Journal of proteome research

Article

pubs.acs.org/jpr

Integrating mRNA and Protein Sequencing Enables the Detection and Quantitative Profiling of Natural Protein Sequence Variants of *Populus trichocarpa*

Paul E. Abraham,[†] Xiaojing Wang,[§] Priya Ranjan,[†] Intawat Nookaew,[‡] Bing Zhang,[§] Gerald A. Tuskan,[‡] and Robert L. Hettich*[,†]

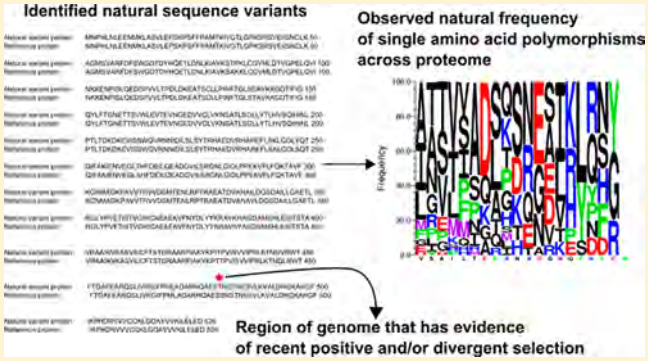
[†]Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States

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

[§]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, United States

Supporting Information

ABSTRACT: Next-generation sequencing has transformed the ability to link genotypes to phenotypes and facilitates the dissection of genetic contribution to complex traits. However, it is challenging to link genetic variants with the perturbed functional effects on proteins encoded by such genes. Here we show how RNA sequencing can be exploited to construct genotype-specific protein sequence databases to assess natural variation in proteins, providing information about the molecular toolbox driving cellular processes. For this study, we used two natural genotypes selected from a recent genome-wide association study of *Populus trichocarpa*, an obligate outcrosser with tremendous phenotypic variation across the natural population. This strategy allowed us to comprehensively catalogue proteins containing single amino acid polymorphisms (SAAPs), as well as insertions and deletions. We profiled the frequency of 128 types of naturally occurring amino acid substitutions, including both expected (neutral) and unexpected (non-neutral) SAAPs, with a subset occurring in regions of the genome having strong polymorphism patterns consistent with recent positive and/or divergent selection. By zeroing in on the molecular signatures of these important regions that might have previously been uncharacterized, we now provide a high-resolution molecular inventory that should improve accessibility and subsequent identification of natural protein variants in future genotype-to-phenotype studies.



KEYWORDS: *natural variants, mass spectrometry, shotgun proteomics, RNA sequencing, genotype-specific protein database*

INTRODUCTION

With the widespread availability of next-generation sequencing technologies (e.g., Illumina and Moleculo), researchers now have the tools to rapidly scan millions of single-nucleotide variants (SNVs) across thousands of genomes to find genetic variants associated with a particular phenotype. Over the past decade, such genome-wide association (GWA) studies have become increasingly popular in human and plant genetics.[1−3] Studies assessing natural variation across humans are gradually unraveling the genetic mechanisms underlying human traits, diseases, and behaviors.[4,5] This is also clearly an exciting time for plant research, where GWA studies are uncovering the genetic basis of various agronomical traits and thereby providing a means to increase crop quality and yield.[6,7]

With advances in sequencing output and computational data analysis, it has become much easier for researchers to exploit GWA data sets (i.e., SNV genotyping and trait measurements) to rapidly discover millions of SNVs within a genome.[8] However, the current challenge today is not discovering these genetic variations but rather understanding how the identified SNVs affect protein function and, even more challenging, how theses altered proteins collectively affect phenotype. Therefore, missing from many studies are the investigations of how variation in local genome structure relates to the cellular phenotypes.

With the availability of RNA sequencing (RNA-seq), whole-transcriptome studies are now capable of creating rich and abundant gene expression libraries, along with information on splice variants across a dynamic range of quantification.[9] Although identifying the transcribed portion of the genome reveals which genes have been expressed, there are many post-transcriptional regulatory mechanisms occurring after mRNAs are manufactured that influence production and maintenance of their subsequent protein products. As such, these types of

Received: September 3, 2015
Published: October 20, 2015

DOI: 10.1021/acs.jproteome.5b00823
J. Proteome Res. 2015, 14, 5318−5326

investigations would benefit tremendously by pairing transcriptome with proteome analyses.

At present, the evolving field of proteomics continues to provide accurate insights on the diversity and complexity of phenotypes.[10] Among the different methods used for proteomic studies, shotgun proteomics has become a vital tool for the high-throughput characterization of proteins in biological samples.[11−15] Although shotgun proteomics enables exquisite high-resolution peptide sequencing, fundamentally, this interrogation relies heavily on the availability of a high-quality reference genome and thereby is biased toward well-curated, predicted gene annotations.

