Journal of proteome research

Article

pubs.acs.org/jpr

Moving Away from the Reference Genome: Evaluating a Peptide Sequencing Tagging Approach for Single Amino Acid Polymorphism Identifications in the Genus *Populus*

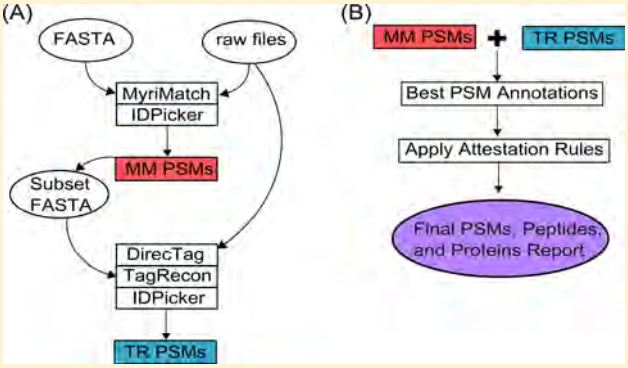
Paul Abraham,[†,‡] Rachel M. Adams,[†,‡] Gerald A. Tuskan,[§] and Robert L. Hettich*[,‡]

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

[‡]Chemical Sciences Division and [§]Biological Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States

Supporting Information

ABSTRACT: The genetic diversity across natural populations of the model organism, *Populus*, is extensive, containing a single nucleotide polymorphism roughly every 200 base pairs. When deviations from the reference genome occur in coding regions, they can impact protein sequences. Rather than relying on a static reference database to profile protein expression, we employed a peptide sequence tagging (PST) approach capable of decoding the plasticity of the *Populus* proteome. Using shotgun proteomics data from two genotypes of *P. trichocarpa*, a tag-based approach enabled the detection of 6653 unexpected sequence variants. Through manual validation, our study investigated how the most abundant chemical modification (methionine oxidation) could masquerade as a sequence variant (Ala→Ser) when few site-determining ions existed. In fact, precise localization of an oxidation site for peptides with more than one potential placement was indeterminate for 70% of the MS/MS spectra. We demonstrate that additional fragment ions made available by high energy collisional dissociation enhances the robustness of the peptide sequence tagging approach (81% of oxidation events could be exclusively localized to a methionine). We are confident that augmenting fragmentation processes for a PST approach will further improve the identification of single amino acid polymorphism in *Populus* and potentially other species as well.



KEYWORDS: *plant proteomics, Populus, single amino acid polymorphisms, mass spectrometry, peptide sequence tagging, high energy collisional dissociation*

1. INTRODUCTION

In 2006, black cottonwood (*Populus trichocarpa*) became the first tree to have its genome fully sequenced and thus emerged as a model for tree genomics.[1] In contrast to other plant models such as *Arabidopsis* and rice, which are predominately self-fertilizing and consequently maintain low levels of allelic polymorphism, the *Populus* genus is primarily composed of dioecious, self-incompatible woody plants.[2] Obligate out-crossing combined with wind-pollination and prolonged reproductive life generates highly heterozygous populations with low levels of linkage disequilibrium. This type of mating system results in high levels of gene flow and extensive nucleotide variability within and across *Populus* species, providing an excellent model system to investigate the relationship between naturally occurring single-nucleotide polymorphisms (SNPs) and phenotypic variation.[3]

Through association genetics, the discovery of nucleotide variations among genotypes has the potential to reveal allelic polymorphisms underlying complex, adaptive traits. SNPs can be located either within a protein-coding region or outside coding regions. On average, SNP frequency in protein-coding regions is high in forest trees, generally in the order of 1 per 1000 base pairs; SNP frequency in *Populus* is somewhat higher, with an estimate of 1 SNP every 200 base pairs.[4] Nucleotide polymorphisms that occur inside coding regions may (non-synonymous) or may not (synonymous) change the amino acid sequence of the corresponding protein. Because synonymous changes are largely invisible to selective pressure and have little biological implications, they are categorized as silent nucleotide variations. On the other hand, nonsynonymous changes can be under strong selective pressure and because they can directly impact gene function, they are the primary focus of most physiological or pathological association studies.[5]

Although SNPs in *Populus* have been extensively studied over the past decade, little attention has been paid to single amino acid polymorphisms (SAAPs) of proteins at the proteome level. In fact, only a few efforts have been made to survey SAAPs

Received: March 4, 2013
Published: June 24, 2013

across the *Populus* proteome. In one particular study, tandem mass spectrometry (MS/MS)-based shotgun proteomics was employed for large-scale proteome analysis.[6] In general, the available protein databases used for such studies are incomplete with respect to sequence variation information. Without taking SNP variations into account, proteomic investigations generally fail to identify any protein form containing a SAAP. One of our previous studies was able to append predicted protein sequence variations to the original database and detect novel protein forms.[6] The main disadvantage of this approach, however, is that *a priori* knowledge of SNPs is required. Moreover, this approach is preconditioned on both the coverage and quality of the predictions when they are available. Therefore, we argue that a more attractive approach considers unexpected single amino acid polymorphisms.

The high-throughput discovery of protein sequence variants (truncations, post-translational modifications, or mutations), especially unexpected variants, has seen tremendous advancements in recent years.[7] Many database sequencing algorithms have been recently designed to effectively identify unanticipated (blind) sequence variants at a global level. One class of such algorithms use *de novo* sequencing to infer full-length peptide sequences from tandem mass spectra without requiring a sequence reference database.[8−10] A strength of this approach is that the concept of variant peptides is not relevant; each spectrum is given an equal opportunity to match any combination of amino acids, regardless of whether the researcher anticipated detecting the sequence or not. This technique, however, greatly increases the number of candidate peptides compared to each spectrum, consequently incurring not only significant costs to processing time but also unacceptable false discovery rates (FDR).[11] In addition, mass spectrometrists have developed and routinely used a hybrid approach between traditional database searching and *de novo* approaches: here peptide sequence tagging (PST) algorithms can detect unexpected sequence variants as extensions of partial sequences identified from in a database.[12−16] In particular, the proteome informatics group led by David Tabb recently released a two-step methodology involving the DirecTag[16] algorithm for highly accurate PST tag generation, followed by the TagRecon[17] software for the detection of peptide sequence variants through tag reconciliation. In brief, short sequence "tags" are directly inferred from a tandem mass spectrum, and then tags are automatically reconciled against representative peptides from a protein database while making allowances for unexpected mass shifts (i.e., mutations and post-translational modifications).[18] PSTs serve as a filter to effectively reduce the number peptide-spectrum matches being scored, which in turn improves costs in processing time, sensitivity, and specificity.[19]

To evaluate a peptide sequence tagging approach for *Populus* with the ultimate goal of globally identifying unknown SAAPs, we employed DirecTag and TagRecon software. Using the state-of-the-art LTQ Orbitrap Pro platform, we profiled and compared two genotypes of *P. trichocarpa* and identified a large number of unexpected peptide variants that would have otherwise been missed by a traditional database search. The sequence variants leveraged from TagRecon demonstrates the value of using peptide sequence tagging algorithms to interrogate proteomics data sets, provided that a SAAP location could be confidently identified. Therefore, while our initial aim was to comprehensively identify SAAPs, we focused on our most abundant sequence variant to show that confident site localization remains an important yet challenging task. Because others have shown that HCD fragmentation improves the coverage of peptide sequences overall, in particular for tryptic peptides up to 15 amino acids in length, we exploited HCD fragmentation to further refine a subset of the data set.

