–39]. In light of this knowledge, the identification of *Populus* sucrose synthase proteins in two replicate pellet (predominantly membrane) analyses adds credence to the theory that membrane-associated *Populus* sucrose synthase partakes in cellulose biosynthesis in xylem cells. Other previously reported wall-remodeling proteins that we identified include polygalacturonase, laccase (diphenol oxidase) and fasciclin and related adhesion glycoproteins.

3.4 Characterization of the xylem nuclear proteome

Nuclear fractionation of xylem was undertaken to identify low-abundance nuclear-localized proteins including nucleic acid binding and regulatory proteins. About 77% of WoLF PSORT-based nuclear localization predictions are correlated with experimental evidence ("wolfpsort" column in Supporting Information Table 1). The efficiency of the nuclear enrichment method was inferred to be significantly high based on the abundance of specific nuclear marker proteins, as Fig. 4A demonstrates for histone and histone-associated proteins. Compared with crude and pellet fractions, the nuclear fraction had a significantly higher representation of proteins relating to such nuclear processes as replication, recombination and repair; nuclear structure, intracellular trafficking, secretion and vesicular transport; cell cycle control, cell division and chromosome partitioning; chromatin structure and dynamics and transcription (Fig. 5). The nuclear protein enrichment was also validated through a Western hybridization experiment using an anti-histone antibody (Fig. 4B).

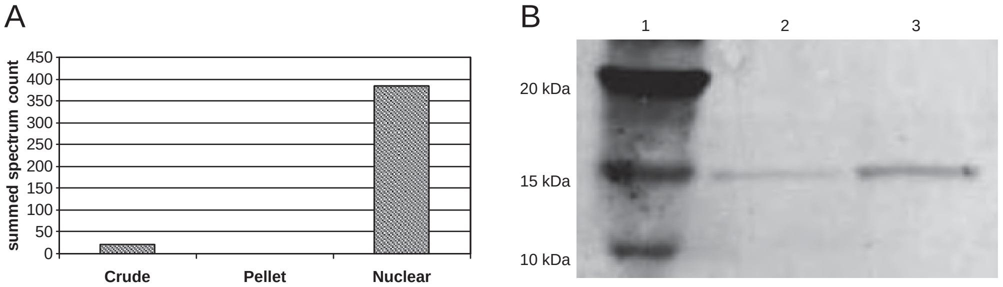

Figure 4. (A) Quantitative estimate of histone and histone-associated proteins found exclusively from crude, pellet or nuclear fractions. (B) Western blot of nuclear and crude protein fractions using anti-histone antibody. Lane 1, molecular weight standards; lane 2, 1 μg protein extracted from xylem cells; lane 3, 1 μg of protein extracted from enriched xylem nuclei. The blot was probed with anti-Histone H3 antibody.

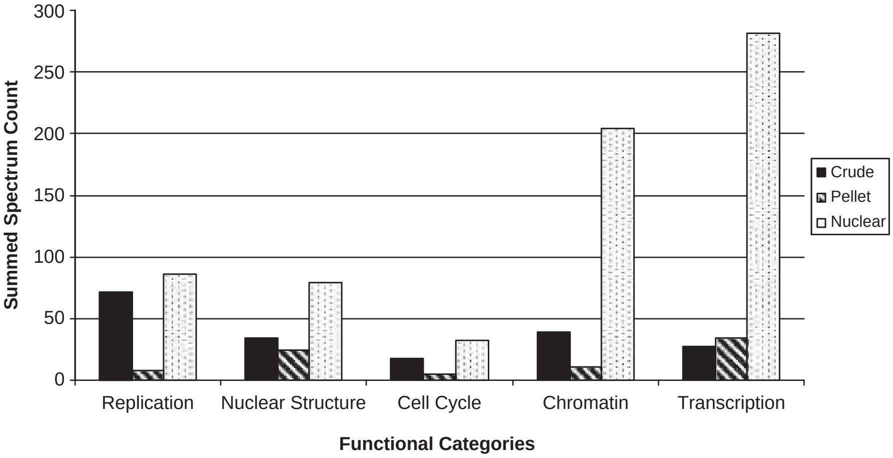

Figure 5. Distribution of exclusive crude, pellet or nuclear proteins within five functional categories that are related to nuclear processes. The data set consists of only uniquely identified proteins (DS or DF). The *x*-axis labels represent various predicted nucleus-associated functional categories. The *y*-axis represents the summed spectrum count of proteins in each functional category.