The main impediment to the detection of sequence variants using mass spectrometry has been the lack of proteomic databases that include sample-specific variant sequences; consequently, such investigations are often taxonomically restricted to model organisms. This limitation prevents mass spectrometry-based proteomic approaches from providing the necessary information for GWA studies. To circumvent this problem, researchers have relied on *de novo* search algorithms to infer full-length peptide sequences from tandem mass spectra without requiring a protein sequence reference database.[16−18] Although this strategy has seen moderate success,[19,20] there are several critical challenges that cannot be simply addressed; e.g., these approaches create large databases that increase the false positive rate and thereby decrease peptide identification sensitivity. With the availability of RNA-seq data, more complete, sample-specific protein databases are becoming available. Today, RNA-seq transcriptomes can be obtained using next-generation sequencers in a rather straightforward, low cost manner. The integration of RNA-seq to establish nucleotides sequences that are directly translated into protein sequences and thus identifiable in proteomic measurements appears promising.[21−25]

The scientific community has heavily invested in generating -omic resources for *Populus*, given its scientific and economic importance. Currently, *Populus* genomics resources are among the most mature of any plant species and include (1) the first reference genome of a perennial plant[26] and (2) a repository of more than 48 million high-quality SNVs[27] and a whole-genome resequencing initiative of 1084 genomes.[28] Therefore, by characterizing the diversity at the protein level across many genotypes, there is a potential to identify underlying mechanisms underpinning complex phenotypes, which could have noteworthy implications in the bioenergy and forest products industries.

Here, we use two individual genotypes identified in a recent GWA study[28] in the genus *Populus* to demonstrate how next-generation RNA-seq data can be leveraged to generate the first genotype-specific protein databases to identify SNVs and short insertion and deletions (INDELs) in *Populus*. By exploiting available genome resources, we also characterized the diploid landscape of the *Populus* genome at the protein level and can therefore now provide novel insights into this inherent variation. Given this comprehensive analysis strategy, we generated a web-based resource to help visualize the results at the genome, transcriptome, and proteome levels.

■ MATERIALS AND METHODS

Plant Material

Two *Populus trichocarpa* genotypes, DENA-17-3 and VNDL-27-4, were grown under standard greenhouse conditions as previously outlined.[29] From these trees, young leaf including the petiole and midrib (LPI 4−6), 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 trees per genotype and pooled together for each sample type to reduce the effects of sample-to-sample variation.

Protein Extraction and Digestion

As previously described,[19] leaf, root, and stem tissues were ground under liquid nitrogen using a mortar and pestle. For each organ-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 (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, 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. Pelleted proteins were resuspended in 250 μL of 8 M urea, 100 mM Tris-HCl, pH 8.0 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 iodoacetamide (20 mM) to prevent reformation of disulfide bonds. Samples were digested via two aliquots of sequencing-grade trypsin (Promega, 1:75 [w:w]) at two different sample dilutions, 4 M urea (overnight) and subsequent 2 M urea (5 h). Following digestion, samples were adjusted to 200 mM NaCl, 0.1% formic acid (FA) and filtered through a 10 kDa cutoff spin column filter (Vivaspin 2, GE Health) to remove under-digested proteins. The peptide-enriched flow through was then quantified by BCA assay, aliquoted, and stored at −80 °C.

LC−MS/MS

For the analysis of the proteome samples, 25 μg of each peptide mixture was 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 i.d.) packed with 15 cm of C18 RP material and analyzed via 22-h MudPIT 2D LC−MS/MS. The three solvent solutions used for chromatography were 5% ACN/0.1% FA (solvent A), 70% ACN/0.1% FA (solvent B), and 500 mM ammonium acetate/5% ACN/0.1% FA (solvent C). Ten steps were 120 min each with the following profile: 5 min of 100% solvent A, 2 min of x% buffer C, 3 min of 100% solvent A, a 10 min gradient from 0 to 10% solvent B, a 75 min gradient from 10 to 35% solvent B, and a 25 min gradient from 35 to 50% solvent B. The last step was 155 min with following profile: 5 min of 100% solvent A, 5 min of x% buffer C, 5 min of 100% solvent A, a 10 min gradient from 0 to 10% solvent B, a 75 min gradient from 10 to 35% solvent B, a 25 min gradient from 35 to 50% solvent B, a 15 min gradient from 50 to 100% solvent B, and 15 min of solvent A. Solvent C percentages (x) in steps 1−11 were as follows: 5, 7, 10, 12, 15, 17, 20, 25, 35, 50, and 100%. Peptide sequencing analysis was performed with an LTQ-Orbitrap-Velos-Pro mass spectrometer (ThermoScientific). For each sample, three technical replicates were