2. EXPERIMENTAL PROCEDURES

2.1. Plant Material

Two *Populus trichocarpa* genotypes, "DENA" and "VNDL", were grown under standard greenhouse conditions as previously outlined.[20] From these trees, 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. Tissue was harvested from six individual ramets per genotype and pooled together for each sample tissue type to reduce the effects of biological variation.

2.2. Protein Extraction

Leaf, root, and stem tissues were ground under liquid nitrogen using a mortar and pestle. For each tissue type, a 1 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 (Branson model SSE-1, 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 precleared via centrifugation at 4500*g* for 10 min, quantified by BCA assay (Pierce Biotechnology), and reduced with 25 mM dithiothreitol (DTT). Three milligrams of crude protein extract was then precipitated by trichloroacetic acid (TCA), pelleted by centrifugation, and washed with ice-cold acetone to remove excess SDS as previously described.[6]

2.3. Protein Digestion

For in-solution protein digestions, pelleted proteins were 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. Denatured proteins were reduced with DTT (5 mM), and cysteines were blocked with 20 mM iodoacetamide at room temperature for 20 min to prevent reformation of disulfide bonds. At a temperature of 37 °C and pH 8.0, protein samples were digested into peptides by adding two aliquots of sequencing-grade trypsin (Promega, 1:75 [w/w]) at two different sample dilutions, 4 M urea (overnight) and subsequent 2 M urea (5 h). 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.

2.4. LC-MS/MS

For the analysis of the proteome samples, 25 μg of each peptide mixture were bomb-loaded onto a biphasic MudPIT back column packed with ∼5 cm strong cation exchange (SCX) resin followed by ∼3 cm C18 reversed phase (RP) (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 nanoelectrospray emitter (100 μm ID) packed with 15 cm of C18 RP material and analyzed via 24-h MudPIT 2D-LC-MS/MS as previously described.[6] Peptide sequencing analysis was performed by the LTQ Orbitrap Pro mass spectrometer (Thermo Scientific). For

each sample, three technical replicates were performed. Peptides were fragmented by CID or HCD, but in either case precursor ions were measured in the Orbitrap analyzer to obtain high resolution spectra (15 000 at m/z 400). Mass spectra were acquired in a data-dependent "top 20" mode: each survey scan was followed by MS/MS spectra of the 20 most abundant precursor ions (3 m/z isolation window). For peptide fragmentation, normalized collision energy of 35 and 40 eV was used for CID and HCD, respectively. Each fragmented precursor ion was dynamically excluded from targeting for 60 s. A dynamic exclusion repeat of 1 and a mass width of 0.2 m/z were applied to maximize peptide sequencing. All high resolution (7500 at m/z 400) MS/MS spectra were acquired in the Orbitrap analyzer (XCalibur version 2.1).

2.5. Peptide Identification

All 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* (v3, released in 2012, available at http://www.phytozome.net/cgi-bin/gbrowse/poplar/, containing 73 013 primary and alternate spliced gene models), (2) mitochondria and chloroplast proteins,[1] and (3) common contaminant proteins (i.e., porcine trypsin, human keratin, etc.). A decoy database, consisting of the reversed sequences of the target database, was appended to evaluate the false-discovery rate (FDR) at the peptide level. For standard database searching, the peptide fragmentation spectra (MS/MS) were searched with MyriMatch algorithm v2.1. MyriMatch was configured to derive fully tryptic peptides and consider the following mass shifts: a static modification on cysteine (+57.02 Da) and an N-terminal dynamic modification of +43.00 Da (carbamylation). For the directed searches, MyriMatch was configured to consider a dynamic modification corresponding to an oxidation (+15.99 Da) on either a methionine or alanine. Verbose detailed lists of all the configuration parameters used for both standard and directed searches are listed in Supporting Information.

For peptide sequence tagging, DirecTag was parametrized to reduce MS/MS spectra to the top 100 peaks. For identifiable spectra that yielded partial sequence tags (tag length = 3), DirecTag stored the top 50 sequence tags and their scores into a tab-delimited file. TagRecon reconciled the inferred sequence tags against a subset protein database (i.e., proteins identified by MyriMatch) while making allowances for 2 mass shifts. TagRecon was configured (mutation mode) to consider only one unanticipated mass shift corresponding to amino acid substitutions using the BLOSUM62 matrix. In addition to mutations, TagRecon was configured to consider the following mass shifts: a static modification on cysteine (+57.02 Da) and an N-terminal dynamic modification of +43.00 Da. A detailed list of the configuration parameters used for DirecTag and TagRecon are listed in Supporting Information.

2.6. Protein Inference

IDPicker filtered the resulting peptide-spectrum matches (PSMs) from all searches at a 2% FDR. While search algorithms rigorously assess the statistical significance of each PSM, high-throughput validation of modified peptides remains an open problem. In this study, we applied tested attestation principles for validating modified peptides in a complex mixture.[21] To obtain a data set of the highest quality, we enforced the following filtering guidelines:

(1) Mutated peptides were removed if they mapped to a contaminant protein.
(2) Mutations of lysine or arginine residues cannot occur at trypsin cut sites.
(3) If a spectrum matched to a mutated peptide (TagRecon) as well as a nonmutant (MyriMatch) peptide, the mutated PSM must improve upon the score of the unmodified PSM by 10%.
(4) A distinct mutated peptide sequence must match to at least three different spectra.
(5) Mutations that can also be explained as common sampling processing artifacts were removed: these included the deamidation (+0.984 Da) of asparagine or glutamine, dehydration (−18.01 Da) of aspartate and glutamate, formylation (+27.99 Da) of threonine or serine, and the oxidation of methionine (+15.99 Da).

Only peptides passing the FDR threshold and the above guidelines were considered for further analysis. Protein identifications with at least two distinct peptide identifications were considered for further analysis. To deal with the redundancy associated with the *Populus* genome, all proteins in the FASTA database (includes *P. trichocarpa* v3, mitochondria and chloroplast predicted proteomes) were grouped by sequence similarity (≥90%) using the UCLUST component of the USEARCH v5.0 software platform. As described in our previous study, grouping proteins by this conservative level of sequence identity serves to maintain biologically relevant peptides.[6]

3. RESULTS AND DISCUSSION

A single polymorphism can potentially impact protein stability and the molecular interactions that dictate protein function. Using standard database searching algorithms, a polymorphism associated with a specific phenotypic variant cannot be identified. As a result, a considerable portion of high-quality tandem mass spectra collected is left unassigned. In an effort to identify sequence variations, we utilized DirecTag and TagRecon to extract and infer peptide sequence tags that match unexpected single amino acid polymorphisms in *Populus*. We employed a recently established experimental strategy that yielded superior protein extraction and identification. For maximal spectral acquisition, high mass-accuracy, and high resolution, we used a dual-pressure linear ion trap analyzer coupled with the Orbitrap analyzer (LTQ Orbitrap Pro). We systematically investigated SAAPs across two *P. trichocarpa* genotypes, DENA and VNDL, to test the value of the sequence tag-based approach in a broader biological context.