3.5 Identification of putative transcription and regulatory factors from the nuclear proteome

Transcription factors (ARF, NAC, MYB, JUMONJI, LIM, BLH3, HD-ZIP, HOMEOBOX domain proteins) and regulatory factors involved in chromatin remodeling (histone acetylates and deacetylases, SWI/SNF, SWIFT/BRCT domain proteins), RNA processing or modifying (RRM, Helicase, DEAD/DEAH, SnoRNP, PABP domain proteins) and cell cycle/replication (cyclin, retinoblastoma-associated, tudor, DNA-repair proteins) represent nearly half (44–47%) of the proteins identified exclusively from the nuclear fraction.

Interestingly, transcript-level data obtained from the *Populus* EST database (PopulusDB) suggests that many of the nuclear localized putative transcription regulators identified from xylem proteome are also preferentially represented in xylem libraries (cambium, stem, wood, tension wood, xylem and petiole libraries) ("Best EST Category" in Supporting Information Table 1). Such candidate genes have no reported reverse or forward genetic mutant observations.

Among the proteins in the "transcription" category, one gene [eugene3.00140349] had a particularly high and specific EST expression support from wood-forming tissues such as cambial zone and tension wood. We found that the protein coded by this gene, a CtBP-like transcription factor, was experimentally detectable from two nuclear MS/MS analyses, as well as predicted to be nuclear by WoLF PSORT program. The closest *Arabidopsis* homolog of this gene is *ANGUSTIFOLIA* [40]. It is has been shown that the *ANGUSTIFOLIA* gene product controls the cortical tubular network and likely the expression of *MERI5* gene, a xyloglucan endotransglycosylase required in cell wall remodeling [41]. Interestingly, our xylem proteome data set also contains the closest *Arabidopsis* MERI5 homolog, gw1.XVIII.2837.1. It is reported that *Populus* xyloglucan endotransglycosylase functions in gelatinous layers of tension wood fibers even days after cell death [7]. Additionally, NAC-domain transcription factor proteins [gw1.I.5485.1 and gw1.XI.947.1] having strong homology to transcription factors known to control secondary cell wall formation in *Arabidopsis* such as SND1 [42] and NST1 [43] proteins were found exclusively in the nuclear fraction. Other interesting nuclear proteins include a Myb DNA-binding protein [gw1.IV.467.1] and a CHB3-SWI/SNF-related protein. It will be valuable to evaluate the roles of such transcription factor genes as master regulators in the *Populus* cell wall remodeling pathway.

4 Concluding remarks

The present study aimed at applying MudPIT to study *Populus* developing xylem tissue. Our results show that the technique successfully isolated and identified ~6000 proteins from xylem tissue, greatly expanding the numbers of protein identifications reported from previous *Populus* proteome studies [8–11]. Subcellular proteomics has a twofold advantage of indicating cellular contexts for functional roles as well as potentially detecting low-abundance proteins. Our attempt to enrich nuclei from xylem was successful as indicated by the presence in this fraction of a high number of detected proteins that are associated with nuclear processes. We identified in this study several candidate secondary cell wall or wood formation regulator proteins highly similar to SND1, NST1, CtCP, CHB3-SWI/SNF-related proteins. In addition, many proteins of as yet unknown function but predicted to be nuclear were experimentally found to be abundant in the nuclear proteome and also had enhanced transcript expression in wood- or xylem-related tissue contexts. Such genes are good candidates for further functional genomics investigations. Differential representation of SuSy proteins in soluble and predominantly membrane fractions was found to be consistent with the knowledge of membrane associatedness of certain SuSy isoforms.

Conserved duplicate gene pairs that originated from the salicoid duplication event in *Populus* were found to be distinguishable at appreciable rates. Data generated by shotgun proteome profiling can also prove useful in gene annotation. Our proteome data set presented first-time-expression support for ~100 predicted *Populus* gene models, attesting to the functional validity of these gene models. Further applications of the proteomics technique will be useful in genome annotation as well as in functional classification of *Populus* genes, many of which have a representative duplicate in the genome. Moreover, the proteome data set discussed in this report, albeit from one particular plant tissue, xylem, will be useful in future efforts to predict and understand gene functions in much the same way that we have used the EST database to obtain transcript level support from wood-forming tissues.

This study presents an additional tool for systems biology investigations in *Populus*. While not explored in the current study, proteomics can also be applied to yield valuable knowledge with respect to post-translational modifications and quantitative differences in protein expression. The ability to rapidly identify and contrast whole proteomes from different cellular fractions across treatments and genetic variations along with supportive bioinformatics capabilities will be key to providing accurate systems biology models of not only cell wall biosynthesis but also other plant developmental and physiological pathways.

