Journal of proteome research

ARTICLE

pubs.acs.org/jpr

Defining the Boundaries and Characterizing the Landscape of Functional Genome Expression in Vascular Tissues of *Populus* using Shotgun Proteomics

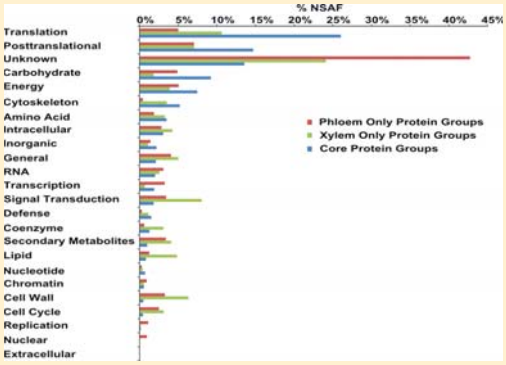
Paul Abraham,[†,‡,§] Rachel Adams,[†,‡,§] Richard J. Giannone,[‡] Udaya Kalluri,[∥] Priya Ranjan,[∥] Brian Erickson,[‡] Manesh Shah,[‡] Gerald A. Tuskan,[∥] and Robert L. Hettich[*,‡]

[∥]Biosciences Division and [‡]Chemical Sciences Division at Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States

[§]Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, Tennessee 37830, United States

Supporting Information

ABSTRACT: Current state-of-the-art experimental and computational proteomic approaches were integrated to obtain a comprehensive protein profile of *Populus* vascular tissue. This featured: (1) a large sample set consisting of two genotypes grown under normal and tension stress conditions, (2) bioinformatics clustering to effectively handle gene duplication, and (3) an informatics approach to track and identify single amino acid polymorphisms (SAAPs). By applying a clustering algorithm to the *Populus* database, the number of protein entries decreased from 64689 *proteins* to a total of 43069 *protein groups*, thereby reducing 7505 identified proteins to a total of 4226 protein groups, in which 2016 were singletons. This reduction implies that ∼50% of the measured proteins shared extensive sequence homology. Using conservative search criteria, we were able to identify 1354 peptides containing a SAAP and 201 peptides that become tryptic due to a K or R substitution. These newly identified peptides correspond to 502 proteins, including 97 previously unidentified proteins. In total, the integration of deep proteome measurements on an extensive sample set with protein clustering and peptide sequence variants provided an exceptional level of proteome characterization for *Populus*, allowing us to spatially resolve the vascular tissue proteome.



KEYWORDS: plant proteomics, single amino acid polymorphisms, populus, mass spectrometry, protein inference, shotgun proteomics, vascular tissue, xylem, phloem

INTRODUCTION

The advent of high-throughput DNA sequencing has revolutionized the assembly of high-quality genomes for prokaryotes and eukaryotes such as plants and humans.[1] The release of reference genomes (http://www.phytozome.net) has paved the way for "omics"-based research which has focused on the identities and functions of the suites of genes and proteins that are important for plant growth and development.[2] In particular, the rapidly developing field of proteomics is already providing remarkable insight into cellular activities at the protein level that complement genomic and transcriptomic investigations.[3−5] That is, obtaining deep protein-level measurements for the identification, quantification, post-translational modification, and localization of proteins has facilitated a more comprehensive understanding of molecular functionality. While there are a variety of proteomic techniques available to measure protein abundance, they differ greatly in their analytical merits of sensitivity, depth of measurement, resolution, and throughput.

The most commonly used platform for plant proteomics has been two-dimensional gel electrophoresis (2-D PAGE) followed by protein sequencing via mass spectrometry.[6] A number of plant proteomic studies published to date have used this platform to map proteomes of various cells, tissues, and organs.[7−9] More recently, online chromatographic mass spectrometry-based proteomics has dramatically extended the throughput and depth of protein identification in complex mixtures by interfacing multidimensional liquid chromatography with nanoelectrospray tandem mass spectrometry (2D-LC−MS/MS).[10−12] Using this gel-free approach, shotgun proteomics (analysis of proteolytic peptide mixtures) has provided detailed qualitative and quantitative observations of cellular metabolic activity for *Oryza sativa* (rice), *Arabidopsis*, and *Populus*.[13−15]

Following the release of the *Populus trichocarpa* genome in 2006, *Populus* emerged as a model system for the study of woody perennial plant biology.[16] The availability of a sequenced genome has prompted vigorous proteomic investigations aimed at elucidating developmental phenomena pertinent to *Populus*.[17−20] Here, we investigate the growth and development of the tree vascular network which involves a complex system that integrates both molecular signaling components and regulation of protein expression. In higher plants, this elaborate network exists in two vascular

Special Issue: Microbial and Plant Proteomics

Received: August 31, 2011

tissues, phloem and xylem. Spanning the entire length of plants, these extensive vascular networks are responsible for the distribution of water and essential nutrients across long distances to vital locations. Insights derived from the detailed identification of proteins and their abundances within *Populus* vascular tissues will undoubtedly yield an improved understanding of the growth and development processes, such as wood biogenesis and drought response.

The full potential of shotgun proteomics in plants is limited in part by the complexities of the proteomic reference database. Most plant genomes contain functional gene redundancies, segmental duplications, single nucleotide polymorphisms (SNPs), and whole-genome reorganizations that have led to gene duplications and adaptive specialization of pre-existing genes (i.e., gene models, protein families and gene duplications that share >90% sequence identity). This inherent redundancy within all plant proteomes confounds the accuracy of the proteome characterization, inflating the total number of proteins identified and/or leading to incorrect biological interpretations. A sophisticated bioinformatics workflow for assigning peptides to proteins and for interpreting resulting protein identifications has to be employed to deal with gene duplications and extended gene families in *Populus*.

Database searching algorithms, such as SEQUEST[21] and MASCOT,[22] which are commonly used to match experimental tandem mass spectra to theoretical fragmentation spectra generated from a predefined proteomic sequence database, cannot resolve peptide spectral matching for any peptide variation unaccounted for in the database. Therefore, when dealing with higher eukaryotes such as humans[23] and plants, a major issue for tandem MS and peptide identification algorithms is the high level of sequence variation, including naturally occurring post-translational modifications (PTMs) and SNP-based single amino acid polymorphisms (SAAPs). In many proteomic measurements, such as those for microbial species, modifications and peptide isoforms do not dramatically affect proteome identification and thus are ignored. In contrast, the complexities of plant proteomics demand attention to these protein alterations, as they have a significant impact on the quality of the proteome characterization.[24] Thus, the degree of sequence variation in *Populus* was explored to identify a number of unassigned quality spectra that result from these common peptide modifications.

In this study, current state-of-the-art experimental and computational approaches were employed to obtain a broad proteome profile of *Populus* vascular tissue. The experimental context includes (1) a large *Populus* sample set consisting of two genotypes grown under normal and tension stress conditions,[25] (2) bioinformatics clustering to effectively handle gene duplication, and (3) an informatics approach to track and identify single amino acid polymorphisms. Together, the integration of deep proteome measurement on an extensive sample set with protein clustering and characterization of peptide sequence variants has provided a level of proteome characterization for *Populus* that has not yet been observed.

2. MATERIALS AND METHODS

2.1. Plant Material

Clonally propagated stem cuttings of two *Populus* clones were established under standard cultural greenhouse conditions following procedures outlined by Kalluri et al. (2009).[15] Two clones were sampled "WV94", a pure *Populus deltoides* clone, and "717", a *P. tremula* x *alba* clone. Cuttings were allowed to grow under normal conditions for six months and then half of the trees in each clone were subjected to tension stress by bending the stem from the apical meristem to the mid stem.[25] After two weeks, xylem and phloem tissue samples were collected from the upper (tension) and lower (opposite) sides of bent stems as well as erect control (normal) stems as described in Kalluri et al. 2009.[15] Six ramets per tissue type per genotype were pooled together for proteomic measurements.

2.2. Protein Extraction and Quantification

Xylem and phloem tissue were ground under liquid nitrogen using a mortar and pestle. For each growth condition, a 3 g sample of 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.[26] 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. Large debris was removed from the sample by centrifugation for 5 min at 1200× *g*. The supernatant was again centrifuged for 10 min at 12000× *g*, and the pellet discarded. A final centrifugation step at 100000× *g* for 1 h yielded a crude soluble protein fraction (cytosolic fraction) and a pellet (pellet fraction). Protein concentration was determined using Lowry's method.[27]

2.3. Protein Digestion

The digestion protocol was modified from methods applied in proteomic analysis of xylem tissue.[15] Prior to MS analysis, samples were denatured and reduced with 6 M guanidine/10 mM dithiothreitol (DTT) for 1 h at 60 °C. These denatured and reduced samples were diluted with 50 mM Tris-HCL/10 mM $CaCl_2$ (pH 7.6) to reduce the guanidine concentration to 1 M. Proteins were digested into peptides with 1:100 (w/w) sequencing-grade trypsin (Promega, Madison, WI) 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. Centrifugation (3000× *g* for 10 min) was performed to remove cellular debris from solution. Digested peptides were desalted off-line using C18 solid phase extraction via SepPak Plus C18 cartridges (Waters), eluting peptides using 100% acetonitrile (ACN). Samples were concentrated using vacuum centrifugation (SpeedVac, Savant Instruments, Holbrook, NY), bringing the final volume to ∼500 μL.