3.1. Peptide Identification Using a Standard Database Algorithm

For this study, proteome extracts from three tissues (leaf, root, and stem) were harvested from two *P. trichocarpa* genotypes, DENA and VNDL, and analyzed in triplicate on an LTQ-Orbitrap-Pro mass spectrometer. Using standard parameters, the collected tandem mass spectra (MS/MS) were searched with MyriMatch[22] against the *P. trichocarpa* v3.0 protein database and supplemented with the chloroplast and mitochondrial proteomes. We employed IDPicker[13] to filter the resulting peptide-spectra matches at a maximum FDR of 2% (PSM level) and assemble peptides into a list of proteins (Supporting Information Tables 1 and 2 (SI files 3, 5)). Overall, 69 613 distinct peptide sequences were detected across the entire MyriMatch data set. Because a considerable portion of the observed peptides are shared among multiple proteins, assigning peptides to their respective proteins is a considerable

challenge in *Populus*. As highlighted in a previous study, we recommend addressing this by incorporating additional supporting information (i.e., sequence homology) to better infer the existence of proteins in the sample.[6] Therefore, proteins sharing 90% or more sequence identity within the *Populus* database were collapsed into protein groups.

Of the original 25 550 redundant proteins observed, a total of 9601 protein groups were identified and of those, 3399 were singletons (i.e., one-membered groups). Because both genotypes were grown under identical growth conditions, we expected to observe substantial overlap in the proteins that were expressed in both genotypes. Indeed, the measured proteins for both VNDL and DENA shared a high level of overlap (~80%). For the purpose of evaluating the depth of coverage achieved, we compared the number of protein groups identified against our recently published study that had achieved the deepest proteome coverage in the genus *Populus*.[23] Overall, >2000 additional protein groups were detected in the current study.

3.2. Identification of Sequence Variants Using Peptide Sequencing Tagging

Sequence variations, manifested by single amino acid polymorphisms, provide clues to the genetic structures that induce a pathological or physiological trait. To our knowledge, SNPs are widely measured at the transcriptome level but rarely at the proteome level. For the reasons outlined above, we employed a peptide-sequence tagging approach to identify SAAPs in *Populus*.

Figure 1 illustrates the three-step experimental workflow used to identify unexpected sequence variants in *Populus*. The first step uses the MyriMatch search engine to identify a confident list of proteins (no unexpected sequence variants considered) for each biological sample (Figure 1A), and IDPicker was employed to ensure that only confident identifications were retained. This step serves to dramatically reduce the candidate list of proteins (a subset FASTA database) for the blind search that follows, with the purpose of improving processing time, sensitivity, and specificity of the analysis. In the second step, DirecTag infers sequence tags from the MS/MS scans from each raw file, followed by TagRecon mass matching the inferred sequences to the subset protein database while making allowances for unanticipated mass shifts in peptides. IDPicker was employed to filter the resulting peptide-spectra matches at a maximum FDR of 2% (PSM-level) and assemble peptides into a list of proteins (Figure 1A and Supporting Information Tables 3 and 4 (SI files 4, 6)). For the final step, peptide-spectrum matches observed in MyriMatch and TagRecon were compared to obtain a final data set of the highest quality (Figure 1B and Supporting Information Table 5 (SI files 7−12)). In addition, several proven attestation principles[21] were applied to further validate peptide sequence variants (see Experimental Procedures). Table 1 presents a summary of the attested results after merging the data from the two database search engines.

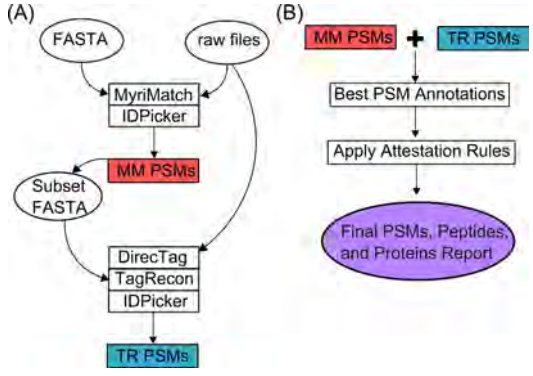


Figure 1. Computational workflow for the identification of peptide sequence variants. This flowchart illustrates the three-step strategy used to identify peptide sequence variants from a DirecTag and TagRecon approach. (A) First, a traditional database search was performed by MyriMatch to confidently identify a list of proteins, which was used to minimize the search database. In the second step, the DirecTag software provides an inferred sequence tag for every MS/MS spectra. TagRecon then reconciles the inferred sequence tags against the subset database to identify unexpected mass shifts in peptides sequences. Following every search, the IDPicker software applied a variety of score combinations to filter the resulting identifications at 2% FDR. (B) The MyriMatch and TagRecon results were compared to identify the best PSM for every MS/MS spectra. The data set was further refined by validating every sequence variant with proven attestation rules.

Table 1. Results after Merging MyriMatch and TagRecon Data Sets

	VNDL leaf	DENA leaf
averaged spectra	144117	153420
summed spectra	432352	460259
peptides	23495	25200
distinct peptides	20944	22694
proteins	13688	13530
protein groups (≥90%)	4662	4643
peptide variants	2362	2800
distinct peptide variants	1834	2161
peptide variant summed spectra	31045	31366
	VNDL stem	DENA stem
averaged spectra	152954	176356
summed spectra	458863	529069
peptides	26870	27976
distinct peptides	24610	25705
proteins	15725	15480
protein groups (≥90%)	5500	5374
peptide variants	2142	2291
distinct peptide variants	1744	1822
peptide variant summed spectra	24512	24784
	VNDL root	DENA root
averaged spectra	180226	182388
summed spectra	540677	547164
peptides	29426	33766
distinct peptides	27380	31099
proteins	15210	16476
protein groups (≥90%)	5334	5827
peptide variants	1934	2257
distinct peptide variants	1537	1790
peptide variant summed spectra	20047	22205

In general, the percentage of identified peptides (frequency) and spectra (abundance) containing a sequence variant did not seem dependent on the genotype (Table 2). We identified a total of 6653 peptide sequence variants (~10% of total identified peptides); 4391 and 4900 sequence variants for VNDL and DENA, respectively. Overall, these sequences mapped to 22 067 proteins and 8088 protein groups, which means a peptide sequence variant was identified in 86% of the

dx.doi.org/10.1021/pr400192r | J. Proteome Res. 2013, 12, 3642−3651

Table 2. Frequency and Abundance of Sequence Variants in *Populus*

genotype/organ	peptides	spectra	frequency, %	abundance, %
VNDL/leaf	2362	31045	11.3	7.2
VNDL/stem	2142	24512	8.7	5.3
VNDL/root	1934	20047	7.1	3.7
DENA/leaf	2800	31366	12.3	6.4
DENA/stem	2291	24784	8.9	4.7
DENA/root	2257	22205	7.3	4.1

proteins observed and 84% of protein groups. Although the percentage of peptide sequence variants identified seems relatively small, this can be explained by the experimental aspects of the approach. In general, the median sequence coverage observed in shotgun proteomic experiments employing a trypsin-based schema is often between 20% and 25%.[24] Consequently, we anticipated a limited sampling of SAAPs across individual proteins. Nevertheless, we identified a sequence variant for nearly every protein detected. Future studies may be warranted to specifically focus on achieving maximal sequence coverage by modifying the experimental strategy to incorporate multiple proteases,[25,26] which would provide more specificity to the frequencies of SAAPs per protein.

