MCP Papers in Press. Published on October 16, 2012 as Manuscript M112.022996

Putting the Pieces Together: High-performance LC-MS/MS Provides Network-, Pathway-, and Protein-level Perspectives in *Populus*

Authors: Paul Abraham[†, 1, 2], Richard J. Giannone[†, 2], Rachel M. Adams[1, 2], Udaya Kalluri[3], Gerald A. Tuskan[3], Robert L. Hettich[2]

[†]Authors contributed equally

[1]Graduate School of Genome Science and Technology, University of Tennessee, Knoxville TN 37830; [2]Chemical Sciences Division at Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA; [3]Biosciences Division

* To whom correspondence should be addressed. Mailing address: (RLH) Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6131. Phone: 865-574-4986

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

ABBREVIATIONS:

ADF, actin depolymerizing factor; BCA, bicinchoninic acid assay; FDR, false discovery rate; FASP, filter-aided sample preparation; KOG, eukaryotic clusters of orthologous groups; LPI, leaf plastichronic index; LTQ, linear ion trap mass spectrometer; MudPIT, multidimensional protein identification technology; nSpC, normalized spectra counts; 2PG, 2-phosphoglycolate; 3PG, 3-phosphoglycerate; PR, photorespiration; PV, prevalence value; RP, reversed phase; RuBisCO, ribulose-1, 5-bisphosphate carboxylase/oxygenase; SCX, strong cation exchange; 2D-LC-MS/MS, two-dimensional-liquid chromatography-tandem mass spectrometry; UGPase, UDP-glucose pyrophosphorylase; UGDH, UDP-glucose dehydrogenase

ABSTRACT

High-performance mass spectrometry (MS)-based proteomics enabled the construction of a detailed proteome atlas for *Populus*, a woody perennial plant model organism. Optimization of experimental procedures and implementation of current state-of-the-art instrumentation afforded the most detailed look into the predicted proteome space of *Populus*, offering varying proteome perspectives: 1) network-wide, 2) pathway-specific, and 3) protein-level viewpoints. Together, enhanced protein retrieval through a detergent-based lysis approach and maximized peptide sampling via the dual-pressure linear ion trap mass spectrometer (LTQ Velos), have resulted in the identification of 63,056 tryptic peptides. The technological advancements, specifically spectral-acquisition and sequencing speed, afforded the deepest look into the *Populus* proteome, with peptide abundances spanning 6 orders of magnitude and mapping to

~25% of the predicted proteome space. In total, tryptic peptides mapped to 13,574 protein assignments across four organ-types: mature (fully expanded, leaf plastichronic index (LPI) 10-12) leaf, young (juvenile, LPI 4-6) leaf, root, and stem. To resolve protein ambiguity, identified proteins were grouped by sequence similarity (≥ 90%), thereby reducing the protein assignments into 7,538 protein groups. In addition, this large-scale data set features the first systems-wide survey of protein expression across different *Populus* organs. As a demonstration of the precision and comprehensiveness of the semi-quantitative analysis, we were able to contrast two stages of leaf development, mature versus young leaf. Statistical comparison through ANOVA analysis revealed 1,432 protein groups that exhibited statistically significant ($p \leq 0.01$) differences in protein abundance. Experimental validation of the metabolic circuitry expected in mature leaf (characterized by photosynthesis and carbon fixation) compared to young leaf (characterized by rapid growth and moderate photosynthetic activities) strongly testifies to the credibility of the approach. Instead of quantitatively comparing a few proteins, a systems view of all the changes associated with a given cellular perturbation could be made.

1. INTRODUCTION

Mass spectrometry (MS)-based proteomics has experienced tremendous growth in recent years, leading to the establishment of numerous protocols, platforms, and workflows for the characterization of protein expression at the genome level[1]. While these advancements have facilitated comprehensive proteomic investigations of simple bacterial isolates and microbial communities, the application of MS-based proteomics for plants and other higher eukaryotes remains underdeveloped. Recently, large-scale proteomic studies have been directed at characterization of *Populus*, a woody perennial model organism. With the recent release and subsequent curation of the *P. trichocarpa* genome[2], these large-scale MS-based proteomic investigations offer the potential to introduce new biological insights into woody perennial plant biology[3,4,5]. For example, we have recently demonstrated the ability to measure ~17% of the *Populus* proteome by coupling multi-dimensional liquid chromatography (MudPIT) with nano-electrospray tandem mass spectrometry (2D-LC-MS/MS)[6]. Relative to the two-dimensional gel-based approaches[7], MudPIT provides enhanced separation and when used in conjunction with MS/MS, surpasses the throughput and number of identifiable proteins detected in complex mixtures[8]. Although we have demonstrated the general effectiveness of this approach, the identification and quantitation of the proteins expressed in a plant cell or tissue are still notoriously complicated by a number of factors, including the size and complexity of plant genomes, abundance of protein variants, as well as the dynamic range of protein identification. To overcome these challenges, improvements are needed in sample preparation, MS instrumentation, and data interpretation.

The architecture of plant cell walls provides resistance to chemical and biological degradation, thus requiring mechanical and detergent-based lysis for optimal proteome analysis. However, this criterion presents a major challenge for plant proteomic research using electrospray mass spectrometry, as detergent-containing solutions can impede enzymatic digestion and cause significant analyte suppression[9]. Therefore, most plant proteomic studies using the ‘MudPIT’ strategy apply mechanical disruption in conjunction with a detergent-free preparation method[10]. Typically, strong chaotropic agents such as urea and guanidine hydrochloride are used for the extraction, denaturation, and digestion of proteins. In a recent study, Mann *et al.* (2009) introduced a filter-aided sample preparation (FASP) method that utilizes and effectively removes sodium dodecyl sulfate (SDS) prior to enzymatic digestion and electrospray analysis[11]. This study demonstrated enhanced retrieval of peptides from biological materials, yielding a more accurate representation of the proteome. We developed a similar experimental approach for extraction of proteins from plant tissue in order to obtain a more comprehensive, unbiased proteome characterization well beyond that achievable with currently available methods. Similar to the FASP method, we demonstrate the power of SDS for proteomic sample preparation, not only in its ability to more-thoroughly lyse cells, but also its ability to better solubilize both hydrophilic and hydrophobic proteins. This powerful attribute gives proteolytic enzymes maximum opportunity to generate peptides specific to their cleavage potential so that at least a few representative peptides can be obtained for proteins that would have otherwise been discarded or lost due to insolubility, e.g., membrane-bound proteins. Rather than performing a buffer exchange with urea, depletion of SDS is achieved by precipitating proteins out of solution using trichloroacetic acid.

Characterization of protein expression in plants is further complicated by the heterogeneous mixture of various cell types, each with a unique proteome signature and individualized response to environmental chemical or physical signals. This inherent complexity of plant proteomes and the large dynamic range in protein abundance overwhelms current analytical platforms[12]. Moreover, biochemical regulatory networks in plants are more elaborate and dynamic than in microbial species; consequently, many biological components are left undiscovered, including modified peptides and low-abundance proteins[13,14,15]. Recent developments in ion-trap MS instrumentation, namely the dual-pressure linear ion trap mass spectrometer (LTQ Velos), have demonstrated improved ability to comprehensively characterize complex proteomics samples[16]. Featuring a newly designed ion source and a two-chamber ion trap mass analyzer, the LTQ Velos achieves greater dynamic range, sensitivity, and speed of spectral acquisition when applied to complex proteomic samples. Cumulatively, the technological advancements afford substantial increases in the detection and identification of both proteins and unique peptides when compared to existing state-of-the-art technologies. Therefore, to satisfy the need for depth of proteome characterization in plants, we apply the newly developed LTQ Velos for mass spectrometry measurements of the *Populus* proteome.

For most terrestrial plants, life begins and ends in the same physical location. For woody perennial plants, this sedentary lifestyle may last thousands of years. One consequence of this lifestyle is that each plant typically experiences dramatic changes in its ambient environment throughout its lifetime and, at any given time, equilibrium between endogenous growth processes and exogenous constraints exerted by the environment must be tightly controlled. To survive under varying environmental conditions, temporal plastic responses evoke patterns of protein expression that progressively influence morphological, anatomical and functional traits of three principal organs -- leaf, root and stem. Collectively and individually, these organs operate to perceive and respond to periodic and chronic environment conditions. Currently, a comprehensive understanding of the spatial variation in protein expression patterns across the organ types is lacking for woody perennial plants, where most large-scale proteome analyses with *Populus* were performed on isolated organs, tissues, organelles, or subcellular structures. For this reason, we combined the state-of-the-art LTQ-Velos platform with the SDS/TCA sample preparation methodology to generate a high-coverage proteome atlas of the principal organ types from *Populus*.

2. EXPERIMENTAL PROCEDURES

2.1 Plant Material. The *Populus tremula* x *alba* clone, ‘717’, was grown under standard greenhouse conditions outlined in Kalluri *et al*. (2009). From these trees, mature fully expanded leaf including the petiole and midrib (leaf plastichronic index (LPI) 10-12) and young leaf including the petiole and midrib (LPI 4-6) samples, fine roots less than 2 mm in diameter and young photosynthetically active stem segments less than 5 mm in diameter were collected, immediately frozen in liquid nitrogen and stored at -80°C until use. Harvesting tissue samples from 6 month old trees afforded little biomass and thus confined the experimental design to only a single biological replicate per organ type. To reduce the effects of biological variation, tissues across 6 individual trees were pooled together for each organ type.