2.4. LC−MS/MS

Peptide analysis was performed using two-dimensional liquid chromatography (strong cation-reverse phase) interfaced with a linear ion trap mass spectrometer (LTQ Thermo Fisher, San Jose, CA) as previously described.[28] In total, 24 different samples were analyzed with 2−3 technical replicates for each sample. For each sample, 100 μg of peptides were bomb-loaded onto a 150-μm inner-diameter back column packed with 4 cm of strong cation exchange column (SCX; Luna, 5 μM particle, 100 Å pore size [Phenomenex]). Prior to MS analysis, each column was washed off-line for 30 min with a gradient from aqueous solvent (95% H_2O/5% ACN/0.1% formic acid) to 50% organic solvent (30% H_2O/70% ACN/0.1% formic acid). Following the SCX wash, the back column was then attached to a 100-μm inner-diameter front column packed with C18 reverse phase (Aqua, 300 Å pore size [Phenomenex]) integrated with a nanospray tip. The column system was positioned on a nanospray source (Proxeon, Denmark) that was aligned in front of the LTQ mass spectrometer. Liquid chromatography was performed by an Ultimate 3000 HPLC pump (LC Packings; a division of Dionex,

dx.doi.org/10.1021/pr200851y |J. Proteome Res. XXXX, XXX, 000–000

San Francisco, CA) at a flow rate of ~300 nL/min at the nanospray tip. All samples were analyzed via 2D-LC over 24 h by 11 consecutive pulses at increasing ammonium acetate salt concentration (0−500 mM), with each salt pulse followed by a 2-h reverse phase gradient elution. During the chromatographic separation, the LTQ was operated in a data-dependent mode and under the direct control of the Xcalibur software (Thermo Fisher Scientific). Tandem mass spectra were acquired in a data-dependent mode provided by Xcalibur software. The data-dependent acquisition used the following parameters: collision-activated dissociation of 5 parent ions were performed following every full scan; 2 microscans were averaged for every full MS and MS/MS spectrum; a 3 m/z isolation width was permitted; 35% collision energy was used for fragmentation; and a dynamic exclusion repeat of 1 with duration of 3 min.

2.5. MS Bioinformatic Analysis

2.5.1. MS/MS Protein Identification. For peptide identification, experimental MS/MS spectra were compared to theoretical tryptic peptide sequences generated from a database containing (1) the protein database of *P. trichocarpa* (v2.0, available at http://www.phytozome.net/cgi-bin/gbrowse/poplar/, containing 45778 proteins), (2) predicted small proteins (20565; 10−200 amino acids in length),[29] 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 determine the false-discovery rate (FDR) for protein identifications.[30] All MS/MS spectra were searched using the SEQUEST algorithm (tryptic cleavages, ≤ 4 missed cleavages allowed) with a parent ion mass tolerance of 3.0 m/z units and a fragment mass tolerance of 0.5 m/z units. Resulting peptide identifications from SEQUEST were filtered and organized into protein identifications using DTASelect version 1.9.[31] Peptide identifications required XCorr values of at least 1.8 (+1), 2.5 (+2), or 3.5 (+3) and a DeltaCN ≥ 0.08. Unless otherwise stated, only proteins identified with two fully tryptic peptides were considered for biological interpretation. These filtering criteria yielded peptide FDRs less than 1%. Estimates of relative protein abundances were based on normalized spectral abundance factors (NSAFs), a semiquantitative label-free approach.[32] NSAF values for each protein were calculated and a value of 0.01 was added to each value to compensate for null values.[33,34] Each NSAF value was then multiplied by a factor of 10000 to convert the NSAF decimal value to a value much easier to visualize.[35] All adjusted NSAF values observed by LC−MS analysis from each growth condition were extracted into a single worksheet.

2.5.2. Creation of Protein Groups. *P. trichocarpa* database proteins sharing extensive sequence homology were assigned to protein groups using a freely available software package, USEARCH v4.0.[36] Proteins that shared more than 90% of their sequence with another protein in the database were clustered by pairwise sequence comparisons using the UCLUST program (similarity threshold of 0.9). A similarity threshold of 0.9 was chosen to reflect the level of intraproteomic similarity in the *Populus* proteome: two genome-wide duplication events have increased the level of redundancy, in which nearly two-thirds of the protein-coding genes share sequence similarity (>90%).[37] Moreover, plotting similarity thresholds ranging from 0.5 to 1.0 against the percent proteome reduction via clustering provided further support in choosing 90% as a cutoff (data not shown). Protein groups are defined by the longest protein sequence, the seed, which shares ≥90% sequence identity to each protein in that cluster. All groups were manually verified to ensure that obvious redundancies, such as alternatively spliced variants, remained together. In contrast with the database searching parameters detailed in section 2.5.1, peptides belonging to proteins that passed DTASelect's one peptide filter were reorganized into protein groups. Those identified peptides that were unique to a particular protein group were marked as *protein-group unique*, meaning these peptides did not belong to another group of proteins. All peptides that were originally database-unique were necessarily protein group-unique, but grouping the peptides of homologous proteins allowed nondatabase-unique peptides to be considered unique if they belonged to only one protein group identification, that is, a protein-group-unique peptide. NSAF values were calculated for each protein group by normalizing the sum of the total spectral counts for each peptide belonging to a protein group. For peptides belonging to multiple protein groups, the spectral counts were recalculated based on the proportion of uniquely identified peptides between the protein groups sharing the peptide in question.[35,38] While NSAF values typically account for biases resulting from protein length, here NSAF values were calculated for each protein group by using the length of the seed sequence. Adjusted NSAF values were calculated for each protein group.

These protein-group unique peptides were instrumental in both parts of a two-tiered filtering process performed on the identified protein groups. The first filter required evidence of at least one protein-group unique peptide and two total peptides to support the unambiguous identification of a protein group. The second filter removed low abundance protein groups with adjusted NSAF values less than 5.

2.5.3. Single Nucleotide Polymorphism Analysis. High-throughput SNP discovery through deep (approximately 30× depth per genotype) resequencing of 19 trees yielded 16 million SNPs in the *Populus* genome (485 Mb) (unpublished results). For this analysis, a subset of these SNPs present in 2 *P. trichocarpa* and 2 *P. deltoides* genotypes were considered. Of the 17 million amino acid positions found in *P. trichocarpa's* 45778 protein-coding gene models, ~400000 amino acid positions due to nonsynonymous SNPs (SAAP) were investigated. All possible combinations of SNP-influenced peptides (SAAP peptides) were predicted and subjected to *in silico* tryptic cleavage using PeptideSieve software[39] with the following parameters: maximum mass criterion of 5000, minimum sequence length of 6, maximum sequence length of 50 and allowing for 4 missed cleavages. Some of the nonsynonymous amino acid changes resulted in new tryptic cleavage sites or resulted in disappearance of these sites. These were taken into consideration while predicting the peptides. To detect the expression of a SAAP peptide, experimental MS/MS spectra from one MS run were compared to theoretical tryptic peptide sequences generated from a target database consisting of the protein database of *P. trichocarpa* (v2.0) and all predicted SAAP peptides. Each SAAP peptide was concatenated to the target database as a new protein entry, in which ten tryptophan residues flanked both sides of the peptide sequences. For SAAP peptides that originated from the N-terminus of a protein, the tryptophan residues were excluded from the beginning of the SAAP peptide. Similarly, for each SAAP peptide that originated from the C-terminus of a protein, the tryptophan residues were excluded at the end of the SAAP peptide. In total, 7775313 additional entries from SAAP peptides were included. All MS/MS were searched with SEQUEST and filtered by DTASelect as described above. Similar to section

2.5.1, filtering criteria were controlled to yield peptide FDRs less than 1%.

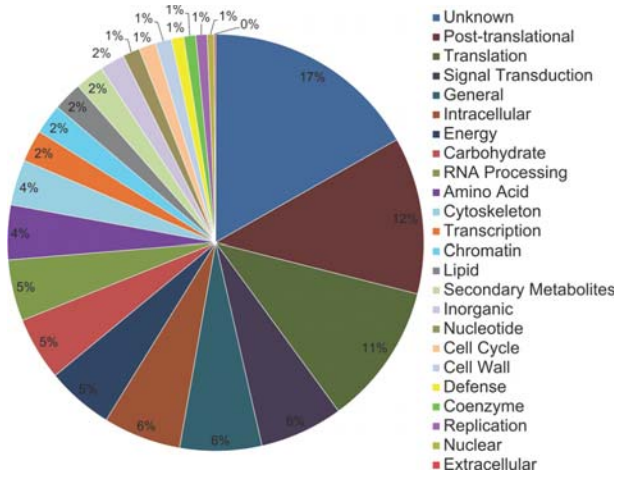


Figure 1. Distribution of detected proteins by their functional classification. The data indicates the most abundant functional categories for the combined xylem and phloem vascular tissue proteomes.