3.3. Types of Variant Peptides in *Populus*

The procedure described above identified a total of 76 types of sequence variants (each type denoting an amino acid with a mass shift corresponding to a mutation). Noticeably, the occurrence of variants in both genotypes is similar (Pearson correlation = 0.99). A complete list of the variants and their number of peptides and abundances observed are in Supporting Information Table 6 (SI files 13, 14). Of those listed, the top 20 most abundant have been highlighted in Table 3.

Peptides and fragment ions containing an oxidation mass shift (+15.99 Da) were the most prevalent variant type, representing ~38% of the total assigned spectra for variant peptides. While this observation may suggest that the two most prominent SAAPs are Ala→Ser and Phe→Tyr, we critically evaluated the results by validating each variant through manual verification of the MS/MS spectra. In the course of this inspection, we observed that the site of +16 Da mass shifts were often in close proximity to a methionine residue (see Figure 2), which is frequently oxidized during sample processing. Correspondingly, the site of a ΔA = 32 Da mass shift, which can correspond to double-oxidation event or two singly oxidized alanine residues, was also often found near methionine residues. Therefore, the source of the most frequent and abundant SAAPs could perhaps be explained away as a "shadow" of the most common sampling processing artifact.

Though the presence of a mass shift changes the ion fragmentation pattern of the corresponding ions, the fragmentation process is often incomplete. Some mass shifts will lead to unique fragmentation patterns, enabling a site to be unambiguously located. On the other hand, a mass shift that can occur at adjacent residue sites can introduce ambiguity and lead to incorrect localization; the candidate peptide variants will have similar theoretical fragmentation patterns and thus similar statistical scores. As the distance between the two sites increases, complementary site-determining b- and y-type ions together should increase a scoring algorithm's ability to mitigate the ambiguity. Therefore, we objectively evaluated how this ambiguity diminishes as the adjacency decreases.

Table 3. Top 20 Sequence Variants Observed for the *Populus* Data Set

variants identified	UniMod annotation	VNDL peptides	DENA peptides	VNDL summed spectra	DENA summed spectra
A[16]	Ala→Ser	978	1147	22439	22459
F[16]	oxidation; Phe→Tyr	345	387	7024	7424
M[−3]	Met→Lys; Met→Gln	247	261	5399	5351
N[−27]	Asn→Ser	201	229	3366	3437
V[14]	Val→Xle	153	169	2434	2998
S[14]	methyl; Ser→Thr	44	67	2353	798
A[32]	Ala→Cys	106	64	2206	2091
V[2]	Val→Thr	130	121	1872	1890
A[28]	Ala→Val	69	155	1811	1355
S[42]	acetyl; Ser→Glu	99	89	1424	1417
S[41]	Ser→Lys; Ser→Gln	82	81	1268	1344
G[30]	Gly→Ser	59	71	1217	1337
V[32]	Val→Met	74	106	1192	1227
S[−16]	deoxy; Ser→Ala	82	82	991	1480
N[15]	Asn→Glu	70	49	916	1129
L[−14]	Xle→Val	41	69	861	567
N[−13]	Asn→Thr	27	76	861	1192
G[14]	Gly→Ala	73	60	856	843
T[−30]	Thr→Ala	57	65	766	832
V[−28]	Val→Ala	68	27	741	1289

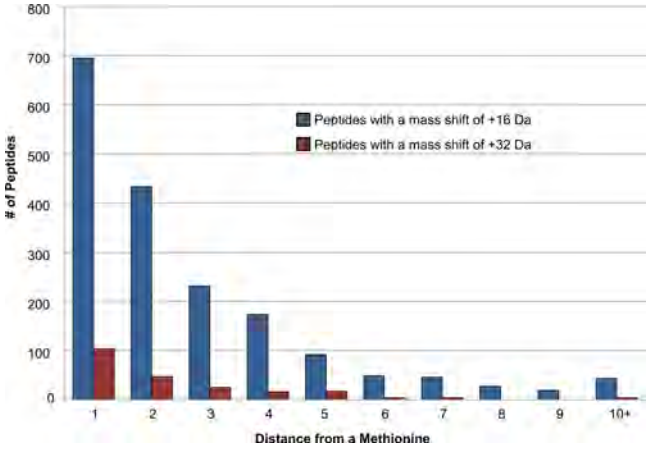


Figure 2. Proximity of +16 Da and +32 Da mass shifts to methionine residues. Detected peptides containing a methionine residue and either a +16 Da (blue) or +32 Da (red) mass shift on a non-methionine residue were plotted. This frequency distribution illustrates the degree of adjacency between the modification site and neighboring methionine residues.

The analysis was constrained to DENA leaf samples, which contained the highest frequency and abundance of ΔA = 16 Da mass shifts. Because high mass accuracy of fragment ions can help unambiguously annotate fragment ion peaks, an MS run using a "high−high" strategy, in which full scans (MS) and tandem mass spectra (MS/MS) are detected in the Orbitrap analyzer at high resolution and high mass accuracy, was simultaneously evaluated with an MS run that acquired MS/MS scans in the ion trap ("high−low"). The collected spectra were searched by MyriMatch using a directed method (see