DOI: 10.1021/acs.jproteome.5b00823
J. Proteome Res. 2015, 14, 5318−5326

Table 1. Summary of SNV Positions

genotype	total	3′UTR	5′UTR	coding	intergenic	intron
DENA-17-3	67 050	16 478 (24.6%)	6425 (9.6%)	39 863 (59.5%)	2185 (3.3%)	2099 (3.1%)
VNDL-27-4	63 360	14 875 (23.5%)	5843 (9.2%)	38 283 (60.4%)	2159 (3.4%)	2200 (3.5%)

Table 2. Summary of INDEL Positions

genotype	total	3′UTR	5′UTR	coding	intergenic	intron
DENA-17-3	8475	4443 (52.4%)	1931 (22.8%)	981(11.6%)	663 (7.8%)	457 (5.4%)
VNDL-27-4	4122	2303 (55.9%)	988 (24%)	210 (5.1%)	360 (8.7%)	264 (6.3%)

measured. 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% was used for CID. 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 (15000 at m/z 400) MS1 spectra were acquired in the Orbitrap analyzer (XCalibur version 2.1).

Peptide and Protein Identification

All experimental MS/MS spectra were compared to theoretical tryptic peptide sequences generated from a genotype-specific FASTA database containing the full protein complement of the RNA-Seq-derived protein sequences for DENA-17-3 (63 241 proteins) and VNDL-27-4 (56 841 proteins) and common contaminant proteins (i.e., porcine trypsin and human keratin). A decoy database, consisting of the reversed sequences of the target database, was appended in order to discern the false-discovery rate (FDR) at the spectral level. For standard database searching, the peptide fragmentation spectra (MS/MS) were searched with MyriMatch algorithm v2.1.[30] MyriMatch was configured to derive fully tryptic peptides with the following parameters: unlimited missed cleavages, parent mass tolerance of 10 ppm, a fragment mass tolerance of 0.5 m/z unit, a static modification on cysteine (carbamidomethylation; +57.0214 Da), an N-terminal dynamic modification of (carbamylation; +43.0058 Da), and a dynamic modification corresponding to an oxidation (+15.9949 Da) of methionine. IDPicker v3.1.573[31] software filtered the peptide identifications from MyriMatch, by combining different scoring information, to determine where to place the threshold to maximize identifications for a 2% peptide spectrum match-level FDR. For protein inference, database search results for each technical replicate search were merged together, and protein identifications with at least two distinct peptide identifications (i.e., amino acid sequence, charge state, and modifications are unique) were considered for further analysis. Overall, the average peptide-level FDR for each sample was <1%.

RESULTS AND DISCUSSION

Creating a Genotype-Specific Protein Reference Database Using RNA Sequencing Information

Although SNV information can be leveraged from the whole-genome resequencing of DENA-17-3 and VNDL-27-4 to identify predicted sequence variations at the protein level, the number of introduced variations makes it difficult to control the FDR in the search results for these expanded databases. We instead integrated mRNA level SNVs to identify variant protein sequences, which are directly upstream of protein production and closely relevant to proteomics data.[32] With the availability of a *Populus* RNA sequencing library containing over 0.5 million putative SNPs,[33] we were able to generate genotype-specific protein databases for the two *Populus trichocarpa* genotypes, DENA-17-3 and VNDL-27-4, using *customProDB*,[34] an R package that incorporates SNVs and short INDELs identified from RNA-Seq data into a protein database. In brief, TopHat2[35] was used to map the RNA-Seq reads to the *Populus* reference genome (v3.0; 41 335 genes and 73 013 transcripts). SAMtools were used to call variations and, since many variations are false positive, calls were filtered on the basis of the following criteria: (1) SNV quality above 20; (2) mapping quality above 25; (3) coverage above 6 reads; and (4) alternative base supported by a minimum of 3 reads. An extensive list of all SNV and INDEL positions can be found in Supplemental Tables 1 and 2 for DENA-27-3 and VNDL-27-4, respectively. In general, 60% of all SNVs are located in coding regions (Table 1). Table 2 lists the total number and the location of INDEL positions. In both cases, we employed a loose cutoff RPKM (reads per kilobase of transcript per million reads mapped; an estimation of gene expression) value of 1 to remove the transcripts with low expression levels. Identified nonsynonymous SNVs and INDELs which were located in the coding region were then used to create genotype-specific protein sequence databases (Table 3).