2.2 Protein Extraction and Quantification. Leaf, root and stem tissues were ground under liquid nitrogen using a mortar and pestle. For each organ-type, a 1.5 g sample of ground tissue was suspended in SDS lysis buffer (4% SDS in 100 mM of Tris-HCl), boiled for 5 min, sonically disrupted (40% amplitude, 10 s pulse with 10 s rest, 2 min total pulse time), and boiled for an additional 5 min. Crude protein extract was pre-cleared via centrifugation at 4500 x g for 10 min, quantified by BCA assay (Pierce Biotechnology), and reduced with 25 mM dithiothreitol (DTT). Three milligrams of crude protein extract were then precipitated by trichloroacetic acid (TCA), pelleted by centrifugation, and washed with ice-cold acetone to remove lipids and excess SDS as previously described[17].

2.3 Protein Digestion. As previously described, pelleted proteins were then resuspended in 250 µL of 8 M urea, 100 mM Tris-HCl, pH 8.0 (denaturant) using sonic disruption to fully solubilize the protein pellet and incubated at room temperature for 30 min[17]. Denatured proteins were reduced with DTT (5 mM) and cysteines were blocked with iodoacetamide (20 mM) to prevent reformation of disulfide linkages. Samples were digested via two aliquots of sequencing-grade trypsin (Promega, 1:75 [w/w]) at two different sample dilutions, 4 M urea (overnight) and 2 M urea (4 hr). Following digestion, samples were adjusted to 200 mM NaCl, 0.1% formic acid and filtered through a 10 kDa cutoff spin column filter (Vivaspin 2, GE Health) to remove underdigested proteins. The peptide-enriched flow through was then quantified by BCA assay, aliquoted, and stored at −80°C until analysis.

2.4 LC-MS/MS. Peptide analysis was performed using online two-dimensional liquid chromatography interfaced with a linear ion trap mass spectrometer (LTQ Velos by Thermo Scientific). Peptides from each of the 4 organ-types were analyzed over 5-6 technical replicates per sample for a total of 23 MS runs. For each sample, 100 µg of peptides were bomb-loaded onto a biphasic MudPIT back column packed with ~5 cm strong cation exchange (SCX) resin for charge-based separation of peptides followed by ~3 cm C18 reversed phase (RP) for online desalting (Luna and Aqua respectively, Phenomenex). Each peptide-loaded column was first washed off-line to remove residual urea and NaCl and then placed in-line with an in-house pulled nanospray emitter (100 micron ID) packed with 15 cm of C18 RP material and analyzed via 24-hr MudPIT 2D-LC-MS/MS as previously described[6]. Data-dependent acquisition of tandem mass spectra employed the following parameters: collision-activated dissociation (35% energy) of 10 parent ions (MS/MS, 2 µscans) following every full scan (2 µscans) with a 3 m/z isolation width, and a dynamic exclusion repeat of 1 with duration of 60 second (XCalibur version 2.1).

2.5 Database Searching and Peptide Identification. Experimental MS/MS spectra were compared to theoretical tryptic peptide sequences generated from a FASTA database containing (1) the full protein complement of *P. trichocarpa* (v2.2, released in 2011, available at http://www.phytozome.net/cgi-bin/gbrowse/poplar/, containing 45,778 proteins), (2) mitochondria and chloroplast proteins[2], and (3) common contaminant proteins (i.e., bovine trypsin and human keratin). A decoy database, consisting of the reversed sequences of the target database, was appended in order to discern the false-discovery rate (FDR) at the peptide level. Using this protein database of 93,330 entries, peptide fragmentation spectra (MS/MS) were assigned peptide sequences with the SEQUEST algorithm v.27[18], employing the following parameters: ≤ 4 missed tryptic cleavages allowed, a parent ion mass tolerance of 3.0 m/z units, a fragment mass tolerance of 0.5 m/z units, and a static modification on cysteine (iodoacetamide; +57 Da). Resulting peptide identifications from SEQUEST were filtered and organized into protein identifications using DTASelect v.1.9[19] with each peptide identification requiring XCorr values of at least 1.8 (+1), 2.5 (+2), or 3.5 (+3) and a DeltaCN ≥ 0.08.

2.6 Protein Inference / Normalization for Semi-quantitative Analysis. In order to deal with the redundancy associated with the *Populus* genome, all proteins in the FASTA database (includes P. trichocarpa v2.2, mitochondria, and chloroplast predicted proteomes) were grouped by sequence similarity (≥ 90%) using the UCLUST component of the USEARCH v. 5.0 software platform[20]. As described by Abraham et al., grouping proteins by this very conservative level of sequence identity serves to maintain biologically-relevant peptide information that would have otherwise been lost due to proteomic redundancy[6]. Once identified proteins were consolidated into the predefined protein groups, spectra counts were balanced and converted to normalized spectra counts (nSpC)[17], which are derived from normalized spectral abundance factors[21], values that are commonly used for semi-quantitative proteomic analyses. To assess differences between organ types, only those proteins with substantive nSpC, as determined by prevalence value (PV)[22], were carried on to subsequent analyses. Briefly, each protein identified is given a PV, which is determined by averaging the nSpC values across all samples. Next, PVs were plotted as a histogram to graphically capture the distribution of assigned spectra, such that one could assess the cumulative spectra assigned at varying PV cut-offs. Through iterative removal of proteins below each PV cut-off, only proteins considered to be highly representative or reproducible remained. Using this approach, an ideal PV cut-off of 2.0 was determined.

2.7 Hierarchical Clustering, ANOVA, and Pathway Analysis. Protein groups passing the PV filter were log2-transformed to obtain a normal distribution of data that could be subsequently analyzed by the hierarchical clustering and ANOVA analysis packages of JMP Genomics v.4.1 (SAS Institute). For hierarchical clustering, transformed data across all organ types were compared via the Fast Ward clustering algorithm using the STD option to standardize protein abundance values across all organs on a protein-by-protein basis, which essentially converts abundance values to standard deviations above or below the row mean. Proteins exhibiting similar trends across all organs were grouped into clusters and visualized by heat map to ascertain organ-specific protein representation.

Though difficult to assess semi-quantitative differences in protein abundance across all organs, due in part to the distinct proteomes expressed by spatially distant organs, a semi-quantitative comparison between a single organ at two different developmental stages is perhaps a more robust and intuitive measurement. As follows, log2-transformed nSpC values collected across both young and mature leaves were analyzed for statistically significant differences in protein abundance using JMP Genomics' ANOVA feature. Protein groups exhibiting significant ($p \leq 0.01$) abundance differences were identified and represented in both tabular and graphical format, the latter of which utilized KEGG pathway mapping software iPATH2.0[23] to visualize differences between both developmental stages on a more global, metabolic pathway-centric level. Proteins exhibiting increased abundance in mature leaf were represented as red edges on the KEGG maps while those increased in young leaf were represented as green edges. Edges were further color-coded based on fold change, with the color intensity correlated to those proteins with larger fold changes.

3. RESULTS AND DISCUSSION

3.1 Global Protein Identification in *Populus*

A protein sample derived from plant tissue is likely to consist of over 10,000 different protein species present at any time and thus the complexity far exceeds an analogous sample derived from any prokaryotic species. The first step in accurate and deep proteome characterization in these mixtures must consist of an optimal cell lysis and protein solubilization strategy. For plant tissue, we devised a shotgun proteomics workflow that combines the advantages of extensive proteome solubilization in SDS with the benefits of in-solution digestion.

In an effort to generate a high-density proteomic atlas that accurately captures the predicted *Populus* proteome, individual proteome maps of the four major organ-types were integrated. In total, we performed multiple (5-6 each) LTQ Velos ion-trap mass spectrometry measurements on proteome extracts from root, stem and both mature (fully expanded, leaf plastichronic index (LPI) 10-12) and young leaf (LPI 4-6) samples. The resulting tandem mass spectra (MS/MS) were searched (SEQUEST) against the most recent protein database of *P. trichocarpa*, containing 45,778 predicted proteins and supplemented with the chloroplast and mitochondrial proteomes.

In plants, the task of assigning identified peptides to their respective proteins is not trivial. Due to the peptide-centric nature of shotgun proteomics, peptides that map to multiple proteins in a reference database can lead to ambiguous identifications. Within higher eukaryotes, this imposes a considerable challenge because shared or *degenerate* peptides, which result from segmental duplications, homologous proteins or splicing variants and comprise a large fraction of total extracted peptide library[24,25]. To date, there are different methods for aggregating MS evidence for protein assembly[26]. In a previous study, we proposed that the most advantageous framework to classify and validate protein identifications in higher eukaryotes should include the following: 1) a means to report the minimum of proteins implicated by at least one unique peptide and 2) the ability to account for database redundancies by clustering similar proteins into groups by sequence homology[6].

Using the principle of parsimony with Occam's razor constraints[27,26], 7,720 *Populus* proteins were confidently identified (classified as distinct or differentiable), and 4,520 proteins were categorized as indistinguishable (Supplemental Table 1). Although widely used, the guidelines in the suggested nomenclature make data interpretation more complicated and less accurate, especially in highly redundant proteome databases like *Populus*[6].

For this reason, we propose a strategy that incorporates additional supporting information (i.e., sequence homology) to better infer the existence of proteins. While this approach can be applied to shotgun proteomic studies of plants in general, it confers demonstrable advantages for *Populus* specifically (see Experimental). Proteins sharing 90% or more sequence identity within the *Populus* database were clustered into groups. Each protein group was defined by a single representative protein sequence called a seed, where each seed shares ≥90% sequence identity with all other members of that group. Observed peptides from the originally searched protein entries were then directly referenced back to the clustered database. For the current data set that included 63,056 tryptic peptides, ~25% were previously shared within the original *Populus* database (non-unique/degenerate) but were reclassified as unique to a particular protein group in the newly constructed database. This illustrates the advantage of implementing a "protein group-centric" approach, such that including information about sequence homology allows the interpreter to readily assess the relatedness between shared peptides of indistinguishable proteins derived from gene duplication and splice variants. Moreover, as clustered proteins are ≥90% similar to one another, members of a particular group likely exhibit similar functional roles which, when applied to semi-quantitative proteomics, allows for a more robust analysis of functional signatures across conditions, time points or organ types. In other words, this strategy effectively reduces the complexity of the functional analysis and biological interpretation of plant data.