dx.doi.org/10.1021/pr400192r | J. Proteome Res. 2013, 12, 3642−3651

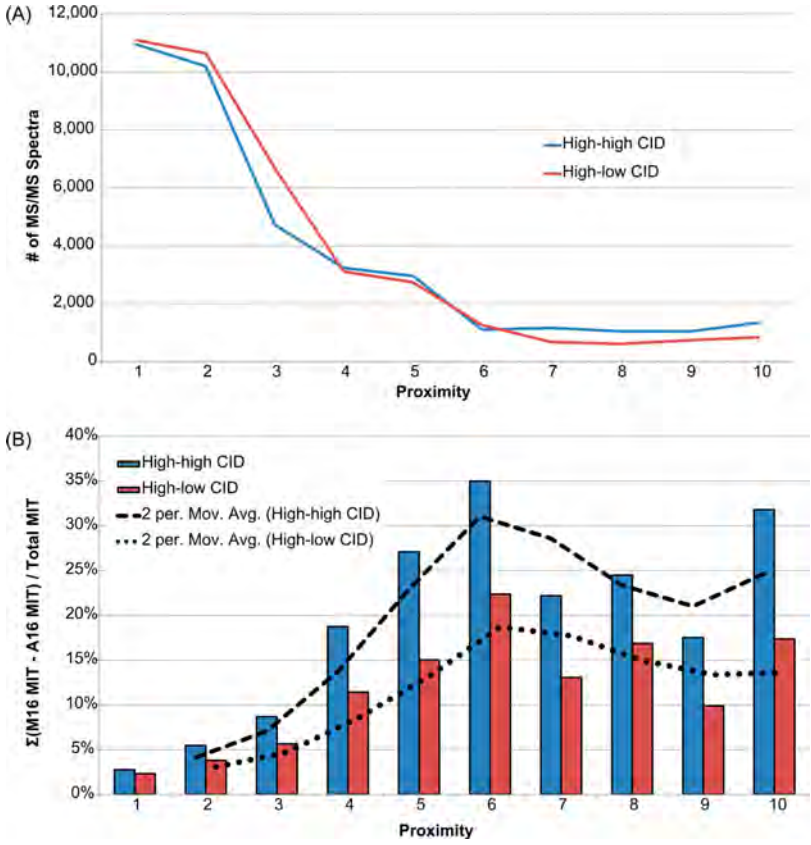


Figure 3. Identifying the level of ambiguity between adjacent mass shift sites. MS/MS spectra collected using a high−high (blue) and high−low strategy were interpreted by MyriMatch to identify all permissible +16 Da modifications on either alanine or methionine residues. Only contentious spectra (CS), MS/MS spectra that matched to the same peptide sequence but differed in the placement of the modification (i.e., at alanine or methionine), were plotted. (A) The frequency distribution of CS illustrates that level of ambiguity is strongly dependent on the distances between two potential modifications sites. (B) A matched ion intensity (MIT) was calculated for the two site positions, and the difference between the matched ion intensity (MIT) values was calculated for each CS as a function of the proximity. A moving average trendline was provided for both the high−high (dashed-line) and high−low (dotted-line) strategy to highlight the earliest maximal difference in the matched ion intensities.

Experimental Procedures); only a user-defined mass shift was considered. For both MS runs, two directed searches were performed: either a methionine (+16 Da) or an alanine (+16 Da) was allowed as a dynamic modification. By searching for the modifications independently, the search algorithm interpreted each spectrum, identified the mismatch region containing a permissible modification, and determined the most probable position of the mass shift on either the methionine or alanine. This approach enabled the identification of spectra that were annotated similarly, having the same underlying peptide sequence but differing by the location of the mass shift, either on a methionine or a neighboring alanine. For discussion purposes, these spectra will be referred as "contentious spectra" (CS). In total, the MS searches identified nearly the same number of CS for each analysis strategy −37 776 and 38 399 for high−high and high−low, respectively.

As anticipated, the number of CS declined as the distance between the methionine and alanine sites increased (Figure 3A). This observation is the result of an overall increase in the number of discriminatory b- and y-ions, which provides a more definitive spectral fingerprint. Also shown in this figure, the frequency of CS decreased at a similar rate for the two MS strategies. This was expected, as both strategies perform collision-induced dissociation (CID); the MS/MS spectra will contain the same percentage of backbone fragmentation. Interestingly, both MS strategies show a clear inflection point when the proximity was about six amino acid residues. We suspect that this point represents the distance that provides the most discrimination between the two types of mass shifts, (1) those belonging to a methionine sulfoxide and (2) those more likely due to a SAAP. For distances greater than six, the mass shift locations likely approach the terminal ends of the peptide sequence. In general, mass shifts located near the ends of a peptide sequence tend to be assigned less reliably than those near the center, which explains why a level of ambiguity remains. These observations are further corroborated by comparing the total matched ion intensity (MITs) of the b- and y-ion series for each peptide sequences that differed only by the location of a +16 Da mass shift. That is, for each ambiguous spectrum, we calculated the difference between the total MIT of the methionine (+16 Da) sequence and the total MIT of the alanine (+16 Da) sequence. Figure 3B shows the distribution of the percent difference between two potential sites for each distance. As shown, the maximum difference between the two theoretical mass shift sites occurred when the site locations were six amino acids apart. Although we suspected a high level of uncertainty for proximal sites, we demonstrated that the likelihood of precise site localization is severely diminished when the number of site-determining b- and y-ions fall below 12. Notably, the vast majority of the CS (68% high−high and 70% high−low) belong to peptides containing two potential possibilities that are less than four

dx.doi.org/10.1021/pr400192r | J. Proteome Res. 2013, 12, 3642−3651

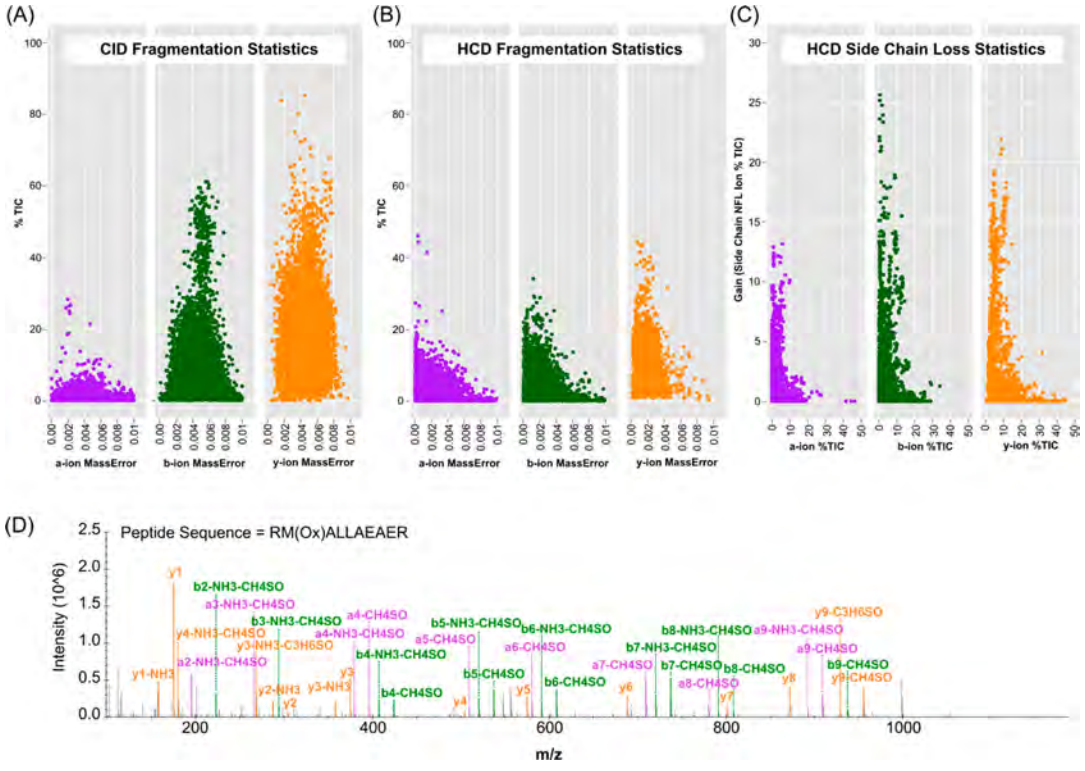


Figure 4. Fragmentations statistics of CID and HCD spectra. Only peptide spectrum matches (PSMs) meeting the following criteria were graphed: (1) PSMs identified by both CID and HCD strategies and (2) PSMs containing at least one methionine and a modified alanine (+16 Da) residue. a (purple), b (green), and y (yellow) series were plotted. For each CID spectrum (A) and HCD spectrum (B), the percentage of the total ion current (TIC) attributable to a particular fragment ion series was plotted. (C) If a spectrum contained peaks which could unambiguously assigned as neutral losses from methionine sulfoxide, the additional intensity coverage for ambiguous spectra was calculated. (D) As an example, the HCD spectrum with the maximum additional coverage achievable (31%) was provided. Here, only the top 20 most abundant fragment ions were highlighted.

amino acids apart. Cleary, these spectra have little or no site-determining information for proper site placement, which would be necessary for confident SAAP identification.