Table 3. Summary of RNA-seq Genotype-Specific Protein Databases

genotype	total	RPKM > 1	SNVs	INDEL
DENA-17-3	63 241	46 423	15 278	1540
VNDL-27-4	56 841	41 462	15 124	255

Identifying Protein Sequence Variants Using Tandem Mass Spectrometry

Using the genotype-specific protein databases generated in the previous section, we then employed the MyriMatch database search algorithm[30] to identify tandem mass spectra belonging to peptide sequence variants across three different tissue types (i.e., leaf, root, and stem) for two genotypes, DENA-17-3 and VNDL-27-4. While search algorithms assess the statistical significance of each peptide-spectrum match (PSM), we applied several additional criteria to ensure only high-confident, gene-specific variant peptides remain in the dataset. To improve confidence in peptide identifications, we implemented the following filtering guidelines: (1) the data was filtered to achieve an average spectrum FDR of <0.5%, (2) all sequence variant peptides matched at least 3 different spectra, (3) each distinct sequence variation was observed in only one gene, and (4) evidence for the SNVs unambiguously mapped to one nucleotide position. After applying these filtering criteria, we

DOI: 10.1021/acs.jproteome.5b00823
J. Proteome Res. 2015, 14, 5318−5326

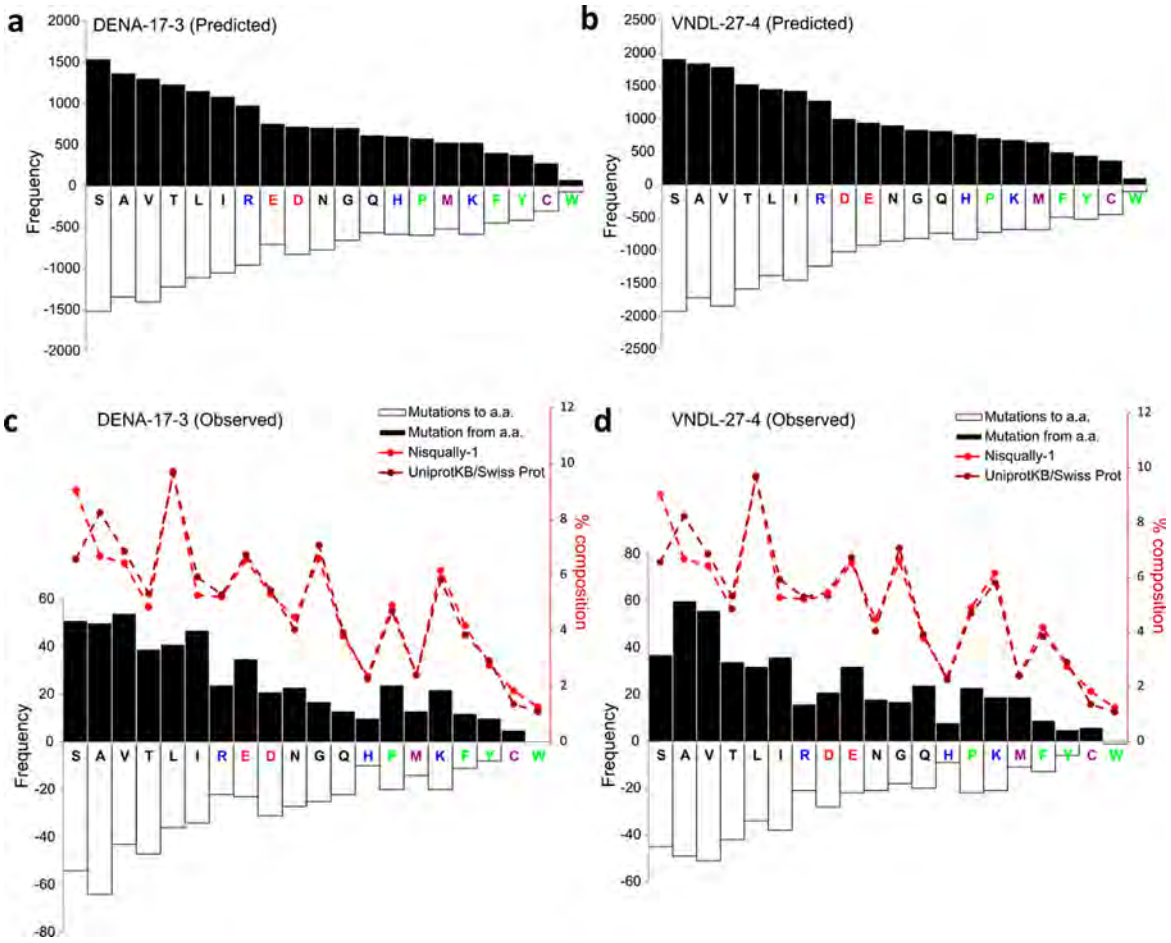


Figure 1. Frequency of a substitution from an amino acid (black) to an amino acid (white). The amino acids were sorted by frequency for the following distributions: (a) predicted frequency of substitutions in DENA-17-3, (b) predicted frequency of substitutions in VNDL-27−4, (c) identified frequency of substitutions in DENA-17-3, and (d) identified frequency of substitutions in VNDL-27-4. For panels c and d, the amino acid compositions from the *Populus* reference genome (Nisqually-1 v3.0) and the entire UniProtKB/Swiss-Prot protein knowledgebase have been provided. The residues are colored to represent the following amino acid classes: neutral and aliphatic (black), basic (blue), acid (red), sulfur-containing amino acids (purple), and other (green).