Based on this approach, a total of 11,692 protein assignments (Supplemental Table 2) across all organ-types were reduced into 7,538 protein groups at an average false-discovery rates of <1% at the peptide level (Supplemental Table 3). Protein groups were populated by as many as 21 members, with one-membered groups (i.e., singletons) representing only 36% of the total. In total, we were able to measure 25% of the predicted proteins for *Populus*. Generating complete proteome maps of higher organisms is a difficult task as it is unlikely the entire ensemble of polypeptide species encoded by a genome will be expressed at any given time. Nevertheless, this integrated data set provides an “information backbone” that captures baseline protein expression across spatially and functionally distinct pathways. This holistic view of plant-wide protein expression will provide a better understanding of the detected components (i.e., proteins, pathways, etc.) in the context of relationships between organs.

3.2 Depth of Analysis of the *Populus* Proteome

Having established robust peptide/protein identification criteria, we sought to assess the depth of our data set by two critical figures of merit, proteome sequence coverage and dynamic range. Despite differences in organ background, similar total protein sequence coverage (median=19%) was achieved (Supplemental Fig. 1), a value comparable to recent work employing a similar approach to analyze yeast[30]. Of the four organ-types, the mature leaf proteome consisted of proteins with lower total sequence coverage. Concomitantly, there were fewer proteins with high sequence coverage. We speculate that the heterogeneity of the expected protein population expressed in mature leaf (i.e., membrane-related proteins, post-translational modifications, etc.) is perhaps less suited for the current trypsin-based schema. Transmembrane prediction using Hidden Markov Models (TMHMM)[31] analysis revealed similar identification rates of proteins with transmembrane domains (6-7% across all organs), suggesting that the systematic decrease in protein sequence coverage is more likely due to changes in the frequency of post-translational modifications or some other phenomena related to the types of proteins being expressed.

Electrospray ionization presents the mass spectrometer with a dynamic population of peptides, of which only a fraction is selected for sequencing[32]. Consequently, highly abundant peptides limit the sampling and identification of low abundant peptides. Because the LTQ Velos platform includes advances that benefit the analysis of low-abundant and low signal-to-noise precursors, we sought to quantitatively assess the capabilities of the instrument. By comparing our current data set against our previous in-depth *Populus* study[6], which used the LTQ XL platform, we examined the achievable depth of proteome characterization. When examining the distribution of the identified precursor ions versus local signal-to-noise ratios, the LTQ Velos platform increased the identification of low signal-to-noise precursor ions compared to the LTQ XL (Supplemental Fig. 2). Furthermore, peptide populations created from complex mixtures often tax the sequencing speed of MS instruments such that the mass spectrometer is incapable of targeting every eluting peptide and thus misses “sequenceable” peptides. As anticipated, the faster acquisition speed facilitated a 2-fold increase in the number of scans collected and assigned as well as the total number of proteins identified (Fig. 1A and 1B). Given these improvements provided by the Velos platform, we anticipated a sizeable increase in the analytical dynamic range. Indeed, protein dynamic range spanned 5-6 orders of magnitude, representing a 1-2 order of magnitude increase when compared to the LTQ XL platform (Fig. 1B). Together, these increases in sensitivity and speed augment the Velos' analytical dynamic range, providing demonstrably better depth of proteome characterization.

Similar experimental strategies directed towards deep proteome coverage in higher eukaryotes, like yeast, have measured a remarkably large dynamic range of protein expression[33]. Unlike yeast[34], there is no available information regarding known cellular concentrations (copies/cell) for proteins spanning the entire abundance range in *Populus*. Therefore, it is a challenge to accurately assess the biological dynamic range achieved by this approach. Nevertheless, the dual-pressure ion trap design includes substantial improvements that benefit the analysis of complex protein mixtures.

3.3 Spatial Proteomics: Profiling Organ-Specific Proteomes

A function-level view of the different *Populus* organ proteomes was generated by sorting protein groups into functional categories as defined in the eukaryotic clusters of orthologous groups (KOG) database and weighting by normalized spectral count (nSpC)[17] (Fig. 2). KOG categories of "unknown function" and "post-translational modification and chaperone" had the highest representation in all organs. With regard to specific organs, "signal transduction mechanisms" and "chloroplast components" were the most abundant functional categories in mature leaves, "translation and RNA processing" in young leaves, "cytoskeleton" in stem and "unknown function" in roots.

We next identified protein groups from our data set that overlapped different organs, as well as those that were only found in one organ (Fig. 3A). In the current study, a "core" proteome shared among the four different *Populus* organs was identified, consisting of 2,060 protein groups. The spatial distance between organs appeared to influence the degree of overlap between the different proteomes. For two organs that have a distal relationship, such as root and mature leaf, the overlap between proteomes decreases.

Protein groups found in only one organ may be linked to specialized, organ-specific processes. In total, we identified 688 protein groups unique to root, 831 to stem, 370 to mature leaf, and 629 to young leaf. For a more detailed comparison of the different organ proteomes, a Pearson correlation matrix assessed the correlation between the different organ proteomes (Fig. 3B). The pairwise comparisons resulted in Pearson correlation values that range from 0.06 (mature leaf vs. root) to 0.72 (young leaf vs. stem). The correlation coefficients support the results in Figure 3A and together corroborate the hypothesis that the degree of proteomic overlap between different *Populus* organs is reflected by their shared function.

For a network-wide perspective of *Populus* metabolism, we employed the use of iPath2.0[23] (http://pathways.embl.de) to navigate and explore the predicted KEGG metabolic pathways (Supplemental Figure 3). Using the entire data set (7,538 protein groups), a metabolic pathway diagram was constructed to highlight the core proteome relative to all protein groups measured (Supplemental Fig. 3A). Overall, the core molecular network spanned every major functional category belonging to central metabolism. These protein groups likely belong to catalytic and regulatory interactions that govern the life of a plant cell, and may include signaling networks that choreograph cross talk between plant cells in response to environmental perturbations. Supplemental Figures 3B-E depict metabolic grids of individual organ proteomes. Even though similar coverage of the metabolic network was observed for each organ, the most revealing feature of these maps is the existence of molecular networks and the protein groups that are characteristic of a specific organ. For each organ proteome, a number of unique protein groups were identified and, rather than mapping ubiquitously, they generally assembled into discrete pathways. Although beyond the scope of this study, future work could integrate metabolomics to measure net fluxes of material into and out of pathways, capturing the relationship of enzymes and their substrates/products[35].

3.4 Quantitative Analysis of *Populus* Organ Proteomes

In order to generate a quantitative proteome map of the different *Populus* organs, we first filtered the data to account for the stochastic nature of the peptide sampling process[36]. That is, a significant proportion of the data consists of low-abundant proteins and because accurate label-free quantitation is difficult to perform on low-abundant proteins, we applied a threshold filter for their subsequent removal. Rather than choosing an arbitrary abundance value to eliminate low abundance proteins, an empirical prevalence value was identified to obtain a “cut-off” criterion that distinguishes changes in protein expression from background noise and false positives[22]. Applying this filter to the entire data set reduced the number of quantifiable protein groups from 7,538 to 3,242 (see Experimental). Notably, while only ~43% of the data set remains, we retained ~98% of the total assigned nSpC values for quantitative analysis.

Using these parameters, we sought to identify the distribution of protein expression across the different organs. Hierarchical cluster analysis was applied to the 3,242 protein groups, resulting in 14 clusters that can be visualized in Figure 4. Across all clusters, the number of protein groups ranged from 395 (cluster 7) to 74 (cluster 12) (Supplemental Table 4). While cluster membership reflects the relative diversity of protein function, the overall activity of each cluster is revealed through the relative percentages of the total assigned spectra (Fig. 5A). With such values, the quantitative representation of each organ within a cluster can be defined.

To interpret the biological significance of each cluster, cluster membership was plotted against KOG category (Fig. 5B). First, we examined the protein groups that were predominately expressed in only one organ: cluster 1 (root), cluster 2 (mature leaf), cluster 8 (stem), and cluster 14 (young leaf). For the set of protein groups that were predominantly expressed in roots, the three most abundant functional categories observed were “unknown function”, “post-translation modification and chaperones”, and “amino acid transport and metabolism”. For those protein groups whose function remains unknown, an attempt to elucidate a biological role was dependent on whether a protein could be associated with a particular protein family in the Pfam database[37]. Although functional annotations based solely on family membership must be interpreted with caution, high-quality association with a protein family would, in fact, indicate what functional units are present and thus suggest a biological role. By investigating protein family membership, a functional role for the two most abundant proteins (POPTR_0013s10350.1 and POPTR_0013s10380.1) in the unknown category could be hypothesized. When searched against the Pfam database, both proteins matched with high confidence to a phosphorylase superfamily that includes 5'-methylthioadenosine phosphorylase. A previous publication[38] suggests that the ortholog represented in *A. thaliana* responds to changes in the level of cytokinin production in various cell types. Within plants, the phytohormone-related cytokinin is an important regulator of plant growth and, notably, 5'-methylthioadenosine, the substrate for the above-mentioned protein, has been linked to cytokinin metabolism[39]. For another set of proteins (POPTR_0008s13030.1 and POPTR_0008s13040.1), which are also among the highest expressed in the unknown category, *A. thaliana* orthologs have been shown to correlate with cytokinin levels in roots[40]. A search against the Pfam database resulted in high confident annotations for both proteins, matching against the Bet v I allergen protein family. In fact, members of the Bet v I allergen protein family are storage proteins that occur across dicotyledonous plants[41] and have been shown to be cytokinin-binding proteins[42]. Together, the expression patterns observed, as well as annotations provided through protein family memberships, suggest biological roles impacting multiple aspects of plant development, including cell growth and sink/source relationships.