As others have shown, these observations highlight how precise site localization can be challenging for search algorithms when there are few site-determining fragment ions.[27] Presently, additional software is available to calculate the probability of correct localization for each site.[28−31] Though calculating a probability-based score provides a measure of certainty, spectra with insufficient site-determining ions (i.e., peptides with proximal residue sites and spectra featuring incomplete fragmentation) remain logistical problems. In other words, precise site localization in CID fragmentation spectra can be difficult when the distance between the two likely sites is less than six amino acids apart. Nevertheless, an alternate approach, as outlined below, is available to provide additional information for discriminating between SAAPs and what we suspect is the most common chemical modification mistaken for SAAPs: methionine oxidations.

3.3.1. Identification of Methionine Sulfoxide Sites Using High Energy Dissociation (HCD). For peptide-sequence tagging, we employed collision-induced dissociation (CID), which is by far the most frequently used technique in proteomics for peptide sequencing. When CID fragmentation techniques are applied, the widely accepted model that describes the dissociation process designates b- and y-ion series as the most prevalent types.[32,33] The primary fragment ions and their contribution to the overall intensity coverage for a single CID run are illustrated in Figure 4A. In principle, complete coverage of the entire b- and y-ion series ions allows full annotation of the amino acid sequence of a peptide. As detailed in the section above, this information may be insufficient for definitively localizing mass shifts. However, there are alternative fragmentation processes that could benefit this task.

Introduced in 2007, higher energy collisional dissociation (HCD) fragmentation became available on the Orbitrap platforms.[34] In a dedicated collisional cell, peptide ions are subjected to a beam-type fragmentation process, where primary fragment ions retain kinetic energy and are therefore more likely to fragment again. In general, HCD ion types are expected to follow the fragmentation rules modeled from CID. Therefore, regular ions (b- and y-type ions) derived from backbone fragmentation are expected to be among the most abundant types observed. Besides a slightly lower contribution of the b- and y-ion series to the total TIC collected in each scan, the observed primary fragment ions and their overall intensities in a HCD run are comparable to CID (Figure 4B). A prominent difference, however, is larger contribution of the a-type ion series, which are derived from b-ions by losing CO. Moreover, as a direct consequence of the beam-type fragmentation process, the primary fragment ions are subjected to additional fragmentation pathways and consequently give rise to various ion types beyond those typically observed in CID.[35] A large portion of such ions are those involving neutral losses; the loss of water and ammonia are by far the most frequently observed. Another frequently observed class is the neutral loss of an amino acid side chain. In fact, the side chain of methionine sulfoxide is prone to cleavage[36], producing ions with a specific neutral fragment loss (NFL). Because search

dx.doi.org/10.1021/pr400192r | J. Proteome Res. 2013, 12, 3642−3651

algorithms only consider backbone fragmentation (i.e., a-, b-, and y-ions) and some of their neutral losses (NH_3 and H_2O), a large percentage of the content in HCD spectra remain unassigned. Though many of these peaks belong to internal fragment ions and immonium ions, there are peaks which can be unambiguously assigned as neutral losses from methionine sulfoxide, based on the knowledge of how they fragment and the calculation of their fragment masses. Therefore, we exploited HCD fragmentation to identify the presence and precise location of methionine oxidations.

Again, the analysis was constrained to DENA leaf samples, and measurements were collected by the LTQ Orbitrap Pro mass spectrometer, which features improved sensitivity and HCD capability compared to its predecessors. HCD fragmentation was performed in the dedicated octopole collisional cell, and fragment ions were detected in the Orbitrap. To test the suitability of this approach, the collected spectra were searched by MyriMatch using a directed method: alanine (+16 Da) was considered as the only dynamic modification. With this approach, the search algorithm considers the location of the mass shift irrespective of neighboring methionine sites. Methionine was intentionally neglected during the peptide-spectrum matching process to eliminate the MyriMatch scoring system from the discrimination process. HCD spectra that matched a peptide sequence containing a modified alanine (+16 Da) and at least one methionine were further interpreted. This step restricted the analysis to 4943 spectra, which matched to 1175 peptides. When annotating HCD peptide-spectrum matches, we looked for the presence of the characteristic neutral loss ions from the primary fragment ions (a, b, and y) of a peptide containing methionine sulfoxide (Figure 4D). As mentioned previously, the loss of water and ammonia from primary fragment ions are frequently observed. Therefore, these additional small molecule losses were taken into consideration when applicable.

For each spectrum, we calculated the percent gain in matched ion intensity when considering peaks attributable to the cleavage of a methionine sulfoxide side chain. Figure 4C depicts their overall contribution for each ion series: 96% of the spectra and 81% of the peptides exhibit at least one neutral loss from a methionine sulfoxide residue. With only a slight increase in the relative abundance of b-ions, the trends observed for each ion series (Figure 4C) agree with their expected contribution in a typical HCD run (Figure 4B). The most prominent fragmentation process observed was the neutral loss of methane sulfenic acid (CH_4SO). This chemical species exhibited a higher percentage of side-chain cleavage relative to the frequencies of the other fragment ions and could be observed in 83% of all MS/MS spectra exhibiting side-chain loss. Despite only occurring when a fragment ion contains a methionine sulfoxide residue, i.e., CH_4SO, C_3H_6SO and C_3H_8SO, the three species could be found relatively abundant in the spectra, 3%, 1%, and 1%, respectively. While their mean contribution to the overall intensity coverage was 5%, the maximum additional coverage achievable was 31% (Figure 4D). The gain in spectral information is promising: if searching algorithms could consider these characteristic permutations during the identification process, the false localization rate of oxidation events would be minimized. It should be noted that the HCD fragmentation process is beneficial not only for the localization of methionine oxidations but also for other modification events that have characteristic neutral losses, such as phosphorylations.

4. CONCLUSION

Here we implemented automated sequence tag inferences (DirecTag) and reconciliation (TagRecon) for the identification of unanticipated sequence variants at a global level. Together, careful search space selection and the availability of high mass-accuracy data maximized the sensitivity of the experiment and improved the overall integrity of the data set. The large-scale study yielded a broad and quantitative view of the frequencies and abundances of various single amino acid polymorphisms in *Populus*. Despite the limited sequence coverage afforded in a typical shotgun approach, peptide sequence variants were nearly observed in every protein. Overall, we were able to generate a data set containing 6653 attested peptide sequence variants.