identified 44 848 nonredundant tryptic peptides (Supporting Information). In order to avoid any confusion arising from tissue-specific protein abundances, peptide abundance profiles were represented by summed spectral counts across tissue types. Peptide abundances that were confidently observed in DENA-17-3 and VNDL-27-4 were highly correlated (Pearson's correlation coefficient = 0.84), suggesting there are minor variations in protein abundances between the two genotypes.

To determine the expected level of observed variation at the protein level between each of the two subject genotypes, we used a kinship analysis to calculate relatedness.[8] The kinship matrix, which was based on 8.2 million SNVs with a minor allele frequency greater than 0.05, revealed that the two genotypes are slightly more similar to each other (kinship coefficient at ∼75%) than they are to the reference genome (Nisqually-1), yet both genotypes had approximately 74% similarity to Nisqually-1. Given the similarity of the two subject genotypes to the reference genome, we expected only a modest amount of gain in new peptide identifications using the RNA-Seq-derived protein sequences.

However, based on the above observation and the genotype-specific protein databases, we identified 626 and 547 new peptide sequences for DENA-17-3 and VNDL-27-4, respectively. A complete list of the variants and related information can be found in Supplemental Tables 3 and 4 for DENA-17-3 and VNDL-27-4, respectively. Although these gains are relatively low given the total number of peptides identified, the probability of identifying new peptides is expected to be further limited due to median sequence coverage (median protein sequence coverage of ∼25%[36]) typically achieved using a single protease, such as trypsin, to generate peptides. Furthermore, using the RNA-Seq-derived protein sequences approach is biased in the direction of differences between the subject genotypes and the reference; that is, we can only detect genes that are present in the reference genome, and therefore any novel DNA sequence in either of the two subject genotypes is hidden from comparative alignments. Nevertheless, this approach revealed peptide identifications that were previously unattainable using *Populus* reference genome as a proxy for these two genotypes. Overall, the variants peptides observed in DENA-17-3 and VNDL-27-4 can be attributed to 792 unique genome coding positions, with 201 positions shared in both genotypes. In total, there were 324 and 271 nucleotide variants only observed in DENA-17-3 and VNDL-27-4, respectively.