Within the set of protein groups that were predominately expressed in stem (cluster 8), the three most abundant KOG functional categories observed were “unknown function”, “cytoskeleton”, and “amino acid transport and metabolism”. A set of the most abundant proteins shared a common biological thread; they all are involved in cell wall formation. Among this set, UDP-glucose pyrophosphorylase (UGPase; POPTR_0004s07280.1) and UDP-glucose dehydrogenase (UGDH; POPTR_0004s11760.1) were identified at similar abundance values. In plants, the enzyme UGPase is metabolically positioned at the point of sucrose synthesis/breakdown and, at this important carbon flow junction, produces UDP-glucose, which is required for sucrose synthesis or other polysaccharides, such as hemicellulose or pectin[43]. Because actively growing stem tissues (i.e., phloem and xylem) do not serve a nutritional storage role, we hypothesize that the enzyme UGDH utilizes the UDP-glucose produced by UGPase to form UDP-glucuronate, which is a precursor for hemicellulose and pectin formation, two xylem-related polymers. Lastly, the protein with the largest abundance value is an actin depolymerizing factor (ADF; POPTR_0009s03320.1) that perhaps plays a role in control of woody tissue development. In *Populus*, ADF activity is thought to be essential for the development of phloem and xylem[44].

The three most abundant KOG functional categories observed in mature leaf (cluster 2) were “unknown function”, “chloroplast”, and “carbohydrate transport and metabolism”. A mature leaf harbors a highly active network of chloroplasts, an organelle where light energy is collected and converted into stored chemical energy that ultimately fixes atmospheric carbon dioxide to carbohydrates. Hence, fully developed leaves possess the highest photosynthetic rate, chlorophyll accumulation and respiration levels[45]. Indeed, the three most abundant proteins (ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) large subunit; Chloroplast 11241, PSAD photosystem I reaction center subunit; POPTR_0008s15100.1 and glyceraldehyde 3-phosphate dehydrogenase; POPTR_0014s13660.1) observed in the data set reflect the main purpose of this specialized organ and substantiates its characteristic role in photosynthesis and carbohydrate metabolism.

The three most abundant KOG functional categories observed in young leaf (cluster 14) were “unknown function”, “translation”, and “RNA processing and modification”. Unlike mature, fully expanded leaves, the more juvenile leaves appear to be ontogenetically closer to the shoot apical meristem. Rather, they appear to utilize most of their resources in active growth and development. The two most abundant proteins in young leaf were of unknown function but by investigating protein family membership with Pfam, both proteins (POPTR_0124s00210.1 and POPTR_0018s09610.1) matched with high confidence to members of the GDSL-like Lipase/Acylhydrolase family. In general, the GDSL-like lipase superfamily is thought to play an important role in the regulation of plant development and, recently, in the metabolism of cutin and wax[46,47]. In plants, cutin biosynthesis is a crucial component in the formation of outermost epidermal cell wall surface, the cuticle. Within expanding young leaves, the enzymatic mechanisms involved in the production of cutin monomers have been well-studied[48,49]. However, little progress has been made in identifying the enzymes involved in the transportation and building of the cutin matrix within the epidermal cell extracellular matrix. Lipase-type enzymes have been suggested to be involved in the cutin polymerization step within the extracellular matrix[50]. The results here, in correlation with the ANOVA analysis below, suggest that these abundant lipase proteins are involved in the formation of the plant cuticle.

3.5 Quantitative Analysis of *Populus* Leaf Development

The semi-quantitative power of LC-MS/MS-based proteomics was employed to detail the proteomic differences between *Populus* leaf at two different developmental stages -- young (YL) and fully expanded, mature leaf (ML). To accomplish this task, protein group normalized spectral counts (nSpC) collected across all replicates of each leaf-type were analyzed by one-way ANOVA (see Experimental). Protein groups represented in the analysis include only those with significant sample-to-sample representation, as assessed by prevalence value (PV; see above). In total, 2881 protein groups from both young and mature leaves were statistically compared (Supplemental Table 5). Roughly half (1432 protein groups) were found to exhibit statistically significant ($p \leq 0.01$) differential abundance patterns, with 395 groups showing increased abundance in mature leaf compared to 1037 in young leaf (Fig. 6A and B). These values support the proposition above that mature leaf has "settled" into its organ-specific function (photosynthesis) and thus requires a reduced complement of proteins, relative to young leaf. In contrast, young leaves are still developing, as evidenced by the up-regulation of general biosynthetic pathways (i.e., DNA synthesis, transcription, translation, etc.).

In order to better visualize the functional differences between young and mature leaves, differentially abundant protein groups from the ANOVA analysis (Supplementary Table 4) were mapped to KEGG-derived metabolic pathways using iPath2.0. Only those *Populus* protein groups with assigned function (i.e., KEGG KO or Enzyme EC number) could be mapped, leaving out several highly abundant, differentially expressed proteins of unknown function. Despite this limitation, developmentally responsive protein groups matched to 1444 metabolic map elements (392 in mature leaf vs. 1052 in young leaf, redundant entries included) and allow for a more pathway-centric view of the functional reactions specific to leaf developmental stage (Fig. 7A and B).

Functional enzymes exhibiting differential abundance patterns are highlighted on the pathway maps as varying degrees of either red (up-regulated in mature leaf, Fig. 7A) or green (up-regulated in young leaf, Fig. 7B), depending on their fold change. Protein expression that differs by a factor of 10 or more in either direction is represented by the brightest of either color. As a testimony to the power and accuracy of semi-quantitative proteomics, LC-MS/MS-derived protein abundance patterns highlight several contiguous pathway components, a majority of which respond in an appropriate, concerted fashion specific to leaf developmental stage. This pathway-centric view thus expands upon a general list of up- and down-regulated proteins, allowing for more complete synthesis of systems biological information. However, this is not to suggest that the latter is unnecessary, especially as only a subset of leaf stage-responsive proteins could be effectively mapped to a particular metabolic pathway.

3.5.1 Metabolic Pathway Mapping of Mature Leaf Highlights a Primary Focus on Energy Harvesting

Protein groups exhibiting increased abundance in mature leaf, relative to young leaf, substantiate the general view of a leaf as a specialized energy harvesting organ. As highlighted in box “PS” (Fig. 7A), all major components of photosynthesis (KEGG pathway KO00195) show significant increased protein abundance in mature leaf relative to young leaf. In fact, photosynthesis is one of the most up-regulated systems present in mature leaf, an observation further evidenced by totaling the nSpC of each of the four sub-complexes: 1) Photosystem II – 5324 nSpC in mature leaf compared to 907 in young leaf (5.87x up-regulated), 2) Photosystem I – 7801 to 624 nSpC (12.50x up-regulated), 3) Cytochrome b6/f complex – 1873 to 370 nSpC (5.06x up-regulated) and 4) ATP synthase – 5435 to 762 nSpC (7.13x up-regulated). Furthermore, photosynthetic antenna proteins, the chlorophyll-binding components of light harvesting complexes 1 and 2 (KEGG pathway KO00196), also showed a 4.6x increase in abundance compared to young leaf (2469 vs. 534 nSpC). Taken together, mature leaf photosynthetic function was up-regulated by a factor of 7.2x relative to young leaf (22,902 vs. 3,197 nSpC).

Photosynthesis is inextricably linked to carbon fixation, a process by which photonic energy harnessed from sunlight is used to replenish supplies of NADPH and ATP, both of which power the redox-based reduction of atmospheric carbon dioxide to sugar molecules. Thus, the observed increase in proteins related to photosynthesis in mature leaf must correspond to an increase in the rate of carbon fixation. As follows, enzymes relevant to the carbon fixation pathway (KO00710), which are highlighted in box “CF” (Fig. 7A), exhibit increased abundance in mature leaf relative to young leaf. By totaling the nSpC of the proteins involved in Calvin cycle, C3-based carbon fixation activity is up-regulated by a factor of roughly 6x (16,982 vs. 2818 nSpC in mature and young leaf, respectively), a value that is in line with the degree of photosynthesis up-regulation reported above. Furthermore, RuBisCO, the key enzyme in carbon fixation, accounted for 9,397 nSpC across both mature and young leaves, but was enriched over 4.6x in mature leaf (7,721 vs. 1,676 nSpC).

Though the primary function of RuBisCO is to fix atmospheric CO_2 to ribulose 1,5-bisphosphate (RuBP) through its carboxylase activity, this enzyme can also function as an oxygenase in a process termed photorespiration (PR). In this regard, O_2 rather than CO_2 is assimilated, leading to the production of 3-phosphoglycerate (3PG) and 2-phosphoglycolate (2PG), the latter of which must be metabolized to 3PG for re-entry into the Calvin Cycle. This complicated, multi-organelle pathway, however, equates to an expenditure of metabolic energy to both convert 2PG to 3PG and to recapture carbon (CO_2) and nitrogen (NH_4^+) lost in the process. Although the rate of photorespiration is exacerbated under hot/dry conditions, it occurs at substantial rates (~25%) even under moderate growth conditions[51]. As highlighted in box "PR" (Fig. 7A), there was up-regulation of the photorespiration pathway (glyoxylate and dicarboxylate metabolism, KO00630), starting with RuBisCO's oxygenase-dependent production of 2PG through its conversion to glycerate (2PG → glycolate → glyoxylate → glycine → serine → hydroxypyruvate → glycerate) and involving the necessary accessories pathways/enzymes (catalase, glutamine/glutamate cycle and tetrahydrofolate cycle) to complete the process. The only enzyme within the pathway not identified to be up-regulated was glycerate kinase. In total, PR in mature leaf was up-regulated by a factor of 5x (11,230 vs. 2251 nSpC) relative to young leaf.