Though we have shown the potential of peptide sequencing tagging in *Populus*, a high-throughput and automated assignment of mass shifts to the correct amino acid remains a challenge for these large-scale studies. A widely acknowledged problem, precise site localization becomes difficult when multiple residues within a single peptide can be modified. When the distance between two potential sites decreases, the theoretical fragmentation peaks of the two candidate annotations become more similar. As a result, there are fewer site-determining ions available to uniquely assign a mass shift to a specific residue. Especially because the CID fragmentation process is often incomplete, the identification of a full series of the b- or y-ion type is rarely achieved. Although search algorithms report the highest scoring modified peptide, insufficient site-determining ions may lead to the incorrect localization of mass shifts. These shortcomings were clearly apparent in our study, as the most abundant chemical modification could masquerade a sequence variant (Ala→Ser). Currently, there are more sensitive approaches available, such as the ASCORE method, which calculate a probability based for specific site locations. Although these scores can generally discriminate alternative sites, a smaller spacing of the two sites within a peptide sequences can lower the performance of the scores. Within our data set, the spacing of alternative sites within a peptide greatly influenced localization of the mass shift: the maximum discriminating evidence did not occur until two alternative oxidation sites were six amino acids apart. More importantly, peptides with potential sites less than four amino acids apart (~70% of all the CS) had insufficient evidence for confident site placement.

Owing to the frequency of methionine oxidations, we exploited HCD fragmentation to assess their location objectively without the need for using probability-based scores. Because the HCD fragmentation behavior of methionine sulfoxide-containing peptides can be quite distinct, we were able to empirically collect information that facilitated the localization of ambiguous +16 Da mass shifts. In contrast to CID spectra, in which only the regular ion series (a, b, and y) are available, HCD spectra contain characteristic neutral fragment loss ions which enabled the explicit identification of methionine sulfoxide residues. If search algorithms could make use of the available additional spectral information, we suspect that HCD will enable improved site placement for *de novo* sequencing and hybrid peptide sequencing tagging-based approaches.

Detecting the molecular signatures at the gene, transcript, and protein level is necessary to make reliable phenotype and genotype associations. Because reference protein databases are

dx.doi.org/10.1021/pr400192r | J. Proteome Res. 2013, 12, 3642−3651

incomplete with respect to protein sequence variation information, proteomic investigations fail to identify any protein form containing a single amino acid polymorphism (SAAP). While the peptide sequence tagging approach described in this study has the potential to expand the biological information we can obtain from a proteomics experiment, particularly with respect to single amino acid polymorphisms, further improvements are needed in the ability to accurately determine site-specific variation. When confident site placement of SAAPs is achieved, this will provide a powerful approach to interrogate allelic frequencies at the protein level and, perhaps, quantitate the relative abundance of both protein isoforms. This approach, for instance, could be used to detect the relative excess or deficit of certain allelic frequencies across a natural population, providing information that may lead to the discovery of a novel protein forms that are responsible for a particular phenotype. Therefore, we argue that, even though this approach currently does not provide unambiguous localization of all mass shifts that correspond to SAAPS, in general this approach provides a level of information that is not made available through genomic or transcriptomic techniques.

■ ASSOCIATED CONTENT

Supporting Information

Additional information as noted in the text. This material is available free of charge via the Internet at http://pubs.acs.org

■ AUTHOR INFORMATION

Corresponding Author

*Phone: 865-574-4986. Fax: 865-576-8559. E-mail: hettichrl@ornl.gov.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This study was conducted as part of the BioEnergy Science Center through funding from the U.S. Department of Energy, Office of Biological and Environmental Research, Genome Sciences Program. P.A. and R.A. would like to acknowledge financial support from the Genome Science and Technology Graduate Program at the University of Tennessee. Oak Ridge National Laboratory is managed for the U.S. Department of Energy by the University of Tennessee − Battelle, L.L.C.

■ REFERENCES

(1) Tuskan, G. A.; DiFazio, S.; Jansson, S.; Bohlmann, J.; Grigoriev, I.; Hellsten, U.; Putnam, N.; Ralph, S.; Rombauts, S.; Salamov, A.; Schein, J.; Sterck, L.; Aerts, A.; Bhalerao, R. R.; Bhalerao, R. P.; Blaudez, D.; Boerjan, W.; Brun, A.; Brunner, A.; Busov, V.; Campbell, M.; Carlson, J.; Chalot, M.; Chapman, J.; Chen, G. L.; Cooper, D.; Coutinho, P. M.; Couturier, J.; Covert, S.; Cronk, Q.; Cunningham, R.; Davis, J.; Degroeve, S.; Dejardin, A.; Depamphilis, C.; Detter, J.; Dirks, B.; Dubchak, I.; Duplessis, S.; Ehlting, J.; Ellis, B.; Gendler, K.; Goodstein, D.; Gribskov, M.; Grimwood, J.; Groover, A.; Gunter, L.; Hamberger, B.; Heinze, B.; Helariutta, Y.; Henrissat, B.; Holligan, D.; Holt, R.; Huang, W.; Islam-Faridi, N.; Jones, S.; Jones-Rhoades, M.; Jorgensen, R.; Joshi, C.; Kangasjarvi, J.; Karlsson, J.; Kelleher, C.; Kirkpatrick, R.; Kirst, M.; Kohler, A.; Kalluri, U.; Larimer, F.; Leebens-Mack, J.; Leple, J. C.; Locascio, P.; Lou, Y.; Lucas, S.; Martin, F.; Montanini, B.; Napoli, C.; Nelson, D. R.; Nelson, C.; Nieminen, K.; Nilsson, O.; Pereda, V.; Peter, G.; Philippe, R.; Pilate, G.; Poliakov, A.; Razumovskaya, J.; Richardson, P.; Rinaldi, C.; Ritland, K.; Rouze, P.; Ryaboy, D.; Schmutz, J.; Schrader, J.; Segerman, B.; Shin, H.; Siddiqui, A.; Sterky, F.; Terry, A.; Tsai, C. J.; Uberbacher, E.; Unneberg, P.; Vahala, J.; Wall, K.; Wessler, S.; Yang, G.; Yin, T.; Douglas, C.; Marra, M.; Sandberg, G.; Van de Peer, Y.; Rokhsar, D. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **2006**, *313* (5793), 1596−1604.

(2) Wullschleger, S. D.; Tuskan, G. A.; DiFazio, S. P. Genomics and the tree physiologist. *Tree Physiol.* **2002**, *22* (18), 1273−1276.

(3) Wullschleger, S. D.; Weston, D. J.; Difazio, S. P.; Tuskan, G. A. Revisiting the sequencing of the first tree genome: Populus trichocarpa. *Tree Physiol.* **2013**, *33* (4), 357−364.