DOI: 10.1021/acs.jproteome.5b00823
J. Proteome Res. 2015, 14, 5318−5326

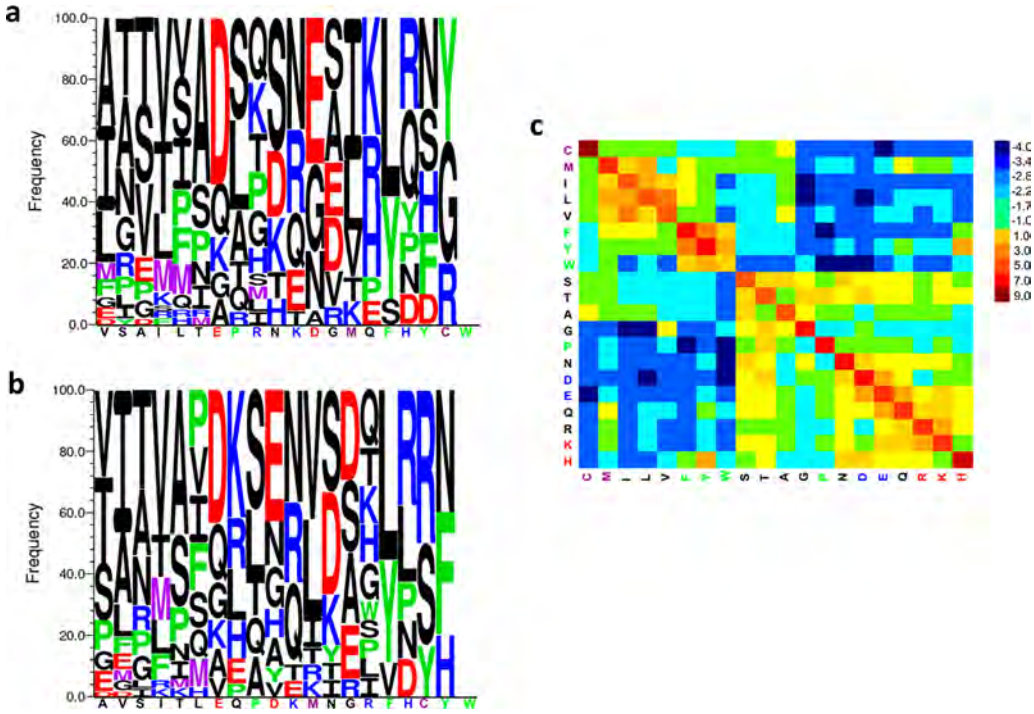


Figure 2. Identified occurrences of SAAPs. The amino acids on the *x*-axis were sorted by the identified frequencies shown in Figure 1. SAAPs in (a) DENA-17-3 and (b) VNDL-27-4 were graphed using Seq2Logo, using the logo type, PSSM-Logo. The residues are colored to represent the following amino acid classes: neutral and aliphatic (black), basic (blue), acid (red), sulfur-containing amino acids (purple), and other (green). (c) BLOSUM62 matrix, provided to illustrate the log probability of the substitution of one amino acid by another.

Characterization of the Identified Sequence Variants in *Populus*

The amino acid composition for each protein observed has been largely influenced by subtle selection pressures [the result of evolutionary forces acting upon random genetic mutations]. A subtle change in amino acid composition has the potential to alter the shape, activity, and/or function of a protein. Broadly, changes to protein amino acid composition may (1) have no impact a protein's functionality, (2) effectively improve protein functionality (e.g., improve thermostability or activity), (3) causatively change protein function, or (4) be deleterious, caused by a destabilization effect on a protein structure or function. Therefore, we characterized each amino acid substitutions for DENA-17-3 and VNDL-27-4. For DENA-17-3, the 525 observed SNVs consisted of 10 insertion and 3 deletion events and 512 nonsynonymous single nucleotide polymorphisms that caused a single amino acid polymorphism (SAAP). For VNDL-27-4, all 472 identified SNVs resulted in a SAAP.

To determine if there were differences in observed SNVs between the two genotypes, we estimated the substitution rate of each amino acid for both genotypes. As depicted in Figure 1a–d, the predicted and observed substitution rate for both genotypes are nearly identical, and the amino acids with the highest substitution rates are generally aliphatic to basic, and the lowest are acidic to aromatic. For both genotypes, we evaluated whether the observed subtraction rate per amino acid is equal to their replacement rate. On the basis of a chi-square test, we found the observed substitution per amino acid was not significantly different than the replacement rate (chi-square *P* values of 0.30 and 0.57 for DENA-17-3 and VNDL-27-4, respectively). Since mutations are generally a random process, the observed substitution frequencies are likely derived from the amino acid composition of the proteome. That is, given the identified distribution of substitutions for both genotypes, the probability of a particular amino acid substitution is likely weighted by its frequency in the genetic code; i.e., if the occurrence of an amino acid residue is high, the frequency of their substitution is expected to be high. Therefore, we related these substitution rates to the amino acid composition of the reference genome (Figure 1c,d), as well as the entire amino acid composition of the current UniProtKB/SwissProt protein knowledgebase (Figure 1c,d; http://web.expasy.org/docs/relnotes/relstat.html). Indeed, the most substituted amino acids are among the most frequently observed amino acids in nature. Since amino acid composition is well-conserved from species to species,[37] evolutionary pressures on the genetic code clearly established fitness with respect to amino acid composition in order to minimize the consequences of mutations to protein stability and/or function, explaining the observed frequencies of amino acid substitutions for these *Populus* genotypes.

In conjunction with amino acid composition, the genetic code ensures that amino acid substitutions due to mutations result in the interchange of similar amino acids, in so doing, buffering the impact on protein stability and/or function.[38] To test this hypothesis, we assessed the occurrence of the types of SAAPs observed for both *Populus* genotypes as well as their frequencies. We observed an immensely diverse set of SAAPs, identifying 128 different types of substitutions (Supplemental Table 5). We evaluated whether the frequency of the types of SAAPs was similar for both genotypes. On the basis of chi-square tests, and as depicted in Figure 2a,b, the frequencies of the types of SAAPs were similar for both genotypes. These frequencies, in conjunction with the Blosum62 matrix (Figure 2c), show the well-known observation that interchangeability is

DOI: 10.1021/acs.jproteome.5b00823
J. Proteome Res. 2015, 14, 5318–5326

an artifact of the relative mutability of an amino acid, where conservative changes to chemically and physically similar amino acids are likely to be nearly neutral and therefore are more likely to be accepted, thereby minimizing the phenotypic effect. For example, hydrophilic residues aspartic acid (D) and glutamic acid (E) have a common second base-pair in their codons and are therefore observed to frequently substitute one another.