3. RESULTS AND DISCUSSION

3.1. Characterizing the Landscape: Global Survey of the *Populus* Proteome

3.1.1. Mapping Deep Measurements to the *Populus* Proteome. To generate a high-coverage proteome profile, we performed shotgun proteomics on a large sample set consisting of subcellular fractions (soluble, pellet) of two tissue types (xylem, phloem) from two *Populus* species: *P. deltoides* and *P. tremula* × *alba*. Using the most recent *Populus* genome draft (v2.0, http://www.phytozome.net/cgi-bin/gbrowse/poplar/), tandem mass spectra from 60 *Populus* proteome measurements collectively identified 7,505 total proteins and 33233 tryptic peptide sequences with an overall false discovery rate of <1% at the protein level. Combining the proteome measurements together provided a global view of protein expression involved in vascular tissue development, resulting in protein assignments for ∼17% of the predicted *Populus* proteome. Approximately 40% of all detected proteins belonged to three specific functional categories based on 24 EuKaryotic Orthologous Groups (KOGs): (1) unknown function, (2) post-translational modification and turnover, and (3) signal transduction (Figure 1). The remaining identified proteins are scattered across the other

Table 1. Total Number of Proteins and Peptides Observed for Each of the 12 Conditions in Two Genotypes

	P. deltoides		*P. tremula X alba*	
sample type	proteins	peptides	proteins	peptides
Xylem Stress (Normal) Soluble Replicate 1	3690	18715	2980	14694
Xylem Stress (Normal) Soluble Replicate 2	3623	18371	2795	12079
Xylem Stress (Tension) Soluble Replicate 1	3088	17278	2889	13497
Xylem Stress (Tension) Soluble Replicate 2	3200	17730	3048	17223
Xylem Stress (Opposite) Soluble Replicate 1	3846	20579	3032	16929
Xylem Stress (Opposite) Soluble Replicate 2	3608	19541	2106	9683
Xylem Stress (Normal) Pellet Replicate 1	2145	8353	1631	6584
Xylem Stress (Normal) Pellet Replicate 2	2304	9491	1467	5549
Xylem Stress (Normal) Pellet Replicate 3	2325	9418	894	2654
Xylem Stress (Tension) Pellet Replicate 1	2267	10923	1714	5491
Xylem Stress (Tension) Pellet Replicate 2	2092	9618	1949	7843
Xylem Stress (Tension) Pellet Replicate 3	2130	9735	1477	4724
Xylem Stress (Opposite) Pellet Replicate 1	2340	9902	1380	4250
Xylem Stress (Opposite) Pellet Replicate 2	2276	9582	1680	5641
Xylem Stress (Opposite) Pellet Replicate 3	2203	9627	1835	6680
Phloem Stress (Normal) Soluble Replicate 1	2322	8401	1676	6198
Phloem Stress (Normal) Soluble Replicate 2	2237	8314	1243	4021
Phloem Stress (Tension) Soluble Replicate 1	2798	10995	1585	5538
Phloem Stress (Tension) Soluble Replicate 2	2840	11519	1412	4557
Phloem Stress (Opposite) Soluble Replicate 1	2396	8875	1328	4270
Phloem Stress (Opposite) Soluble Replicate 2	2432	9094	1657	6128
Phloem Stress (Normal) Pellet Replicate 1	2038	5706	1895	5808
Phloem Stress (Normal) Pellet Replicate 2	2091	5964	788	1591
Phloem Stress (Normal) Pellet Replicate 3	1944	5699	1335	3327
Phloem Stress (Tension) Pellet Replicate 1	2314	7176	529	903
Phloem Stress (Tension) Pellet Replicate 2	2158	6500	347	503
Phloem Stress (Tension) Pellet Replicate 3	2164	5636	500	823
Phloem Stress (Opposite) Pellet Replicate 1	2124	6860	1824	5450
Phloem Stress (Opposite) Pellet Replicate 2	2054	6651	1864	5123
Phloem Stress (Opposite) Pellet Replicate 3	1563	4149	2027	5898

dx.doi.org/10.1021/pr200851y |*J. Proteome Res.* XXXX, XXX, 000–000

functional categories. The number of redundant proteins and peptides identified for each sample type and technical replicate are shown in Table 1.

3.1.2. Genetic Redundancy and Protein Classification. Shotgun proteomics employs a peptide-centric approach that relies on the ability to accurately assemble and assign thousands of measured peptides to reference proteins in biological samples. Although this is the conventional method for identifying proteins in large-scale studies, this approach presents several challenges when assigning peptides to proteins in higher eukaryotes. The most common issue deals with inferring a protein's existence through the identification of peptides that constitute its primary structure. "Protein inference" becomes problematic when two or more proteins share peptides.[40−42] Shared or *degenerate* peptides are natural occurrences that originate from protein homology, conserved protein domains among various proteins, splice variants, and redundant entries due to gene duplication events, all of which are common in plants.[43,44] In fact, the *Populus* genome is highly genetically redundant, such that two-thirds of protein-coding genes share sequence similarity greater than 90%.[37] Therefore, within the large data sets, emphasis must be placed on accurate identification and validation of proteins, accounting for highly conserved, shared peptides.

In previous studies, the categorical nomenclature of Yang et al. (2004)[45] has been adapted to rationally organize the peptide data from each LC−MS/MS experiment. Several research groups have shown that this nomenclature can be coupled with Occam's razor constraints to provide a minimal list of proteins to explain all observed peptides.[42] Using this classification method, we consolidated protein assignments by their level of uniqueness. Proteins that consist of only uniquely identified peptides were classified as *distinct* proteins. Proteins were classified as *differentiable* when they contain at least one peptide that is unique to that locus, as well as one or more peptides that map elsewhere in the proteome. The *indistinguishable* proteins consisted of measured nonunique peptides that map elsewhere in the data set. Within our entire data set, only 50% of the tryptic peptides identified were classified as unique to the database. Therefore, out of the 7505 total protein identifications in the present study, 3510 proteins were uniquely identified (classified as distinct or differentiable) and 3995 proteins were categorized as nonunique or indistinguishable (Supplemental Table 1, Supporting Information).

Using the nomenclature above, we generated a minimal list of proteins that were conclusively determined to be present within the data set. However, due to the inherent ambiguity of the *Populus* proteome, less than 50% of the proteins categorized by the above-mentioned criteria could be used for biological interpretation. In addition, due to the extensive homology within the database, a vast majority of the proteins were classified as indistinguishable. As most of the proteins in this category contain no unique peptides, it was difficult to determine which specific proteins were present in the sample using an MS-based approach. As shown in other studies, one approach for proteins that cannot be distinguished on the basis of identified peptides is to collapse these into protein groups to provide a more accurate and informative data set.[46,47] In an attempt to reconcile this problem, a bioinformatics workflow was incorporated to better handle proteins sharing high sequence homology (≥90%) to increase qualitative accuracy by avoiding the over- and under-identification of homologous proteins.

An illustration of the informatics workflow can be seen in Figure 2A. Briefly, proteins sharing 90% or more sequence identity were clustered into groups by UCLUST, a clustering algorithm

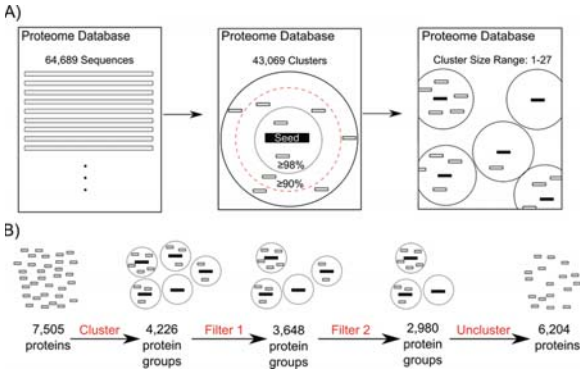


Figure 2. Illustration of bioinformatic workflow. (A) All proteins in the proteomic database were clustered by UCLUST to deal with gene duplications and extended gene families. (B) After the proteins were clustered into protein groups, a conservative two-tiered filtering approach was used to eliminate (1) ambiguous identifications and (2) those at the lower detection limits.

functionally equivalent to BLASTP.[36] Each protein group was defined by a representative protein sequence called a seed, where each seed shares ≥90% sequence identity to each protein in that cluster. By applying the clustering algorithm to the *Populus* database, the number of protein entries decreased from 64689 *proteins* to a total of 43069 *protein groups*. Implementation of clustering to the data set reduced the 7505 observed proteins to a total of 4226 protein groups (see Methods), in which 2016 were singletons (i.e., a one-member group). This reduction implies that ∼50% of the observed proteins were clustered into groups that shared extensive sequence homology. Therefore, this approach effectively consolidates indistinguishable proteins into a meaningful report. Although grouping proteins by high sequence similarity undoubtedly sacrifices some level of protein resolution, it is reasonable to assume that proteins with this level of sequence homology share similar biological functions. Furthermore, integrating the clustering approach with the initial SEQUEST analysis provided a means to categorize which members of a protein group were unique.