(4) Slavov, G. T.; DiFazio, S. P.; Martin, J.; Schackwitz, W.; Muchero, W.; Rodgers-Melnick, E.; Lipphardt, M. F.; Pennacchio, C. P.; Hellsten, U.; Pennacchio, L. A.; Gunter, L. E.; Ranjan, P.; Vining, K.; Pomraning, K. R.; Wilhelm, L. J.; Pellegrini, M.; Mockler, T. C.; Freitag, M.; Geraldes, A.; El-Kassaby, Y. A.; Mansfield, S. D.; Cronk, Q. C.; Douglas, C. J.; Strauss, S. H.; Rokhsar, D.; Tuskan, G. A. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree Populus trichocarpa. *New Phytol.* **2012**, *196* (3), 713−725.

(5) Bunger, M. K.; Cargile, B. J.; Sevinsky, J. R.; Deyanova, E.; Yates, N. A.; Hendrickson, R. C.; Stephenson, J. L., Jr. Detection and validation of non-synonymous coding SNPs from orthogonal analysis of shotgun proteomics data. *J. Proteome Res.* **2007**, *6* (6), 2331−2340.

(6) Abraham, P.; Adams, R.; Giannone, R. J.; Kalluri, U.; Ranjan, P.; Erickson, B.; Shah, M.; Tuskan, G. A.; Hettich, R. L. Defining the boundaries and characterizing the landscape of functional genome expression in vascular tissues of Populus using shotgun proteomics. *J. Proteome Res.* **2012**, *11* (1), 449−460.

(7) Lu, B.; Xu, T.; Park, S. K.; McClatchy, D. B.; Liao, L.; Yates, J. R., III. Shotgun protein identification and quantification by mass spectrometry in neuroproteomics. *Methods Mol. Biol.* **2009**, *566*, 229−259.

(8) Frank, A.; Pevzner, P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **2005**, *77* (4), 964−973.

(9) Searle, B. C.; Dasari, S.; Wilmarth, P. A.; Turner, M.; Reddy, A. P.; David, L. L.; Nagalla, S. R. Identification of protein modifications using MS/MS de novo sequencing and the OpenSea alignment algorithm. *J. Proteome Res.* **2005**, *4* (2), 546−554.

(10) Tsur, D.; Tanner, S.; Zandi, E.; Bafna, V.; Pevzner, P. A. Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* **2005**, *23* (12), 1562−1567.

(11) Kapp, E.; Schutz, F. Overview of tandem mass spectrometry (MS/MS) database search algorithms. In *Current Protocols in Protein Science*; Wiley: New York, 2007; Unit 25.2.

(12) Sunyaev, S.; Liska, A. J.; Golod, A.; Shevchenko, A. MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal. Chem.* **2003**, *75* (6), 1307−1315.

(13) Tabb, D. L.; Saraf, A.; Yates, J. R., III. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem.* **2003**, *75* (23), 6415−6421.

(14) 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.

(15) Shilov, I. V.; Seymour, S. L.; Patel, A. A.; Loboda, A.; Tang, W. H.; Keating, S. P.; Hunter, C. L.; Nuwaysir, L. M.; Schaeffer, D. A. The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol. Cell. Proteomics* **2007**, *6* (9), 1638−1655.

(16) Tabb, D. L.; Ma, Z. Q.; Martin, D. B.; Ham, A. J.; Chambers, M. C. DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *J. Proteome Res.* **2008**, *7* (9), 3838−3846.

(17) Dasari, S.; Chambers, M. C.; Slebos, R. J.; Zimmerman, L. J.; Ham, A. J.; Tabb, D. L. TagRecon: high-throughput mutation identification through sequence tagging. *J. Proteome Res.* **2010**, *9* (4), 1716−1726.

(18) Mann, M.; Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **1994**, *66* (24), 4390−4399.
(19) 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.
(20) Wilmarth, P. A.; Tanner, S.; Dasari, S.; Nagalla, S. R.; Riviere, M. A.; Bafna, V.; Pevzner, P. A.; David, L. L. Age-related changes in human crystallins determined from comparative analysis of post-translational modifications in young and aged lens: does deamidation contribute to Crystallin insolubility? *J. Proteome Res.* **2006**, *5* (10), 2554−2566.
(21) Tabb, D. L.; Fernando, C. G.; Chambers, M. C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* **2007**, *6* (2), 654−661.
(22) Abraham, P.; Giannone, R. J.; Adams, R. M.; Kalluri, U.; Tuskan, G. A.; Hettich, R. L. Putting the pieces together: high-performance LC-MS/MS provides network-, pathway-, and protein-level perspectives in Populus. *Mol. Cell. Proteomics* **2013**, *12* (1), 106−119.
(23) Nesvizhskii, A. I.; Aebersold, R. Interpretation of shotgun proteomic data - The protein inference problem. *Mol. Cell. Proteomics* **2005**, *4* (10), 1419−1440.
(24) MacCoss, M. J.; McDonald, W. H.; Saraf, A.; Sadygov, R.; Clark, J. M.; Tasto, J. J.; Gould, K. L.; Wolters, D.; Washburn, M.; Weiss, A.; Clark, J. I.; Yates, J. R., 3rd Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (12), 7900−7905.
(25) Choudhary, G.; Wu, S. L.; Shieh, P.; Hancock, W. S. Multiple enzymatic digestion for enhanced sequence coverage of proteins in complex proteomic mixtures using capillary LC with ion trap MS/MS. *J. Proteome Res.* **2003**, *2* (1), 59−67.
(26) Aguiar, M.; Haas, W.; Beausoleil, S. A.; Rush, J.; Gygi, S. P. Gas-phase rearrangements do not affect site localization reliability in phosphoproteomics data sets. *J. Proteome Res.* **2010**, *9* (6), 3103−3107.
(27) Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **2006**, *24* (10), 1285−1292.
(28) Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **2006**, *127* (3), 635−648.
(29) Bailey, C. M.; Sweet, S. M.; Cunningham, D. L.; Zeller, M.; Heath, J. K.; Cooper, H. J. SLoMo: automated site localization of modifications from ETD/ECD mass spectra. *J. Proteome Res.* **2009**, *8* (4), 1965−1971.
(30) Chen, Y.; Chen, W.; Cobb, M. H.; Zhao, Y. PTMap–a sequence alignment software for unrestricted, accurate, and full-spectrum identification of post-translational modification sites. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106* (3), 761−766.
(31) Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom.* **2000**, *35* (12), 1399−1406.
(32) Boyd, R.; Somogyi, A. The mobile proton hypothesis in fragmentation of protonated peptides: a perspective. *J. Am. Soc. Mass Spectrom.* **2010**, *21* (8), 1275−1278.
(33) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **2007**, *4* (9), 709−712.
(34) Michalski, A.; Neuhauser, N.; Cox, J.; Mann, M. A systematic investigation into the nature of tryptic HCD spectra. *J. Proteome Res.* **2012**, *11* (11), 5479−5491.
(35) Reid, G. E.; Roberts, K. D.; Kapp, E. A.; Simpson, R. I. Statistical and mechanistic approaches to understanding the gas-phase fragmentation behavior of methionine sulfoxide containing peptides. *J. Proteome Res.* **2004**, *3* (4), 751−759.