These three major mature leaf-enriched metabolic pathways constitute a proof-of-concept with regard to the LC-MS/MS platform described in this paper. As mentioned above, and further corroborated by these proteomic data, mature leaf appears to have "settled" into its primary function. Other less complete pathways were also found to be up-regulated, with a portion of them seemingly involved in reacting to oxygenic stress, most likely induced by the photosynthetic process itself. For example, L-ascorbate peroxidase (EC:1.11.1.11, KO00434) was up-regulated in mature leaf by 3.9x (223 vs. 57 nSpC) while a 2-cysteine peroxiredoxin (EC: 1.11.1.15, KO03386) was up-regulated by a factor of 2.6x (631 vs. 240 nSpC). Both enzymes are known to detoxify reactive oxygen species, with the latter previously shown to be targeted to chloroplasts to provide prevent photooxidative damage to the photosynthetic membrane[52,53]. Furthermore, three enzymes in the pathway for carotenoid and xanthophyll biosynthesis were also slightly up-regulated in mature leaf (lycopene beta-cyclase [KO06443], zeaxanthin epoxidase [KO09838] & 9-cis-epoxycarotenoid dioxygenase [KO9840]) with modest nSpC differences. In fact, the spectral counts from mature to young leaf totaled 26 vs. 0, supporting the premise that the mature leaf has enhanced photosynthetic capabilities, including pigments that modulate harvested photonic energy in periods of high light and act as antioxidants, playing lipid-protective roles in periods of high photosynthetic/oxygen generating activity[54,55].

3.5.2 Metabolic Pathway Mapping of Young Leaf Highlights a Primary Focus on Growth and Development

Young leaf engages in most, if not all, of the photosynthetic-dependent pathways detailed above, albeit with reduced abundance relative to mature leaf. The reduced level of photosynthetic-related proteins is countered by a systematic increase in protein abundance in several major general biosynthetic pathways, consistent with the fact that a young leaf is primarily actively growing and secondarily photosynthesizing. This perhaps intuitive observation is apparent upon viewing the pathway map in Figure 7B. Relative to mature leaf, young leaf shows increases in several metabolic pathways including terpenoid biosynthesis ("TB" | KO00900 | ~2.3x | 177 vs. 78 nSpC), flavonoid biosynthesis ("FB" | KO00941 | ~1.5x | 58 vs. 39 nSpC), pyruvate metabolism ("PM" | KO00620 | ~1.6x | 1443 vs. 888 nSpC), TCA cycle flux ("TCA" | KO00020 | ~2.0x | 1282 vs. 630 nSpC), fatty-acid metabolism ("FA" | KO00061, KO00062, KO00071 | ~3.3x | 321 vs. 98 nSpC) and nucleotide metabolism ("NM" | KO00230, KO00240 | ~3.3x | 735 vs. 221 nSpC).

Although each of these pathways, comprised of their respective pools of differentially expressed proteins, were found to be up-regulated in young leaf, their spectral representation is less abundant relative to observation for regulatory pathways, including transcription (KO03020, KO03022, KO03040), translation (KO03010, KO00970, KO03013, KO03015, KO03008) and protein folding/sorting/degradation (KO03050, KO03018, KO03060, KO04120, KO04141, KO04130, KO04122). Proteins involved in the *Populus* translational apparatus were by far the most abundant in young leaf, exhibiting a ~3.6x increase in abundance (15,691 vs. 4,342 nSpC) relative to mature leaf. This increase was corroborated by significant increases in both transcription (~3.0x, 1,875 vs. 630 nSpC) and protein folding/processing (~3.6x, 6,327 vs. 1,731 nSpC). Taken together, and including a modest increase observed in DNA metabolism (96 vs. 4 nSpC), young leaf regulatory pathways were up-regulated by a factor of roughly 3.6x (23,989 vs. 6,707 nSpC). The measured increase in protein abundance observed for members of these specific regulatory pathways solidify the general observation that young leaf's primary function is growth that is fueled by moderate levels of photosynthesis and carbon fixation.

CONCLUSIONS

Since the release of the *Populus* genome in 2006, a question remains unanswered, i.e., what is the achievable depth and coverage of the predicted proteome space of *Populus* using high-throughput mass spectrometry? Although the application of shotgun proteomics to measure global molecular responses is successful for many proteomic samples of low complexity, the depth of coverage required for a similar inquiry in higher eukaryotes requires more sophisticated sample preparation and advanced instrumentation. Therefore, we sought to address these issues by implementing a myriad of optimizations for nearly every step of the experimental process. These optimizations, while beneficial to plant proteomics in general, are broadly applicable to other organisms of similar complexity such as humans and other higher eukaryotes.

The enormous biological dynamic range inherent to a eukaryotic system demanded incorporation of a detergent-based sample preparation strategy that enhances plant cell lysis and protein extraction, both crucial enablers for in-depth analyses of complex proteomes. Without the appropriate instrumentation, this complexity inevitably leads to a sub-optimal identification of all detectable peptide species. Although a longstanding general challenge in shotgun proteomic experiments, recent technological improvements to the LTQ platform, mainly through enhancements to sequencing speed and sensitivity, doubles the identification rate of these dense peptide populations and enhances the identification of low-abundant protein species. Taken together, the enhanced sample preparation method and the state-of-the-art instrumentation enabled us to achieve one of the deepest proteome analyses in plant organisms to date, spanning six orders of magnitude in protein abundance and requiring only modest levels of sampling (i.e., 5-6 sample replicates per organ).

As demonstrated, the depth of coverage achieved in this study was sufficient for comprehensive characterization of different plant organs that, at the cellular level, have vastly different chemical backgrounds (expressed genes, proteins and metabolites). The streamlined approach applied here affords an unprecedented view of *Populus* protein expression across several major organ-types, with each offering unique proteomic perspectives. This not only facilitates the identification of functional processes unique to a particular organ but also allowed us to define a *Populus* core proteome. Furthermore, in addition to identifying proteins with a known biological function, a large percentage of each proteome consisted of proteins with no known function. Though specific biological roles were not determined for these proteins in this present study, general observations (i.e., organ location, differential regulation, etc.) outlined here provide hypotheses for further interrogation.

In addition to providing qualitative data describing the protein complement of each organ, the collected data also contains semi-quantitative information, reflecting the underlying functional processes and mechanisms in each organ-type as weighted by a conservative estimate of protein abundance. To our knowledge, a systems-wide survey of protein abundances across different *Populus* organs has not been reported. By unveiling several perspectives of protein richness: 1) network-wide 2) discrete pathways and 3) individual proteins, this quantifiable, holistic view of the protein network enables reliable predictions to be made regarding the response of the cellular system to environmental perturbations and experimental manipulations. To demonstrate the precision and comprehensiveness of this approach, we explored proteomic differences between the same organ-type during two growth stages, young and mature leaf. As detailed above, mature leaf appears to function primarily in an autotrophic role consisting of energy generation via its photosynthetic apparatus and reduction of CO_2 to sugar via its carbon fixation pathway. Though other photosynthesis-related pathways were up-regulated in mature leaf (i.e., photorespiration and defense against photosynthetically-derived reactive-oxygen species), proteins involved in photosynthesis and carbon fixation constitute the majority of the quantitative signal. On the other hand, biosynthetic and regulatory functions were relatively up-regulated in young leaf. Even though proteins/pathways for photosynthesis, carbon fixation and photorespiration were detected in young leaf, they were less represented relative to mature leaf. This information suggests that young leaves partition resources between growth and energy production. These observations and data provide a “proof-of-concept” with regard to our

2D-LC-MS/MS platform and suggest the biological validity of these pathway-centric comparisons, opening the door for future hypothesis-driven inquiries into *Populus* and other complex organisms. Obviously caution must be exercised when interpreting these semi-quantitative results, as only one biological replicate was available for statistical assessment. Clearly the inclusion of more biological replicates (3-5) would improve the statistical framework of this discovery-based approach. Nevertheless, the workflow discussed here provides an intellectual springboard for future targeted, quantitative approaches.

Acknowledgement: We thank Manesh Shah for database searching and informatics. This study was funded within the BioEnergy Science Center, a U.S. Department of Energy Bioenergy Research Facility supported by the Office of Biological and Environmental Research, Genome Sciences program, in the DOE Office of Science. University of Tennessee-Battelle LLC for the Department of Energy manages Oak Ridge National Laboratory**.**

Figure Legends:

Figure 1: Peptide and protein dynamic range. Dynamic range of measurement was assessed for each identified peptide and protein across three replicate runs for all four organ types. (A) Maximum ion intensity values obtained for each peptide's extracted ion chromatogram (y-axis) were ranked by intensity and plotted against cumulative number of assigned MS/MS spectra (x-axis). Curves represent individual replicates per organ to identify run-to-run differences in dynamic range. (B) Assembled protein intensity (y-axis), calculated by summing constituent peptide intensities across all replicates, was plotted against the cumulative number of identified proteins (x-axis) to identify the overall protein dynamic range achieved per individual organ. Dynamic range values, represented as magnitudes (base 10), are listed in the figure legend. Light blue: LTQ stem; blue: LTQ-Velos mature leaf; red: LTQ-Velos young leaf; purple: LTQ-Velos root; green: LTQ-Velos stem.