Due to the peptide-centric nature of shotgun proteomics, it was imperative to report peptides in the context of protein groups. As expected, clustering proteins into groups alleviated some of the ambiguity associated with shared peptides. Similar to a peptide being unique to a protein within the database, we found many peptides were unique to a particular *protein group* within the clustered database. In fact, 68% of previously shared peptides that were classified as nonunique to the *Populus* database were reclassified as unique to the clustered database. Moreover, the bioinformatics workflow generated a data set where 84% of the detected peptides were classified as unique. Therefore, rather than disregarding these peptides from the analysis, they were rescued and used for biological insight (Figure 2B). While it may not be clear as to which member of a protein group is actually present in a given sample, the identification of peptides belonging to a particular protein group likely indicates the presence of a shared functional process, especially considering the relatively stringent similarity cutoff (90%) applied to the protein database.[48]

3.1.3. Characterization of the *Populus* Vascular Tissue Proteome. Xylem and phloem tissues are responsible for long-distance transportation and storage of essential minerals and nutrients in plants. A recent study used shotgun proteomics to

dx.doi.org/10.1021/pr200851y |J. Proteome Res. XXXX, XXX, 000–000

examine proteins expressed during xylem development.[15] This approach demonstrated an ability to robustly characterize xylem tissue in *Populus* by vastly increasing the number of proteins identified and characterized relative to previous *Populus* proteome studies.[17] In the current study, a similar experimental approach was applied to identify and contrast the relationship and dissimilarities between the xylem and phloem proteomes. A "core" proteome was extracted from the entire data set, consisting of 2627 protein groups that were confidently identified in both xylem and phloem. The core proteome, encompassing 59% of the total proteins identified in the *Populus* data set, includes proteins representing each KOG category (Figure 3). The core metabolic signature is consistent with other studies that show an overrepresentation of proteins that are involved in energy production and translation.[14] Moreover, a similar quantitative distribution profile was also observed during xylem development.[15,29] In addition, these functionally and spatially separate vascular networks contain tissue-specific proteins: 606 unique xylem proteins-groups and 461 unique phloem protein groups, each having a distinct metabolic profile as shown in Figure 3.

3.1.4. Regulatory Proteins Involved in Vascular Tissue Development. Among the proteins identified in the *Populus* data set there were proteins that have been shown to control the patterning and differentiation of vascular tissues. Interestingly, the receptor protein kinase CLAVATA1 precursor, a part of the CLV3/CLV1 system, was exclusively identified in phloem tissue. The developmental process of the plant vascular network is a complex system that integrates both molecular signaling components and regulation of protein expression. Stem cells in the shoot apical meristem regulate the continuous formation of the different tissues during vascular formation. It has been shown that the receptor protein kinase CLAVATA1 governs stem cell fates in the shoot apical meristem. Along the boundaries of the procambium/cambium space of postembryonic tissue, this process occurs when CLAVATA1 binds to the protein ligand CLE41, which is secreted from the phloem.[49,50]

We also identified *bri1* suppressor 1 (BSU1) protein only in xylem tissue. BSU1 is a ser/thr-protein phosphatase that has been shown to be a positive regulator of brassinolide signaling, thereby playing an important role in the regulation brassinosteroids.[51] It has been shown that brassinosteroids regulate xylem differentiation and vascular patterning from cambium cells.[52] Furthermore, brassinosteroid lack-of-function mutants in *Arabidopsis*[53] and rice[54] disrupt vascular development. It is also known that the plant hormone auxin plays a critical role in the cell-to-cell communication in vascular differentiation.[55] We detected evidence for peptidyl-prolyl cis−trans isomerase protein (PIN1) expression in both xylem and phloem tissue. Currently, many studies suggest that the formation of plant vascular networks is an auxin-transport-based mechanism and the driving force behind this mechanism is the accumulation and polarization of PIN1, an auxin efflux carrier.[56,57] On the basis of our measurements, PIN1 expression in phloem may provide a bidirectional pathway for long-distance transportation, while expression in xylem leads to vascular development and xylem differentiation.

3.1.5. Biosynthesis and Development of Wood Cell Walls. We identified several of the cell wall-related carbohydrate active enzymes within our data set, including cellulose synthases, pectin methyl esterases, and xylogulcan endotransglucosylases and hydrolases. Wood, or secondary xylem, is a water conduit formed from the vascular cambium that provides mechanical support for plants and is the primary source of chemical feedstock

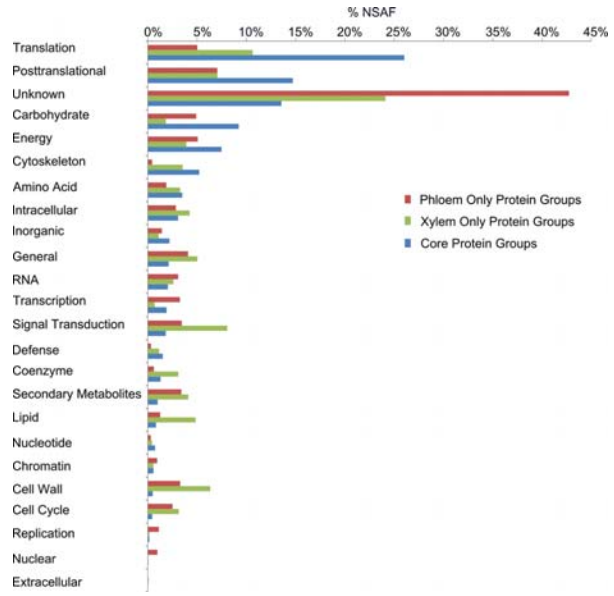


Figure 3. Quantitative distribution of detected proteins by their functional classification. The relative abundance of each functional category was calculated as a percent of the summed protein group abundance within each classification: protein groups found in both xylem and phloem (the core proteome), protein groups found only in phloem, and protein groups found only in xylem.

for the emerging biofuels industry.[58,59] The cell wall is composed of a carbohydrate matrix consisting of cellulose microfibrils that are embedded within a mixture of hemicellulose and lignin, a polymer with subunits of phenylpropanoid.[60,61] Carbohydrate active enzymes (CAZymes) are known components of the construction and remodeling of the carbohydrate matrix.[62] Our proteomics profile identified several genes encoding CAZymes, concurrent with results from EST and microarray analysis.[63−65]

Lignin, the other main constituent of the wood cell wall, is a complex phenolic polymer that provides a physical barrier that protects plants from microbial and physical attack and provides mechanical support. Lignin is polymerized from three primary monomers: *p*-coumaryl alcohol (H), coniferyl alcohol (G) and sinapyl alcohol (S). The monolignols are synthesized from phenylalanine through the phenylpropanoid pathway and, within the *Populus* genome, 95 gene models have been identified as putative phenylproponoid biosynthesis genes.[66] The genetic and biochemical role of most of the 95 gene models remains undefined. Our study identified proteins associated with the monolignol biosynthesis pathway, identifying members for each enzyme family (Table 2).

3.2. Defining the Boundaries: Interrogation of Unassigned MS/MS Spectra

3.2.1. Spectral Quality Assessment. Although remarkable depth of coverage of the *Populus* proteome has been achieved, one of the greatest heuristics that contributes to the success of database-searching approaches also has a complementary limitation: regardless of the quality of peptide-derived spectra, algorithms will only match spectra to peptides that exist within user-defined sequence variations. Peptide sequencing by mass spectrometry is most commonly performed via collisional-induced dissociation (CID), in which peptide ions fragment in a predictable manner to produce dissociation products that yield sequence

Table 2. Protein and Peptide Classification for the Monolignol Biosynthesis Pathway[a]