The authors would like to thank Manesh Shah for website assistance and Drs. Brian Davison and Stan Wullschleger for their technical reviews of this paper. The present study was enabled by research funds to U.C.K. through the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory (ORNL) and the BioEnergy Science Center, a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. ORNL is managed by UT-Battelle, LLC, under contract DE-AC05-00OR22725 for the U.S. Department of Energy.

The authors have declared no conflict of interest.

5 References

[1] Jansson, S., Douglas, C. J., *Populus:* a model system for plant biology. *Annu. Rev. Plant Biol.* 2007, *58*, 435–458.

[2] Higuchi, T., *Biochemistry and Molecular Biology of Wood*, 1st Edn., Springer-Verlag, Berlin-Heidelberg 1997.

[3] Israelsson, M., Eriksson, M. E., Hertzberg, M., Aspeborg, H. *et al.* Changes in gene expression in the wood-forming tissue of transgenic hybrid aspen with increased secondary growth. *Plant Mol. Biol.* 2003, *52*, 893–903.

[4] Karpinska, B., Karlsson, M., Srivastava, M., Stenberg, A. *et al.* MYB transcription factors are differentially expressed and regulated during secondary vascular tissue development in hybrid aspen. *Plant Mol. Biol.* 2004, *56*, 255–270.

[5] Kalluri, U. C., Joshi, C. P., Differential expression patterns of two cellulose synthase genes are associated with primary and secondary cell wall development in aspen trees. *Planta* 2004, *220*, 47–55.

[6] Andersson-Gunneras, S., Mellerowicz, E. J., Love, J., Segerman, B. *et al.* Biosynthesis of cellulose-enriched tension wood in *Populus:* global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis. *Plant J.* 2006, *45*, 144–165.

[7] Nishikubo, N., Awano, T., Banasiak, A., Bourquin, V. *et al.* Xyloglucan endo-transglycosylase (XET) functions in gelatinous layers of tension wood fibers in poplar--a glimpse into the mechanism of the balancing act of trees. *Plant Cell Physiol.* 2007, *48*, 843–855.

[8] Ferreira, S., Hjerno, K., Larsen, M., Wingsle, G. *et al.* Proteome profiling of *Populus euphratica* Oliv. upon heat stress. *Ann. Bot. (Lond.)* 2006, *98*, 361–377.

[9] Du, J., Xie, H. L., Zhang, D. Q., He, X. Q. *et al.* Regeneration of the secondary vascular system in poplar as a novel system to investigate gene expression by a proteomic approach. *Proteomics* 2006, *6*, 881–895.

[10] Plomion, C., Lalanne, C., Claverol, S., Meddour, H. *et al.* Mapping the proteome of poplar and application to the discovery of drought-stress responsive proteins. *Proteomics* 2006, *6*, 6509–6527.

[11] Kieffer, P., Dommes, J., Hoffmann, L., Hausman, J. F., Renaut, J., Quantitative changes in protein expression of cadmium-exposed poplar plants. *Proteomics* 2008, *8*, 2514–2530.

[12] Link, A. J., Eng, J., Schieltz, D. M., Carmack, E. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* 1999, *17*, 676–682.

[13] Washburn, M. P., Wolters, D., Yates J. R., III, Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* 2001, *19*, 242–247.

[14] Koller, A., Washburn, M. P., Lange, B. M., Andon, N. L. *et al.* Proteomic survey of metabolic pathways in rice. *Proc. Natl. Acad. Sci. USA* 2002, *99*, 11969–11974.

[15] Maor, R., Jones, A., Nuhse, T. S., Studholme, D. J. *et al.* Multidimensional protein identification technology (MudPIT) analysis of ubiquitinated proteins in plants. *Mol. Cell. Proteomics* 2007, *6*, 601–610.

[16] Gion, J. M., Lalanne, C., Le Provost, G., Ferry-Dumazet, H. *et al.* The proteome of maritime pine wood forming tissue. *Proteomics* 2005, *5*, 3731–3751.

[17] Lowry, O. H., Rosebrough, N. J., Farr, A. L., Randall, R. J., Protein measurement with the Folin phenol reagent. *J. Biol. Chem.* 1951, *193*, 265–275.

[18] VerBerkmoes, N. C., Shah, M. B., Lankford, P. K., Pelletier, D. A. *et al.* Determination and comparison of the baseline proteomes of the versatile microbe *Rhodopseudomonas palustris* under its major metabolic states. *J. Proteome Res.* 2006, *5*, 287–298.