To interpret the functional impact of each SAAP and INDEL on its protein, we employed PROVEAN (Protein Variation Effect Analyzer[39]), a software tool which predicts whether an amino acid substitution has an impact on the biological function of a protein. Figure 3 shows the distribution of

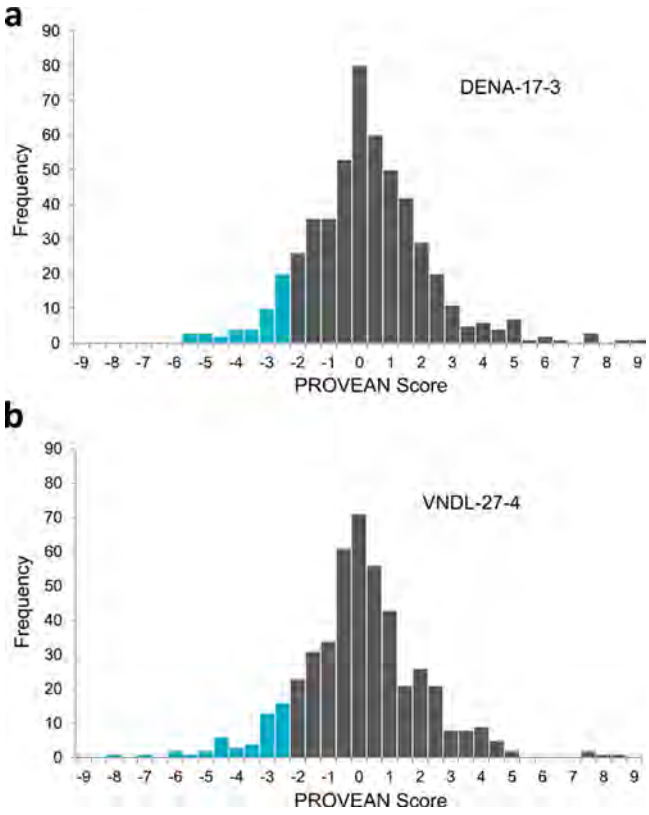


Figure 3. Identification of SAAPs variants that are predicted to be functionally important. Distribution of PROVEAN scores for (a) DENA-17-3 and (b) VNDL-27-4 illustrates the frequency of SAAPs predicted to have a non-neutral (blue; PROVEAN score < -2.5) or neutral effect (black; PROVEAN score ≥ -2.5).

PROVEAN scores for each SAAP for each genotype. A predefined PROVEAN score threshold of −2.5 was selected to allow for the best balanced discrimination between non-neutral and neutral classes.[39] With this threshold, 10% of all SNVs were predicted to have a non-neutral effect. A complete list of the PROVEAN scores and related information can be found in Supplemental Tables 3 and 4 for DENA-17-3 and VNDL-27-4, respectively. Overall, 46 SAAPs and 1 deletion were predicted to have a non-neutral effect on their respective proteins in DENA-17-3. For VNDL-27-4, 49 SAAPs were predicted to have a non-neutral impact. Although these variants are predicted to alter a conserved residue, these are still predictions and therefore may or may not have an impact on protein function.

While determining the functional impact of each variant is beyond the scope of this study, using information from a recent population genomics study of *Populus*,[8] we were able to examine whether identified protein variants belong to genomic regions that appear to be affected by natural selection. Intriguingly, 107 identified SNVs for 84 genes were found in regions of the genome with strong polymorphism patterns consistent with recent positive and/or divergent selection (Supplemental Tables 6 and 7). Of the 107 SNVs, 11% were predicted to have a non-neutral effect, presumably related to purifying selection. Surprisingly, we identified four SAAPs across one gene (Potri.003G043900), which encodes dihydrolipoamide acetyl/succinyl-transferase, and one of the SAAPs in the same gene individually had non-neutral effects. Unfortunately, to our knowledge, there is no validation evidence or natural variants confirming function for this specific protein. However, the ortholog of this particular protein in *A. thaliana* has been shown to have heavy-metal binding properties.[40] Among the highest selection pressures known in ecology, genes annotations associated with heavy-metal homeostasis and symbiosis are overrepresented in classic recent selective sweeps:[41] also known as the "hitch-hiking effect",[42] a selective sweep happens when a new, advantageous mutation is rapidly fixed in a population. Interestingly, Potri.003G043900 was observed by Evans et al.[8] and was shown to be in a region of the genome subjected to a selective sweep. In addition, we identified 25 other genes subjected to a selective sweep. Studying such genes may lead to interesting phenotypes that are important in understanding the nature of natural selection in *Populus*.