protein family	protein	protein parsimony	seed	DU peptides	PGU peptides	NU peptides	total peptides	
PAL	pt017188m	*Differentiable*	pt002727m	15	12	14	41	
	pt002727m	*Differentiable*	pt002727m	19	11	12	42	
	pt026599m	*Differentiable*	pt026599m	16	10	12	38	
	pt011283m	*Differentiable*	pt026599m	7	36	11	54	
	pt011320m	*Differentiable*	pt026599m	6	33	9	48	
C4H	pt030573m	*Indistinguishable*	pt009878m	0	31	0	31	
	pt009878m	*Differentiable*	pt009878m	10	16	0	26	
	pt030574m	*Indistinguishable*	pt009878m	0	31	0	31	
4CL	pt017853m	*Distinct*	pt017853m	2	0	0	2	
	pt038100m	*Distinct*	pt038100m	27	0	9	36	
	pt023497m	*Distinct*	pt023497m	7	0	0	7	
	pt024040m	*Differentiable*	pt024040m	4	0	4	8	
HCT	pt023671m	*Distinct*	pt023671m	7	0	3	10	
	pt038643m	*Differentiable*	pt038643m	13	0	3	16	
C3H	pt002886m	*Indistinguishable*	pt002886m	0	2	3	5	
	pt002890m	*Indistinguishable*	pt002886m	0	2	3	5	
	pt017558m	*Distinct*	pt017558m	13	0	4	17	
CCoAOMT	pt005042m	*Differentiable*	pt005042m	8	17	1	26	
	pt039874m	*Differentiable*	pt005042m	4	17	1	22	
	pt027738m	*Distinct*	pt027738m	7	0	4	11	
CCR	pt005074m	*Indistinguishable*	pt004827m	0	2	0	2	
	pt004953m	*Indistinguishable*	pt004827m	0	4	0	4	
	pt004827m	*Indistinguishable*	pt004827m	0	6	0	6	
	pt005089m	*Indistinguishable*	pt004827m	0	2	0	2	
	pt005064m	*Indistinguishable*	pt004827m	0	2	0	2	
	pt004830m	*Distinct*	pt004830m	12	1	1	14	
	pt039322m	*Differentiable*	pt004830m	1	1	0	2	
	pt004839m	*Distinct*	pt004839m	11	0	1	12	
	pt012284m	*Distinct*	pt012284m	2	0	1	3	
	pt020991m	*Differentiable*	pt020991m	2	0	3	5	
	pt030211m	*Indistinguishable*	pt021000m	0	5	0	5	
	pt021000m	*Indistinguishable*	pt021000m	0	5	0	5	
	pt021032m	*Distinct*	pt021032m	1	0	0	1	
	pt023595m	*Distinct*	pt023595m	6	0	3	9	
	pt023595m	*Distinct*	pt023595m	6	0	3	9	
	pt033373m	*Differentiable*	pt033373m	2	1	2	5	
	pt033727m	*Indistinguishable*	pt033373m	0	1	2	3	
CAld5H	pt032677m	*Differentiable*	pt032677m	3	0	6	9	
	pt025189m	*Differentiable*	pt025189m	7	0	5	12	
COMT	pt000701m	*Indistinguishable*	pt000702m	0	4	15	19	
	pt000702m	*Indistinguishable*	pt000702m	0	4	15	19	
	pt010103m	*Distinct*	pt010103m	3	0	0	3	
	pt015982m	*Distinct*	pt015982m	28	0	15	43	
	pt018020m	*Indistinguishable*	pt018020m	0	5	0	5	
	pt018431m	*Indistinguishable*	pt018020m	0	5	0	5	
	pt020855m	*Indistinguishable*	pt020853m	0	3	0	3	
	pt022214m	*Indistinguishable*	pt020853m	0	3	0	3	
	pt020853m	*Indistinguishable*	pt020853m	0	3	0	3	
	pt020964m	*Indistinguishable*	pt020853m	0	2	0	2	
CAD	pt003155m	*Distinct*	pt003155m	1	0	0	1	
	pt003292m	*Distinct*	pt003292m	11	0	0	11	
	pt004002m	*Distinct*	pt004002m	2	0	0	2	
	pt004753m	*Distinct*	pt004753m	31	0	2	33	
		pt018077m	*Indistinguishable*	pt018073m	0	3	0	3
		pt018073m	*Indistinguishable*	pt018073m	0	3	0	3
		pt039056m	*Distinct*	pt039056m	8	0	0	8

[a] A detailed classification of the peptides detected within 10 protein families contributing to lignin biosynthesis. A protein was marked as *distinct*, *differentiable*, or *indistinguishable* according to its number of database-unique (*DU*) peptides detected. After reorganizing proteins into their protein groups, peptide uniqueness was reevaluated for protein-group uniqueness (*PGU*). The number of nonunique (*NU*) peptides was also reported.

information. Though widely used for its simplicity and effectiveness, more than 50% of MS/MS spectra collected in a typical shotgun proteomic experiment do not result in high-confidence peptide identifications when using automated search algorithms such as SEQUEST or MASCOT. Even though these low identification rates can be partially explained by the presence of spectra arising from concurrent fragmentation of multiple precursor ions, incomplete fragmentation of

dx.doi.org/10.1021/pr200851y |J. Proteome Res. XXXX, XXX, 000–000

peptides, and chemical noise, a large fraction of peptide-derived spectra remain unassigned because of the quality and completeness of the proteome database.[42,67]

Neither prokaryotic nor eukaryotic protein databases typically include protein isoforms or alterations/modifications, and furthermore their omission has a more dramatic effect on higher eukaryotes in which sequence variations and unexpected splice variants are more prevalent. Thus, by not anticipating the presence of these peptides, database search algorithms are more likely to interpret fewer peptide-derived MS/MS spectra when analyzing proteomes of higher eukaryotes. Reanalysis of unassigned tandem mass spectra was performed to determine the magnitude of peptide-derived spectra that remained unmatched to a sequence, thereby providing the proportion of "missing" peptide identifications in a run.

To compare the rates of peptide-spectrum matching (PSM) between eukaryotes and prokaryotes, we contrasted MS data from *Populus* with a simpler bacterium, *Escherichia coli*.[68] In both cases, proteolytic peptides were measured on the same instrument using identical methods to minimize experimental biases. The instrumental acquisition and chromatographic distribution of all MS/MS spectra collected were similar for both organisms (Figure 4A). However, the ability to successfully match experimental MS/MS spectra to theoretical database sequences was superior in *E. coli*. A greater percentage (86%) of *Populus* MS/MS spectra remained unassigned, as compared to only 63% of the MS/MS spectra collected for *E. coli*. A closer look at the proportion of unassigned peptide-derived spectra was used to determine if the observed discrepancies in peptide identifications could be attributed to the incompleteness of the reference database. Spectral quality assessment was used to identify the number of unassigned high-quality spectra, that is, a population of spectra that likely represents mutated, modified or novel peptides. A conservative set of criteria, based on previous implementations of spectral analysis,[69,70] was utilized in the assessment of MS/MS spectral quality. A spectrum was considered high quality if the parent charge state was calculated to be greater than +1 and if the spectrum contained three or more peaks within 20% of the base peak intensity with a minimum intensity of 2500 counts. Using this approach, we performed an assessment of MS/MS spectra quality to distinguish high-quality unassigned spectra from low-quality unassigned spectra in the representative MS runs from *Populus* and *E. coli*. Spectra analysis revealed that, of the total MS/MS spectra collected for *Populus* and *E. coli*, the percentage of high-quality MS/MS spectra (45%) within the representative MS run for *Populus* contained almost twice the percentage (24%) in the *E. coli* run (Figure 4B). Nonetheless, the ability to successfully match the high-quality experimental MS/MS spectra to database sequences remained more common in *E. coli*. A greater percentage of *Populus* high-quality MS/MS spectra (77%) remained unassigned, as compared to only 45% of the high-quality MS/MS spectra collected for *E. coli*. Obviously caution must be exercised not to overinterpret these results, as the level of protein modifications might be very different in these two cases. Clearly the use of *de novo* sequencing and reporter fragment ions (such as those from glycopeptides) might provide a more uniform comparison, but the goal of this initial work was a direct head-to-head comparison of spectra quality under identical and fairly standard experimental and computational conditions employed for typical proteome measurements. In total, these results suggest a critical need to evaluate bioinformatic approaches to rescue the lost, high-quality spectra.

3.2.2. SAAP-Resolved *Populus* Proteomics. One source of unassigned high-quality tandem MS spectra may be peptides containing SAAPs. *Populus* has an estimated one SNP per 20 base pairs (unpublished results), while humans have an estimated one SNP per 1.9 kilobase pairs.[71] The biological implication of a SNP depends on its positional location within the genome and gene structure. Within coding regions, a SNP can be either synonymous which does not alter the amino acid or nonsynonymous which results in an amino acid substitution. Detection of SAAPs not only identifies amino acid changes that have physiochemical consequences but also reveals information regarding sequence, and perhaps phenotypic, variability within a proteome. Therefore, a database containing SNP-based SAAPs and other sequence variations could be highly informative.

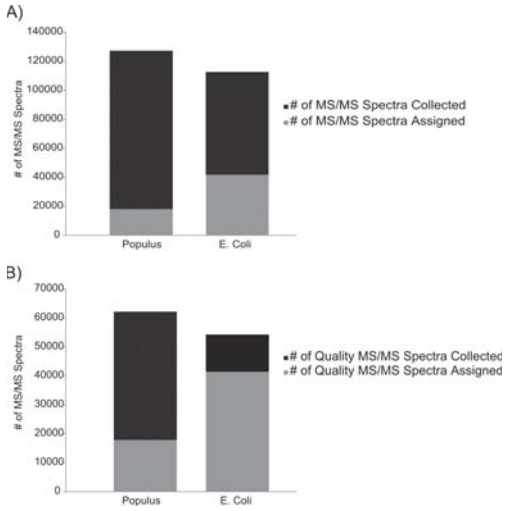


Figure 4. Spectral quality assessment. (A) Comparison of peptide-spectrum matching rates between *E. coli* and *Populus*. (B) Quantitative assessment of the proportion of high-quality MS/MS spectra collected versus those assigned for the representative MS runs from *E. coli* and *Populus*.