Figure 2: Quantitative distribution of detected protein groups by their functional classification. Proteins identified in each tissue type were assigned KOG categories to identify functional trends relevant to each organ type. Category representation was weighted by the sum total of the normalized spectral counts (nSpC) contributed by each protein in the classification. Notable trends include a high proportion of nSpC in mature leaf attributed to chloroplast-based proteins, enrichment of cytoskeletal components in stem, and an increase in translation in young leaf compared to the other tissues. Also noted is the large degree of nSpC representation falling into the unknown category, suggesting a need for improved protein annotation as a whole.

Figure 3: Global proteomic view across all four organs. Numbers of identified protein groups, as represented by a 4-tiered Venn diagram (A), indicate the level of proteomic overlap between organ types. Notable regions include protein groups specific to only one organ-type (solid blue, yellow, green and red) as well as groups identified across all organs (central brown region). (B) Degree of proteomic overlap as visualized by Pearson's correlation analysis of all protein nSpC values averaged across all replicates for each particular organ. The degree of correlation increases as a function of organ proximity.

Figure 4: Hierarchical clustering classifies protein groups by distinct localization trends. Identified protein groups above the determined prevalence value were clustered into groups based on nSpC abundance patterns across all organ types. Abundance values, ranging from -1.56 to +1.56, were calculated by converting nSpC for each protein group, averaged across all replicates, to a value representing the number of standard deviations away from the row mean. Protein groups sharing similar standardized abundance trends were then clustered into distinct families (listed top to bottom - 1, 12, 4, 8, 7, 2, 13, 10, 3, 11, 6, 14, 9 and 5) and denoted in alternate colors. Columns representing each organ-type were then clustered (bottom) based on global data set similarities.

Figure 5: Quantitative distribution of detected protein groups by their functional classification for each hierarchical cluster. Protein group clusters were deconvoluted by organ-type (A) to show each organ's nSpC contribution relative to the total nSpC populating each cluster (across all organs). Table cells are color-coded based on

percent contribution (green:red::low:high) in order to quickly visualize each organ's share of the total nSpC. (B) To view the functional signature of each cluster (z-axis), cluster members were classified into their respective KOG categories (x-axis), with each category's representation weighted, based on the sum of nSpC of contributing protein group members (y-axis).

Figure 6: Differential proteomic analysis of young versus mature leaf by ANOVA. Protein groups identified in young and/or mature leaf above the determined prevalence value were analyzed by ANOVA to compare the functional signature between two distinct developmental stages of leaf. (A) Protein group abundances (nSpC), averaged across all replicates (n=6) per stage, were compared between young leaf (YL, x-axis) and mature leaf (ML, y-axis) and visualized as a scatterplot. Protein groups that showed a significant ($p \leq 0.01$) difference in abundance are colored red. Dotted lines separate "effectively zero" sub-distributions from the main distribution in the top right quadrant. Proteins groups in this main distribution were identified in both developmental stages while proteins in the sub-distributions were likely found in only one stage. To further visualize the statistical metrics of the main distribution, a volcano plot (B) was constructed, comparing the LOG2(nSpC)-based difference between both developmental stages (x-axis) to the level of statistical significance, represented as -LOG10(p-value) (y-axis). As in (A), protein groups that showed a significant ($p \leq 0.01$) difference in abundance are colored red.

Figure 7: Up-regulated metabolic pathways as dictated by *Populus* leaf developmental stage. Proteins exhibiting differential abundance patterns (ANOVA; p ≤

0.01) across both young and mature leaf were mapped to KEGG pathways using iPath v.2.0 and color-coded to indicate the degree of protein abundance differences between each developmental stage. (A) Proteins with significantly increased abundance in mature leaf are labeled in red with brighter shades indicative of larger differences. Highlighted pathways (dashed boxes) include photosynthesis (PS), carbon fixation (CF), and photorespiration (PR). (B) Proteins with significantly increased abundance in young leaf are labeled in green with brighter shades indicative of larger differences. Highlighted pathways include nucleotide metabolism (NM), flavonoid biosynthesis (FB), fatty acid metabolism (FA), pyruvate metabolism (PM), terpenoid biosynthesis (TB), and TCA cycle (TCA).

References:

1. Ahrens, C. H., E. Brunner, et al. (2010). "Generating and navigating proteome maps using mass spectrometry." Nat Rev Mol Cell Biol **11**(11): 789-801.
2. Tuskan, G. A., S. DiFazio, et al. (2006). "The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)." Science **313**(5793): 1596-1604.
3. Plomion, C., C. Lalanne, et al. (2006). "Mapping the proteome of poplar and application to the discovery of drought-stress responsive proteins." Proteomics **6**(24): 6509-6527.
4. Kalluri, U. C., G. B. Hurst, et al. (2009). "Shotgun proteome profile of Populus developing xylem." Proteomics **9**(21): 4871-4880.
5. Shuford, C. M., Q. Li, et al. (2012). "Comprehensive Quantification of Monolignol-Pathway Enzymes in Populus trichocarpa by Protein Cleavage Isotope Dilution Mass Spectrometry." J Proteome Res.
6. Abraham, P., R. Adams, et al. (2012). "Defining the boundaries and characterizing the landscape of functional genome expression in vascular tissues of Populus using shotgun proteomics." J Proteome Res **11**(1): 449-460.
7. Shevchenko, A., M. Wilm, et al. (1996). "Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels." Anal Chem **68**(5): 850-858.
8. Washburn, M. P., D. Wolters, et al. (2001). "Large-scale analysis of the yeast proteome by multidimensional protein identification technology." Nat Biotechnol **19**(3): 242-247.
9. Botelho, D., M. J. Wall, et al. (2010). "Top-down and bottom-up proteomics of SDS-containing solutions following mass-based separation." J Proteome Res **9**(6): 2863-2870.
10. Wang, W., F. J. Tai, et al. (2008). "Optimizing protein extraction from plant tissues for enhanced proteomics analysis." Journal of Separation Science **31**(11): 2032-2039.
11. Wisniewski, J. R., A. Zougman, et al. (2009). "Universal sample preparation method for proteome analysis." Nature Methods **6**(5): 359-362.
12. Corthals, G. L., V. C. Wasinger, et al. (2000). "The dynamic range of protein expression: a challenge for proteomic research." Electrophoresis **21**(6): 1104-1115.
13. Peck, S. C. (2005). "Update on proteomics in Arabidopsis. Where do we go from here?" Plant Physiol **138**(2): 591-599.
14. Baerenfaller, K., J. Grossmann, et al. (2008). "Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics." Science **320**(5878): 938-941.
15. Piques, M., W. X. Schulze, et al. (2009). "Ribosome and transcript copy numbers, polysome occupancy and enzyme dynamics in Arabidopsis." Mol Syst Biol **5**: 314.
16. Second, T. P., J. D. Blethrow, et al. (2009). "Dual-pressure linear ion trap mass spectrometer improving the analysis of complex protein mixtures." Anal Chem **81**(18): 7757-7765.
17. Giannone, R. J., H. Huber, et al. (2011). "Proteomic Characterization of Cellular and Molecular Processes that Enable the Nanoarchaeum equitans-Ignicoccus hospitalis Relationship." PLoS One **6**(8): e22942.
18. Eng, J. K., A. L. Mccormack, et al. (1994). "An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database." Journal of the American Society for Mass Spectrometry **5**(11): 976-989.

19. Tabb, D. L., W. H. McDonald, et al. (2002). "DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics." J Proteome Res **1**(1): 21-26.

20. Edgar, R. C. (2010). "Search and clustering orders of magnitude faster than BLAST." Bioinformatics **26**(19): 2460-2461.

21. Zybailov, B., A. L. Mosley, et al. (2006). "Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae." J Proteome Res **5**(9): 2339-2347.

22. Lochner, A., R. J. Giannone, et al. (2011). "Label-free quantitative proteomics for the extremely thermophilic bacterium Caldicellulosiruptor obsidiansis reveal distinct abundance patterns upon growth on cellobiose, crystalline cellulose, and switchgrass." J Proteome Res **10**(12): 5302-5314.

23. Yamada, T., I. Letunic, et al. (2011). "iPath2.0: interactive pathway explorer." Nucleic Acids Res **39**(Web Server issue): W412-415.

24. Black, D. L. (2000). "Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology." Cell **103**(3): 367-370.

25. Delalande, F., C. Carapito, et al. (2005). "Multigenic families and proteomics: extended protein characterization as a tool for paralog gene identification." Proteomics **5**(2): 450-460.

26. Nesvizhskii, A. I. and R. Aebersold (2005). "Interpretation of shotgun proteomic data - The protein inference problem." Molecular & Cellular Proteomics **4**(10): 1419-1440.

27. Yang, X. Y., V. Dondeti, et al. (2004). "DBParser: Web-based software for shotgun proteomic data analyses." Journal of Proteome Research **3**(5): 1002-1008.

28. Nesvizhskii, A. I., A. Keller, et al. (2003). "A statistical model for identifying proteins by tandem mass spectrometry." Anal Chem **75**(17): 4646-4658.

29. Zhang, B., N. C. VerBerkmoes, et al. (2006). "Detecting differential and correlated protein expression in label-free shotgun proteomics." Journal of Proteome Research **5**(11): 2909-2918.

30. Nagaraj, N., N. A. Kulak, et al. (2012). "System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap." Mol Cell Proteomics **11**(3): M111 013722.

31. Krogh, A., B. Larsson, et al. (2001). "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes." J Mol Biol **305**(3): 567-580.