[19] Mahowald, M. A., Rey, F. E., Seedorf, H., Turnbaugh, P. J. *et al.* Characterizing a model human gut microbiota

composed of members of its two dominant bacterial phyla. *Proc. Natl. Acad. Sci. USA* 2009, *106*, 5859–5864

[20] McDonald, W. H., Ohi, R., Miyamoto, D. T., Mitchison, T. J., Yates, J. R., Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single-dimension LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT. *Int. J. Mass Spectrom.* 2002, *219*, 245–251.

[21] Eng, J. K., Mccormack, A. L., Yates, J. R., An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[22] Elias, J. E., Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 2007, *4*, 207–214.

[23] Moore, R. E., Young, M. K., Lee, T. D., Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* 2002, *13*, 378–386.

[24] Tabb, D. L., McDonald, W. H., Yates J. R., III, DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* 2002, *1*, 21–26.

[25] Sardiu, M. E., Cai, Y., Jin, J., Swanson, S. K. *et al.* Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *Proc. Natl. Acad. Sci. USA* 2008, *105*, 1454–1459.

[26] Liu, H., Sadygov, R. G., Yates J. R., III, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 2004, *76*, 4193–4201.

[27] Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K. *et al.* Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J. Proteome Res.* 2006, *5*, 2339–2347.

[28] Horton, P., Park, K. J., Obayashi, T., Fujita, N. *et al.* WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* 2007, *35*, W585–W587.

[29] Hofmann, K., Stoffel, W., Tmbase – a database of membrane spanning proteins segments. *Biol Chem. Hoppe-Seyler*, 1993, *374*, 166.

[30] Sterky, F., Bhalerao, R. R., Unneberg, P., Segerman, B. *et al.* A *Populus* EST resource for plant functional genomics. *Proc. Natl. Acad. Sci. USA* 2004, *101*:13951–13956.

[31] Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 2006, *313*, 1596–1604.

[32] Yang, X. Y., Dondeti, V., Dezube, R., Maynard, D. M. *et al.* DBParser: Web-based software for shotgun proteomic data analyses. *J. Proteome Res.* 2004, *3*, 1002–1008.

[33] Baerenfaller, K., Grossmann, J., Grobei, M. A., Hull, R. *et al.* Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science* 2008, *320*, 938–941.

[34] Geisler-Lee, J., Geisler, M., Coutinho, P. M., Segerman, B. *et al.* Poplar carbohydrate-active enzymes. Gene identification and expression analyses. *Plant Physiol.* 2006, *140*, 946–962.

[35] Joshi, C. P., Bhandari, S., Ranjan, P., Kalluri, U. C. *et al.* Genomics of cellulose biosynthesis in poplars. *New Phytol.* 2004, *164*, 53–61.

[36] Suzuki, S., Li, L., Sun, Y. H., Chiang, V. L., The cellulose synthase gene superfamily and biochemical functions of xylem-specific cellulose synthase-like genes in *Populus trichocarpa*. *Plant Physiol.* 2006, *142*, 1233–1245.

[37] Haigler, C. H., Ivanova-Datcheva, M., Hogan, P. S., Salnikov, V. V. *et al.* Carbon partitioning to cellulose synthesis. *Plant Mol. Biol.* 2001, *47*, 29–51.

[38] Hardin, S. C., Winter, H., Huber, S. C., Phosphorylation of the amino terminus of maize sucrose synthase in relation to membrane association and enzyme activity. *Plant Physiol.* 2004, *134*, 1427–1438.

[39] Hardin, S. C., Duncan, K. A., Huber, S. C., Determination of structural requirements and probable regulatory effectors for membrane association of maize sucrose synthase 1. *Plant Physiol.* 2006, *141*, 1106–1119.

[40] Stern, M. D., Aihara, H., Cho, K. H., Kim, G. T. *et al.* Structurally related Arabidopsis ANGUSTIFOLIA is functionally distinct from the transcriptional corepressor CtBP. *Dev. Genes Evol.* 2007, *217*, 759–769.

[41] Kim, G. T., Shoda, K., Tsuge, T., Cho, K. H. *et al.* The ANGUSTIFOLIA gene of Arabidopsis, a plant CtBP gene, regulates leaf-cell expansion, the arrangement of cortical microtubules in leaf cells and expression of a gene involved in cell-wall formation. *EMBO J.* 2002, *21*, 1267–1279.

[42] Zhong, R., Demura, T., Ye, Z. H., SND1, a NAC domain transcription factor, is a key regulator of secondary wall synthesis in fibers of Arabidopsis. *Plant Cell* 2006, *18*, 3158–3170.

[43] Zhong, R., Richardson, E. A., Ye, Z. H., Two NAC domain transcription factors, SND1 and NST1, function redundantly in regulation of secondary wall synthesis in fibers of Arabidopsis. *Planta* 2007, *225*, 1603–1611.