To explore the prevalence of SAAPs, a single MS run from within the 60 described above was searched against an expanded *Populus* database that included a list of tryptic peptides generated from predicted SAAP variants in the database. With the high frequency of SAAPs in *Populus*, over 700000 distinct SAAP positions and 7200000 new peptides were included in our database. We found that *Populus* proteins on average contained 17 SAAPs. When identifying SAAPs from MS/MS spectra, it is important to differentiate these from post-translational modifications (PTMs) or peptide modifications generated during sample processing that result in mass shifts which are isobaric to several amino acid substitutions. For example, the covalent addition of a methyl group to a K, R, E, or Q produces a mass shift that is similar to the following amino acid changes: D to E, S to T, V to I/L, and G to A. Therefore, all spectra interpreted as both a PTM and a SAAP should be discarded to lower the identification of false positives. Certainly, utilization of a higher performance mass spectrometer, such as an FTICR-MS or an LTQ-Orbitrap, would provide higher resolution and better mass accuracies to rescue and differentiate many of these ambiguous SAAP/PTM peptides, including sample artifacts such as oxidation and deamidation, but the goal here was to demonstrate that a large eukaryotic genome database containing extensive SNP information could be successfully searched and mined for SAAP information even under conservative, low resolution mass spectrometric conditions. To identify a targeted common set of PTMs (Supplemental Table 2, Supporting Information), MS/MS spectra were analyzed by an

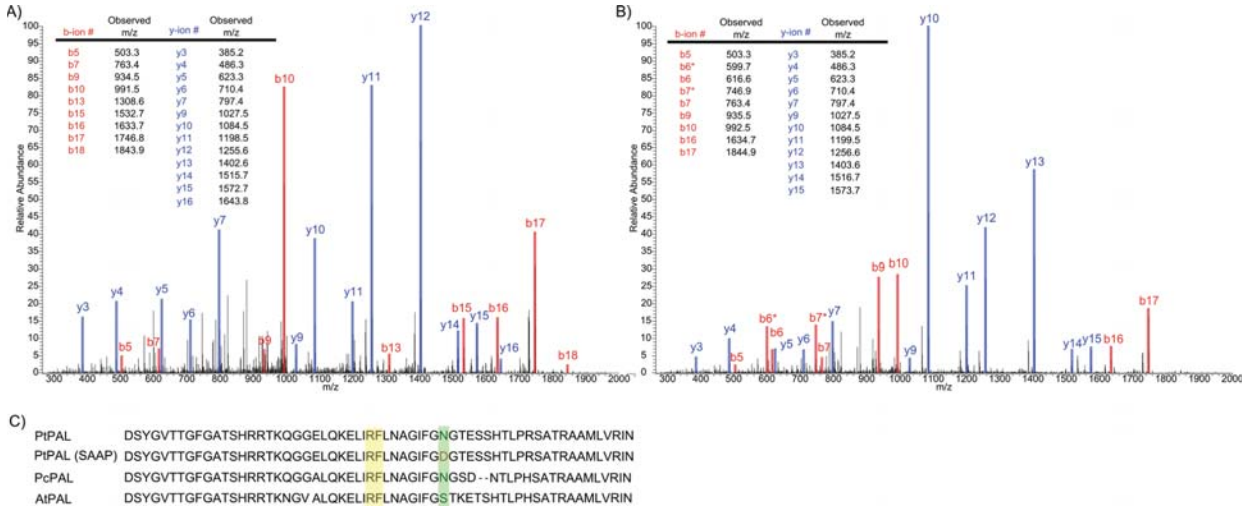


Figure 5. SAAP-resolved peptide identification in PAL. (A) MS/MS spectra of the genomic tryptic peptide (FLNAGIFGNGTESSHTLPR) and the (B) SAAP tryptic peptide (FLNAGIFGDGTESSHTLPR). (C) Partial amino acid (single letter codes) sequence alignment of *P. trichocarpa* (PtPAL) with other members of the phenylalanine ammonia-lyase family (PcPAL, *P. Crisum*, and AtPAL, *A. thaliana*). Only the region near the SAAP-containing peptide is shown. The green box highlights the substrate specificity residues and the yellow box highlights the SAAP position.

automated software tool, InSpecT,[72] at a FDR of 2%. In total, 271 spectra that matched to both a PTM and a SAAP peptide were removed from the analysis. Using conservative search criteria, we were able to identify a total of 1,354 peptides containing a SAAP and 201 peptides that become tryptic due to a K or R substitution (Supplemental Table 3, Supporting Information). Although the new SAAP peptides account for 2% of high-quality unassigned spectra, these newly identified peptides correspond to 502 proteins. Among these, we identified 97 proteins that had not been previously identified. Interestingly, for those proteins containing a SAAP peptide, their overall peptide coverage increased by an average 25%.

Due to the widespread distribution of SAAP peptides in the database, it seems probable that the detected SAAP peptides would map randomly across the proteome. However, our data suggests that the detected population of proteins containing a SAAP peptide map to specific and functionally similar groups. Grouping the SAAP proteins into KOGs, the vast majority of SAAP proteins belonged to the four specific functional categories: unknown function, signal transduction, post-translational modification, and carbohydrate transport and metabolism. Although these functional categories are among the most abundant categories in phloem and xylem, we note that other abundant functional categories, such as general function and translation, do not contain a large number of proteins containing SAAPs. Therefore, it appears that the overrepresentation of nonsynonymous substitutions for the aforementioned functional categories is not a result of their expression levels, but rather that these proteins are under low selective pressure. Although it is unclear how many of these proteins represent evolutionary novelties, future comparative proteomics studies may identify expression patterns that reveal the outcomes of such mutations. In some instances, the location of these mutations could compromise or benefit an enzyme: replacing catalytic, binding, or substrate determining residues with amino acids differing in size, polarity, or hydrophobicity can either disrupt or modulate the activity of an enzyme.

For example, when looking at the monolignol biosynthesis pathway, we identified a SAAP within phenylalanine ammonia lyase (PAL), the entry enzyme into the phenylproponoid pathway. As shown in Figure 5, a mass shift of +1 Da and the experimental b- and y- ion fragmentation pattern coincides with the predicted SAAP substitution of an asparagine (N) with an aspartic acid (D) at position 138. While the effect of the observed polymorphism is unknown, the localization of the substitution within a few amino acids of the substrate-binding site may impact the binding of coumarate to the substrate specificity residues.[73] Because studies have shown that PAL serves as a regulatory control point for the entire pathway,[74] any mutations compromising or altering the activity of the enzyme will, in fact, impact the overall lignin content.

CONCLUSIONS

While it is still unknown what percent of the *Populus* proteome is expressed given a specific time and tissue, combining tandem mass spectra from 60 MS runs yielded a preview of protein expression in xylem and phloem. Perhaps one of the most challenging tasks in proteomic studies of higher eukaryotes is inferring which proteins are present in a particular sample based on the observed peptides. An enhanced bioinformatic workflow alleviated some of the difficulties associated with data interpretation by recasting protein identifications as protein groups, which have a high degree of sequence similarity and therefore most likely share similar biological roles. The resulting data set provided a more accurate and informative perspective that allowed us to characterize the landscape of protein expression in xylem and phloem.

In addition, to fully characterize the boundaries of assignable peptides, we assessed spectral quality and found a large portion of the high-quality spectra remained unassigned. When dealing with higher eukaryotes such as plants, a major issue for tandem MS and peptide identification algorithms is the high level of sequence variation, including naturally occurring PTMs and SNP-based SAAPs. The exact scope and frequency of these detectable protein variants has, to our knowledge, never been reported to date in any plant. By investigating the prevalence of detectable SAAPs, we provide a glimpse of detectable proteins beyond the "basic" proteome (predicted gene products).

dx.doi.org/10.1021/pr200851y |*J. Proteome Res.* XXXX, XXX, 000–000

All together, the integration of deep proteome measurement on an extensive sample set with protein clustering and identification of protein sequence variants pioneered a level of proteome characterization for *Populus* that has not been possible before.

■ ASSOCIATED CONTENT

Supporting Information

Supplementary tables. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6131. Phone: 865-574-4986. Fax: 865-576-8559.

Author Contributions

†These authors contributed equally to this manuscript.

Funding Sources

This manuscript has been authored with funding from 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 nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

■ ACKNOWLEDGMENT

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 in the DOE Office of Science. Oak Ridge National Laboratory is managed by University of Tennessee-Battelle LLC for the Department of Energy.

■ REFERENCES

(1) Deschamps, S.; Campbell, M. A. Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Mol. Breeding* **2010**, *25* (4), 553–570.

(2) Saito, K.; Matsuda, F. Metabolomics for Functional Genomics, Systems Biology, and Biotechnology. *Annu. Rev. Plant Biol.* **2010**, *61*, 463–489.

(3) Gygi, S. P.; Rochon, Y.; Franza, B. R.; Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **1999**, *19* (3), 1720–1730.

(4) Fiehn, O.; Kopka, J.; Dormann, P.; Altmann, T.; Trethewey, R. N.; Willmitzer, L. Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* **2000**, *18* (11), 1157–1161.

(5) de Godoy, L. M. F.; Olsen, J. V.; Cox, J.; Nielsen, M. L.; Hubner, N. C.; Frohlich, F.; Walther, T. C.; Mann, M. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **2008**, *455* (7217), 1251–1260.

(6) Gorg, A.; Weiss, W.; Dunn, M. J. Current two-dimensional electrophoresis technology for proteomics. *Proteomics* **2004**, *4* (12), 3665–3685.

(7) Kieffer, P.; Dommes, J.; Hoffmann, L.; Hausman, J. F.; Renaut, J. Quantitative changes in protein expression of cadmium-exposed poplar plants. *Proteomics* **2008**, *8* (12), 2514–2530.

(8) Giavalisco, P.; Kapitza, K.; Kolasa, A.; Buhtz, A.; Kehr, J. Towards the proteome of Brassica napus phloem sap. *Proteomics* **2006**, *6* (3), 896–909.