32. Kuster, B., M. Schirle, et al. (2005). "Scoring proteomes with proteotypic peptide probes." Nat Rev Mol Cell Biol **6**(7): 577-583.

33. Thakur, S. S., T. Geiger, et al. (2011). "Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation." Mol Cell Proteomics **10**(8): M110 003699.

34. Picotti, P., B. Bodenmiller, et al. (2009). "Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics." Cell **138**(4): 795-806.

35. Kruger, N. J. and R. G. Ratcliffe (2012). "Pathways and fluxes: exploring the plant metabolic network." J Exp Bot **63**(6): 2243-2246.

36. Liu, H., R. G. Sadygov, et al. (2004). "A model for random sampling and estimation of relative protein abundance in shotgun proteomics." Anal Chem **76**(14): 4193-4201.

37. Punta, M., P. C. Coggill, et al. (2012). "The Pfam protein families database." Nucleic Acids Res **40**(Database issue): D290-301.

38. Siemens, J., I. Keller, et al. (2006). "Transcriptome analysis of Arabidopsis clubroots indicate a key role for cytokinins in disease development." Mol Plant Microbe Interact **19**(5): 480-494.

39. Ravanel, S., B. Gakiere, et al. (1998). "The specific features of methionine biosynthesis and metabolism in plants." Proc Natl Acad Sci U S A **95**(13): 7805-7812.

40. Guelette, B. S., U. F. Benning, et al. (2012). "Identification of lipids and lipid-binding proteins in phloem exudates from Arabidopsis thaliana." J Exp Bot **63**(10): 3603-3616.

41. Wen, J., M. Vanek-Krebitz, et al. (1997). "The potential of Betv1 homologues, a nuclear multigene family, as phylogenetic markers in flowering plants." Mol Phylogenet Evol **8**(3): 317-333.

42. Fujimoto, Y., R. Nagata, et al. (1998). "Purification and cDNA cloning of cytokinin-specific binding protein from mung bean (Vigna radiata)." Eur J Biochem **258**(2): 794-802.

43. Kleczkowski, L. A., M. Geisler, et al. (2004). "UDP-glucose pyrophosphorylase. An old protein with new tricks." Plant Physiol **134**(3): 912-918.

44. Cseke, L. J., N. Ravinder, et al. (2007). "Identification of PTM5 protein interaction partners, a MADS-box gene involved in aspen tree vegetative development." Gene **391**(1-2): 209-222.

45. Bohler, S., K. Sergeant, et al. (2010). "Differential impact of chronic ozone exposure on expanding and fully expanded poplar leaves." Tree Physiology **30**(11): 1415-1432.

46. Broun, P., P. Poindexter, et al. (2004). "WIN1, a transcriptional activator of epidermal wax accumulation in Arabidopsis." Proc Natl Acad Sci U S A **101**(13): 4706-4711.

47. Volokita, M., T. Rosilio-Brami, et al. (2011). "Combining comparative sequence and genomic data to ascertain phylogenetic relationships and explore the evolution of the large GDSL-lipase family in land plants." Mol Biol Evol **28**(1): 551-565.

48. Nawrath, C. (2006). "Unraveling the complex network of cuticular structure and function." Curr Opin Plant Biol **9**(3): 281-287.

49. Reina, J. J., C. Guerrero, et al. (2007). "Isolation, characterization, and localization of AgaSGNH cDNA: a new SGNH-motif plant hydrolase specific to Agave americana L. leaf epidermis." J Exp Bot **58**(11): 2717-2731.

50. Kurdyukov, S., A. Faust, et al. (2006). "The epidermis-specific extracellular BODYGUARD controls cuticle development and morphogenesis in Arabidopsis." Plant Cell **18**(2): 321-339.

51. Peterhansel, C. and V. G. Maurino (2011). "Photorespiration redesigned." Plant Physiol **155**(1): 49-55.

52. Baier, M. and K. J. Dietz (1997). "The plant 2-Cys peroxiredoxin BAS1 is a nuclear-encoded chloroplast protein: its expressional regulation, phylogenetic origin, and implications for its specific physiological function in plants." Plant J **12**(1): 179-190.

53. Baier, M. and K. J. Dietz (1999). "Protective function of chloroplast 2-cysteine peroxiredoxin in photosynthesis. Evidence from transgenic Arabidopsis." Plant Physiol **119**(4): 1407-1414.

54. Havaux, M. (1998). "Carotenoids as membrane stabilizers in chloroplasts." Trends in Plant Science **3**(4): 147-151.

55. Niyogi, K. K. (1999). "Photoprotection revisited: Genetic and molecular approaches." Annual Review of Plant Physiology and Plant Molecular Biology **50**: 333-359.

Figure 1:

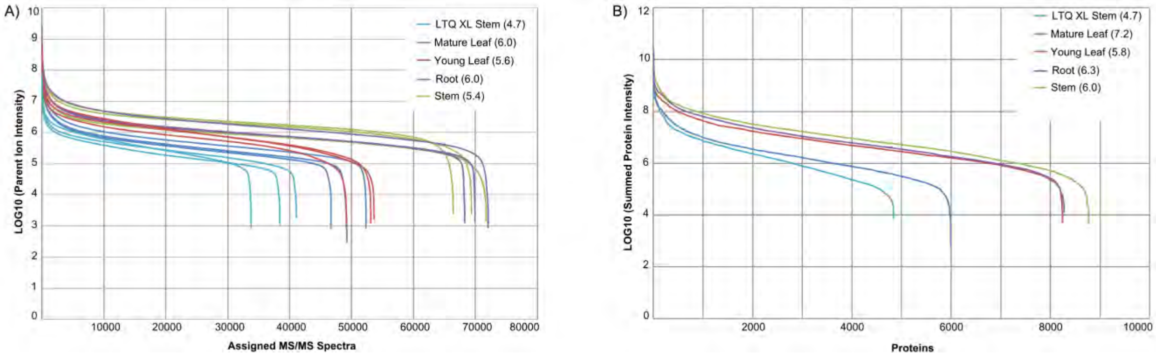


Figure 2:

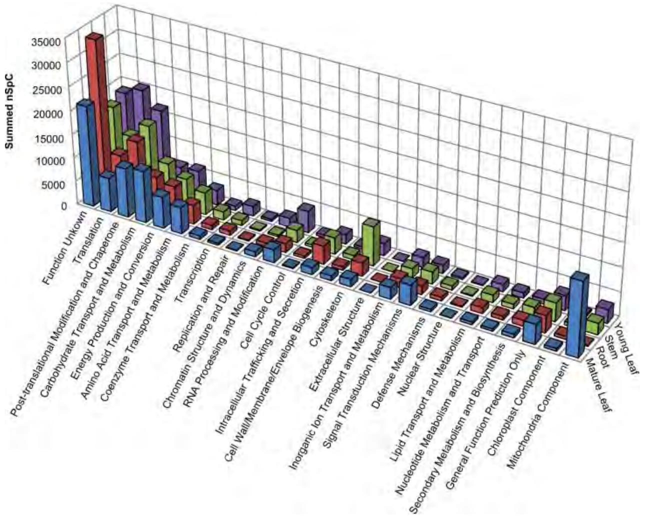
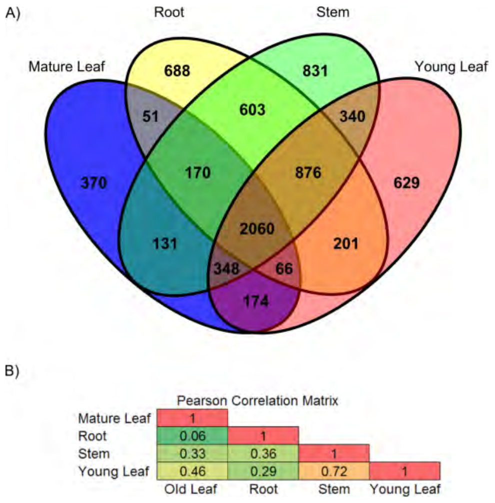


Figure 3:



Pearson Correlation Matrix

Mature Leaf	1			
Root	0.06	1		
Stem	0.33	0.36	1	
Young Leaf	0.46	0.29	0.72	1
	Old Leaf	Root	Stem	Young Leaf

Figure 4:

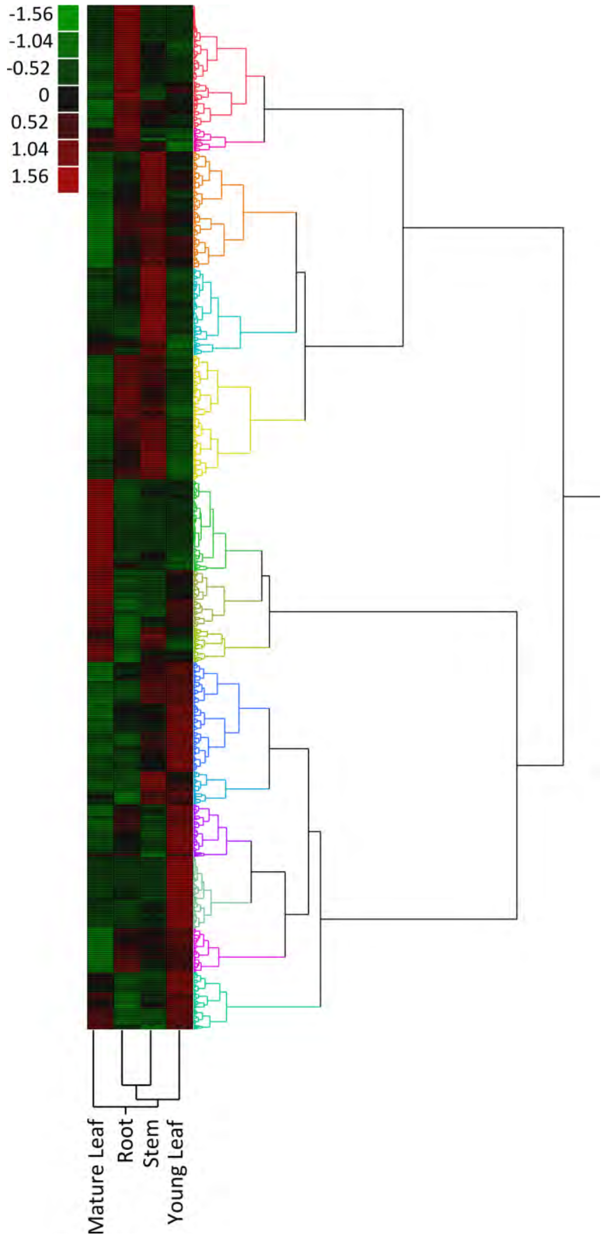
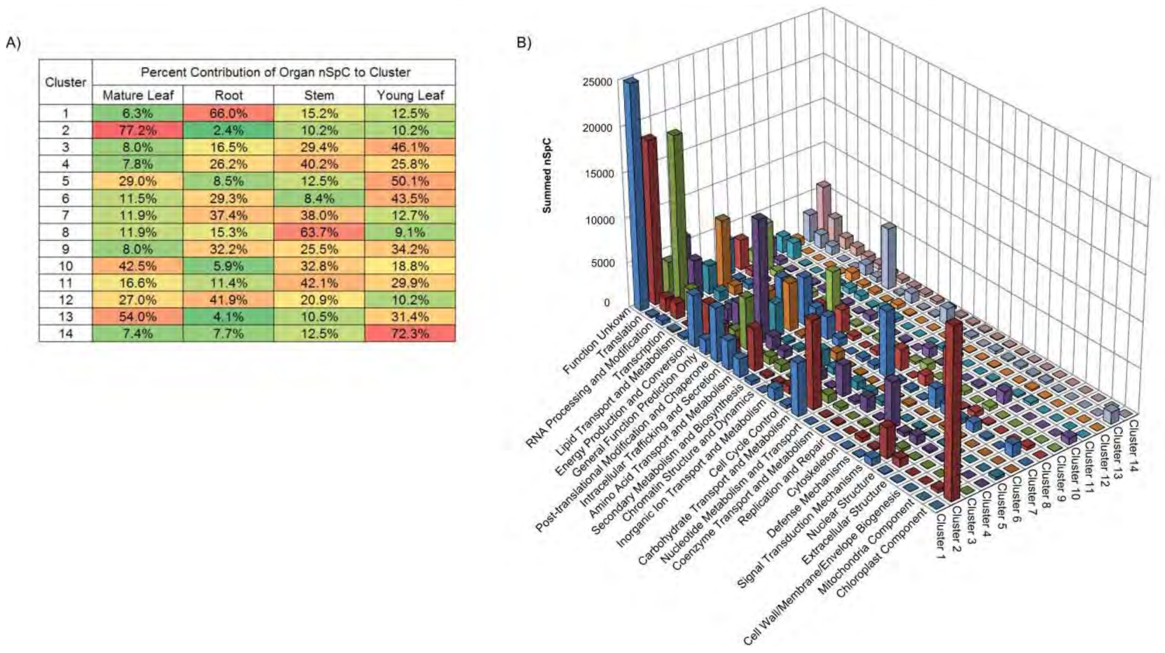


Figure 5:



A)

Cluster	Percent Contribution of Organ nSpC to Cluster			
	Mature Leaf	Root	Stem	Young Leaf
1	6.3%	66.0%	15.2%	12.5%
2	77.2%	2.4%	10.2%	10.2%
3	8.0%	16.5%	29.4%	46.1%
4	7.8%	26.2%	40.2%	25.8%
5	29.0%	8.5%	12.5%	50.1%
6	11.5%	29.3%	8.4%	43.5%
7	11.9%	37.4%	38.0%	12.7%
8	11.9%	15.3%	63.7%	9.1%
9	8.0%	32.2%	25.5%	34.2%
10	42.5%	5.9%	32.8%	18.8%
11	16.6%	11.4%	42.1%	29.9%
12	27.0%	41.9%	20.9%	10.2%
13	54.0%	4.1%	10.5%	31.4%
14	7.4%	7.7%	12.5%	72.3%

Figure 6:

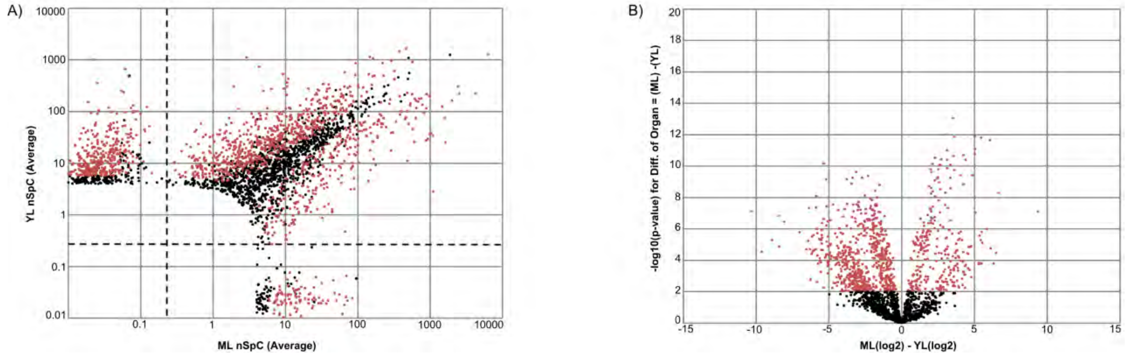
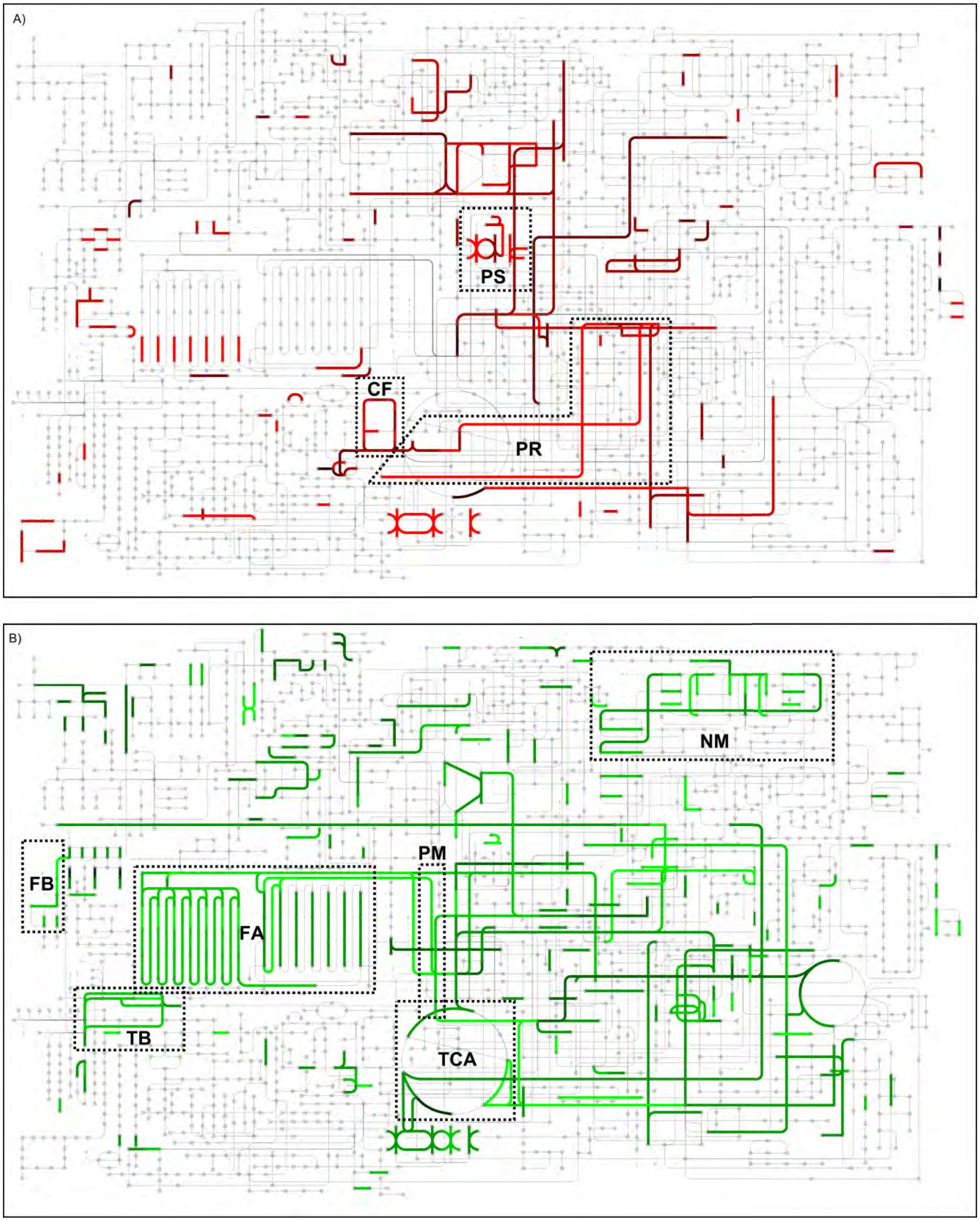


Figure 7:


A)
PS
CF
PR
B)
NM
FB
PM
FA
TB
TCA