(9) Schiltz, S.; Gallardo, K.; Huart, M.; Negroni, L.; Sommerer, N.; Burstin, J. Proteome reference maps of vegetative tissues in pea. An investigation of nitrogen mobilization from leaves during seed filling. *Plant Physiol.* **2004**, *135* (4), 2241–2260.

(10) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **1999**, *17* (7), 676–682.

(11) Washburn, M. P.; Wolters, D.; Yates, J. R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001**, *19* (3), 242–247.

(12) VerBerkmoes, N. C.; Shah, M. B.; Lankford, P. K.; Pelletier, D. A.; Strader, M. B.; Tabb, D. L.; McDonald, W. H.; Barton, J. W.; Hurst, G. B.; Hauser, L.; Davison, B. H.; Beatty, J. T.; Harwood, C. S.; Tabita, F. R.; Hettich, R. L.; Larimer, F. W. Determination and comparison of the baseline proteomes of the versatile microbe Rhodopseudomonas palustris under its major metabolic states. *J. Proteome Res.* **2006**, *5* (2), 287–298.

(13) Koller, A.; Washburn, M. P.; Lange, B. M.; Andon, N. L.; Deciu, C.; Haynes, P. A.; Hays, L.; Schieltz, D.; Ulaszek, R.; Wei, J.; Wolters, D.; Yates, J. R. Proteomic survey of metabolic pathways in rice. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (18), 11969–11974.

(14) Baerenfaller, K.; Grossmann, J.; Grobei, M. A.; Hull, R.; Hirsch-Hoffmann, M.; Yalovsky, S.; Zimmermann, P.; Grossniklaus, U.; Gruissem, W.; Baginsky, S. Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science* **2008**, *320* (5878), 938–941.

(15) Kalluri, U. C.; Hurst, G. B.; Lankford, P. K.; Ranjan, P.; Pelletier, D. A. Shotgun proteome profile of Populus developing xylem. *Proteomics* **2009**, *9* (21), 4871–4880.

(16) Jansson, S.; Douglas, C. J. Populus: a model system for plant biology. *Annu. Rev. Plant Biol.* **2007**, *58*, 435–458.

(17) Plomion, C.; Lalanne, C.; Claverol, S.; Meddour, H.; Kohler, A.; Bogeat-Triboulot, M. B.; Barre, A.; Le Provost, G.; Dumazet, H.; Jacob, D.; Bastien, C.; Dreyer, E.; de Daruvar, A.; Guehl, J. M.; Schmitter, J. M.; Martin, F.; Bonneu, M. Mapping the proteome of poplar and application to the discovery of drought-stress responsive proteins. *Proteomics* **2006**, *6* (24), 6509–6527.

(18) Bylesjo, M.; Nilsson, R.; Srivastava, V.; Gronlund, A.; Johansson, A. I.; Jansson, S.; Karlsson, J.; Moritz, T.; Wingsle, G.; Trygg, J. Integrated Analysis of Transcript, Protein and Metabolite Data To Study Lignin Biosynthesis in Hybrid Aspen. *J. Proteome Res.* **2009**, *8* (1), 199–210.

(19) Nilsson, R.; Bernfur, K.; Gustavsson, N.; Bygdell, J.; Wingsle, G.; Larsson, C. Proteomics of Plasma Membranes from Poplar Trees Reveals Tissue Distribution of Transporters, Receptors, and Proteins in Cell Wall Formation. *Mol. Cell. Proteomics* **2010**, *9* (2), 368–387.

(20) Visioli, G.; Marmiroli, M.; Marmiroli, N. Two-Dimensional Liquid Chromatography Technique Coupled with Mass Spectrometry Analysis to Compare the Proteomic Response to Cadmium Stress in Plants. *J. Biomed. Biotechnol.* **2010**, DOI: 10.1155/2010/567510.

(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* (11), 976–989.

(22) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551–3567.

(23) Nedelkov, D. Population proteomics: investigation of protein diversity in human populations. *Proteomics* **2008**, *8* (4), 779–786.

(24) Zybailov, B.; Sun, Q.; van Wijk, K. J. Workflow for Large Scale Detection and Validation of Peptide Modifications by RPLC-LTQ-Orbitrap: Application to the Arabidopsis thaliana Leaf Proteome and an Online Modified Peptide Library. *Anal. Chem.* **2009**, *81* (19), 8015–8024.

(25) Foston, M.; Hubbell, C.; Samuel, R.; Jung, S.; Fan, H.; Ding, S.-Y.; Zeng, Y.; Jawdy, S.; Davis, M.; Sykes, S.; Gjersing, E.; Tuskan, G. A.; Kalluri, U.; Ragauskas, A. J. Chemical, ultrastructural and supramolecular analysis of tension wood in Populus tremula x alba as a model substrate for reduced recalcitrance. *Energy Environ. Sci* **2011**, DOI: 10.1039/C1EE02073.

(26) Gion, J. M.; Lalanne, C.; Le Provost, G.; Ferry-Dumazet, H.; Paiva, J.; Chaumeil, P.; Frigerio, J. M.; Brach, J.; Barre, A.; de Daruvar, A.; Claverol, S.; Bonneu, M.; Sommerer, N.; Negroni, L.; Plomion, C. The proteome of maritime pine wood forming tissue. *Proteomics* **2005**, *5* (14), 3731–3751.

(27) Lowry, O. H.; Rosebrough, N. J.; Farr, A. L.; Randall, R. J. Protein measurement with the Folin phenol reagent. *J. Biol. Chem.* **1951**, *193* (1), 265–275.

(28) Brown, S. D.; Thompson, M. R.; Verberkmoes, N. C.; Chourey, K.; Shah, M.; Zhou, J.; Hettich, R. L.; Thompson, D. K. Molecular dynamics of the Shewanella oneidensis response to chromate stress. *Mol. Cell. Proteomics* **2006**, *5* (6), 1054–1071.

(29) Yang, X. H.; Tschaplinski, T. J.; Hurst, G. B.; Jawdy, S.; Abraham, P. E.; Lankford, P. K.; Adams, R. M.; Shah, M. B.; Hettich, R. L.; Lindquist, E.; Kalluri, U. C.; Gunter, L. E.; Pennacchio, C.; Tuskan, G. A. Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome Res.* **2011**, *21* (4), 634–641.

(30) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC−MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2003**, *2* (1), 43–50.

(31) Tabb, D. L.; McDonald, W. H.; Yates, J. R., 3rd DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **2002**, *1* (1), 21–26.

(32) Zybailov, B.; Coleman, M. K.; Florens, L.; Washburn, M. P. Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Anal. Chem.* **2005**, *77* (19), 6218–6224.

(33) Zybailov, B.; Mosley, A. L.; Sardiu, M. E.; Coleman, M. K.; Florens, L.; Washburn, M. P. Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J. Proteome Res.* **2006**, *5* (9), 2339–2347.

(34) Gammulla, C. G.; Pascovici, D.; Atwell, B. J.; Haynes, P. A. Differential metabolic response of cultured rice (Oryza sativa) cells exposed to high- and low-temperature stress. *Proteomics* **2010**, *10* (16), 3001–3019.

(35) Giannone, R. J.; Huber, H.; Karpinets, T.; Heimerl, T.; Kuper, U.; Rachel, R.; Keller, M.; Hettich, R. L.; Podar, M. Proteomic Characterization of Cellular and Molecular Processes that Enable the Nanoarchaeum equitans-Ignicoccus hospitalis Relationship. *PLoS One* **2011**, *6* (8), e22942.

(36) Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **2010**, *26* (19), 2460–2461.

(37) Tuskan, G. A.; DiFazio, S.; Jansson, S.; et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **2006**, *313* (5793), 1596–1604.

(38) Zhang, Y.; Wen, Z.; Washburn, M. P.; Florens, L. Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. *Anal. Chem.* **2010**, *82* (6), 2272–2281.

(39) Mallick, P.; Schirle, M.; Chen, S. S.; Flory, M. R.; Lee, H.; Martin, D.; Ranish, J.; Raught, B.; Schmitt, R.; Werner, T.; Kuster, B.; Aebersold, R. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **2007**, *25* (1), 125–131.

(40) Rappsilber, J.; Mann, M. What does it mean to identify a protein in proteomics? *Trends Biochem. Sci.* **2002**, *27* (2), 74–78.

(41) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75* (17), 4646–4658.

(42) Nesvizhskii, A. I.; Aebersold, R. Interpretation of shotgun proteomic data - The protein inference problem. *Mol. Cell. Proteomics* **2005**, *4* (10), 1419–1440.

(43) Delalande, F.; Carapito, C.; Brizard, J. P.; Brugidou, C.; Van Dorsselaer, A. Multigenic families and proteomics: extended protein characterization as a tool for paralog gene identification. *Proteomics* **2005**, *5* (2), 450–460.

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

(45) Yang, X.; Dondeti, V.; Dezube, R.; Maynard, D. M.; Geer, L. Y.; Epstein, J.; Chen, X.; Markey, S. P.; Kowalak, J. A. DBParser: web-based software for shotgun proteomic data analyses. *J. Proteome Res.* **2004**, *3* (5), 1002–1008.

(46) Friso, G.; Majeran, W.; Huang, M.; Sun, Q.; van Wijk, K. J. Reconstruction of metabolic pathways, protein expression, and homeostasis machineries across maize bundle sheath and mesophyll chloroplasts: large-scale quantitative proteomics using the first maize genome assembly. *Plant Physiol.* **2010**, *152* (3), 1219–1250.

(47) Meyer-Arendt, K.; Old, W. M.; Houel, S.; Renganathan, K.; Eichelberger, B.; Resing, K. A.; Ahn, N. G. IsoformResolver: a peptide-centric algorithm for protein inference. *J. Proteome Res.* **2011**, *10* (7), 3060–3075.

(48) Wu, C. H.; Yeh, L. S.; Huang, H.; Arminski, L.; Castro-Alvear, J.; Chen, Y.; Hu, Z.; Kourtesis, P.; Ledley, R. S.; Suzek, B. E.; Vinayaka, C. R.; Zhang, J.; Barker, W. C. The Protein Information Resource. *Nucleic Acids Res.* **2003**, *31* (1), 345–347.

(49) Fisher, K.; Turner, S. PXY, a receptor-like kinase essential for maintaining polarity during plant vascular-tissue development. *Curr. Biol.* **2007**, *17* (12), 1061–1066.

(50) Hirakawa, Y.; Shinohara, H.; Kondo, Y.; Inoue, A.; Nakanomyo, I.; Ogawa, M.; Sawa, S.; Ohashi-Ito, K.; Matsubayashi, Y.; Fukuda, H. Non-cell-autonomous control of vascular stem cell fate by a CLE peptide/receptor system. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (39), 15208–15213.

(51) Mora-Garcia, S.; Vert, G.; Yin, Y.; Cano-Delgado, A.; Cheong, H.; Chory, J. Nuclear protein phosphatases with Kelch-repeat domains modulate the response to brassinosteroids in Arabidopsis. *Genes Dev.* **2004**, *18* (4), 448–460.

(52) Yamamoto, R.; Fujioka, S.; Demura, T.; Takatsuto, S.; Yoshida, S.; Fukuda, H. Brassinosteroid levels increase drastically prior to morphogenesis of tracheary elements. *Plant Physiol.* **2001**, *125* (2), 556–563.

(53) Cano-Delgado, A.; Yin, Y.; Yu, C.; Vafeados, D.; Mora-Garcia, S.; Cheng, J. C.; Nam, K. H.; Li, J.; Chory, J. BRL1 and BRL3 are novel brassinosteroid receptors that function in vascular differentiation in Arabidopsis. *Development* **2004**, *131* (21), 5341–5351.

(54) Nakamura, A.; Fujioka, S.; Sunohara, H.; Kamiya, N.; Hong, Z.; Inukai, Y.; Miura, K.; Takatsuto, S.; Yoshida, S.; Ueguchi-Tanaka, M.; Hasegawa, Y.; Kitano, H.; Matsuoka, M. The role of OsBRI1 and its homologous genes, OsBRL1 and OsBRL3, in rice. *Plant Physiol.* **2006**, *140* (2), 580–590.

(55) Fukuda, H. Signals that control plant vascular cell differentiation. *Nat. Rev. Mol. Cell. Biol.* **2004**, *5* (5), 379–391.

(56) Sauer, M.; Paciorek, T.; Benkova, E.; Friml, J. Immunocytochemical techniques for whole-mount in situ protein localization in plants. *Nat. Protoc.* **2006**, *1* (1), 98–103.

(57) Scarpella, E.; Marcos, D.; Friml, J.; Berleth, T. Control of leaf vascular patterning by polar auxin transport. *Genes Dev.* **2006**, *20* (8), 1015–1027.

(58) Tuskan, G. A.; Walsh, M. E. Short-rotation woody crop systems, atmospheric carbon dioxide and carbon management: A US case study. *Forest Chron.* **2001**, *77* (2), 259–264.

(59) Davison, B. H.; Drescher, S. R.; Tuskan, G. A.; Davis, M. F.; Nghiem, N. P. Variation of S/G ratio and lignin content in a Populus family influences the release of xylose by dilute acid hydrolysis. *Appl. Biochem. Biotechnol.* **2006**, *130* (1−3), 427–435.

(60) Dinus, R. J.; Payne, P.; Sewell, N. M.; Chiang, V. L.; Tuskan, G. A. Genetic modification of short rotation popular wood: Properties for ethanol fuel and fiber productions. *Crit. Rev. Plant Sci.* **2001**, *20* (1), 51–69.

(61) Sannigrahi, P.; Ragauskas, A. J.; Tuskan, G. A. Poplar as a feedstock for biofuels: A review of compositional characteristics. *Biofuels Bioprod. Biorefin.* **2010**, *4* (2), 209–226.

(62) Cantarel, B. L.; Coutinho, P. M.; Rancurel, C.; Bernard, T.; Lombard, V.; Henrissat, B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* **2009**, *37*, D233–D238.

(63) Aspeborg, H.; Schrader, J.; Coutinho, P. M.; Stam, M.; Kallas, A.; Djerbi, S.; Nilsson, P.; Denman, S.; Amini, B.; Sterky, F.; Master, E.;

Sandberg, G.; Mellerowicz, E.; Sundberg, B.; Henrissat, B.; Teeri, T. T. Carbohydrate-active enzymes involved in the secondary cell wall biogenesis in hybrid aspen. *Plant Physiol.* **2005**, *137* (3), 983–997.

(64) Geisler-Lee, J.; Geisler, M.; Coutinho, P. M.; Segerman, B.; Nishikubo, N.; Takahashi, J.; Aspeborg, H.; Djerbi, S.; Master, E.; Andersson-Gunneras, S.; Sundberg, B.; Karpinski, S.; Teeri, T. T.; Kleczkowski, L. A.; Henrissat, B.; Mellerowicz, E. J. Poplar carbohydrate-active enzymes. Gene identification and expression analyses. *Plant Physiol.* **2006**, *140* (3), 946–962.

(65) Andersson-Gunneras, S.; Mellerowicz, E. J.; Love, J.; Segerman, B.; Ohmiya, Y.; Coutinho, P. M.; Nilsson, P.; Henrissat, B.; Moritz, T.; Sundberg, B. 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* (2), 144–165.

(66) Shi, R.; Sun, Y. H.; Li, Q. Z.; Heber, S.; Sederoff, R.; Chiang, V. L. Towards a systems approach for lignin biosynthesis in *Populus trichocarpa*: transcript abundance and specificity of the monolignol biosynthetic genes. *Plant Cell Physiol.* **2010**, *51* (1), 144–163.

(67) Johnson, R. S.; Davis, M. T.; Taylor, J. A.; Patterson, S. D. Informatics for protein identification by mass spectrometry. *Methods* **2005**, *35* (3), 223–236.

(68) Verberkmoes, N. C.; Hervey, W. J.; Shah, M.; Land, M.; Hauser, L.; Larimer, F. W.; Van Berkel, G. J.; Goeringer, D. E. Evaluation of "shotgun" proteomics for identification of biological threat agents in complex environmental matrixes: experimental simulations. *Anal. Chem.* **2005**, *77* (3), 923–932.

(69) Bern, M.; Goldberg, D.; McDonald, W. H.; Yates, J. R., 3rd Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics* **2004**, *20* (Suppl 1), i49–54.

(70) Salmi, J.; Moulder, R.; Filen, J. J.; Nevalainen, O. S.; Nyman, T. A.; Lahesmaa, R.; Aittokallio, T. Quality classification of tandem mass spectrometry data. *Bioinformatics* **2006**, *22* (4), 400–406.

(71) Sachidanandam, R.; Weissman, D.; Schmidt, S. C.; Kakol, J. M.; Stein, L. D.; Marth, G.; Sherry, S.; Mullikin, J. C.; Mortimore, B. J.; Willey, D. L.; Hunt, S. E.; Cole, C. G.; Coggill, P. C.; Rice, C. M.; Ning, Z.; Rogers, J.; Bentley, D. R.; Kwok, P. Y.; Mardis, E. R.; Yeh, R. T.; Schultz, B.; Cook, L.; Davenport, R.; Dante, M.; Fulton, L.; Hillier, L.; Waterston, R. H.; McPherson, J. D.; Gilman, B.; Schaffner, S.; Van Etten, W. J.; Reich, D.; Higgins, J.; Daly, M. J.; Blumenstiel, B.; Baldwin, J.; Stange-Thomann, N.; Zody, M. C.; Linton, L.; Lander, E. S.; Altshuler, D. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **2001**, *409* (6822), 928–933.

(72) Tanner, S.; Shu, H.; Frank, A.; Wang, L. C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **2005**, *77* (14), 4626–4639.

(73) Louie, G. V.; Bowman, M. E.; Moffitt, M. C.; Baiga, T. J.; Moore, B. S.; Noel, J. P. Structural determinants and modulation of substrate specificity in phenylalanine-tyrosine ammonia-lyases. *Chem. Biol.* **2006**, *13* (12), 1327–1338.

(74) Howles, P. A.; Sewalt, V.; Paiva, N. L.; Elkind, Y.; Bate, N. J.; Lamb, C.; Dixon, R. A. Overexpression of L-Phenylalanine Ammonia-Lyase in Transgenic Tobacco Plants Reveals Control Points for Flux into Phenylpropanoid Biosynthesis. *Plant Physiol.* **1996**, *112* (4), 1617–1624.