Exploring the Diploid Landscape in *Populus* Proteomes

The *Populus* proteome is a function of a diploid entity, yet the protein reference database is incomplete with respect to alleles and their relevant biallelic combinations. Beyond identifying variants, the RNA-Seq-derived protein sequences approach offers the ability to identify and quantify the relationship of alleles of genes. Leveraging information from the whole-genome resequencing data from 1084 *Populus* genotypes (8), each polymorphic position for the two subject genotypes was characterized as *homozygous with the reference* (0/0), *heterozygous with the reference* (0/1) or *homozygous alternate* (1/1). As shown in Figure 4, roughly half of the identified SAAPs belong to heterozygous loci, while the other half belong to a homozygous alternate allele that is genotype-specific. In total, 240 heterozygous polymorphic positions were observed for DENA-17-3, and for 156 of these positions, we identified both the reference and alternate allele version of the protein. We identified 216 heterozygous polymorphic positions for VNDL-27-4, and for 149 of these positions, we identified both the reference and alternate allele version of the protein. While the RNA-Seq-derived protein sequences approach allows the detection of heterozygous variants, we were not able to identify the segregation of variants among each paternal copy because the typical peptide length fell within shared protein space among the two haplotypes. Efforts to build haplotypes during genome assembly will likely extend the characterization of haplotypes at the protein level.[43]

Comparing Identified Sequence Variance Across All -Omic Resources

Currently, direct links between genomic, transcriptomic, and proteomic sequencing data are not frequently made in GWA investigations. Given the breadth of information available, it

DOI: 10.1021/acs.jproteome.5b00823
J. Proteome Res. 2015, 14, 5318−5326

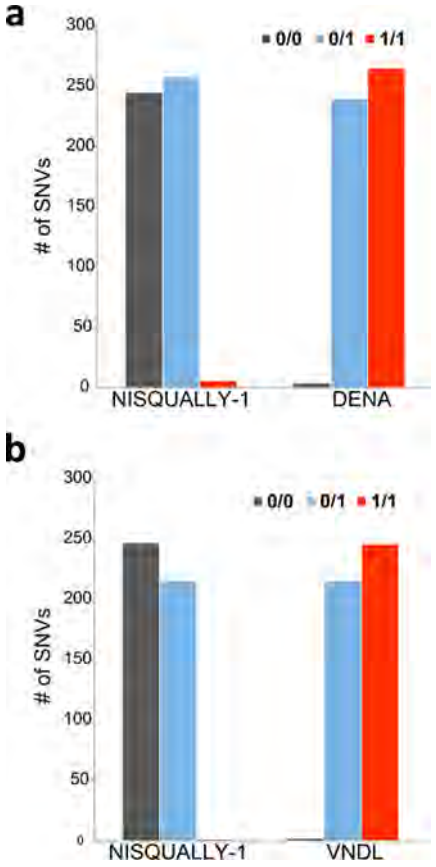


Figure 4. Allele frequency for single nucleotide variant loci. The zygosity of each SNV location was assessed relative to (a) DENA-17-3 and (b) VNDL-27-4. Each locus was classified as one of the following: 0/0 (homozygous reference), 0/1 (heterozygous), or 1/1 (homozygous alternate).

would be highly advantageous to enable the co-visualization of genomics, transcriptomics, and proteomics sequencing to manually validate the discovery of novel protein isoforms. Therefore, to navigate the three available -omic resources (i.e., whole-genome resequencing, RNA sequencing, and protein sequencing), we integrated these -omic data sets using JBrowse,[44] a tool that can be used to compare and correlate the sequence information obtained over a JavaScript-based web browser. The collected -omic resources for DENA-17-3 and VNDL-27-4 have been made publicly accessible and can be viewed at http://besc-portal.ornl.gov/jbrowse/JBrowse-1.11.4/?data=Populus. For example, Figure 5 illustrates how direct links between genomic, transcriptomic, and proteomic data facilitate the validation of variant discovery by observing the precise base position where RNA-sequencing reads and LC–MS/MS-derived peptides cover genes or gene isoforms at the tissue specific level. In this particular illustration, we highlight the mRNA and protein abundance ratios for a heterozygous locus. This is a critical piece of information which could greatly aid and inform the down-selection of possible candidates for connecting genotype to phenotype characteristics. The availability of this integrated data architecture provides a wealth of information that facilitates easy access to visually detailed information across all three -omic resources, highlighting the need for appropriate bioinformatic tools to co-visualize multiple -omic resources.[45]

CONCLUSIONS

Although discoveries of genetic associations will further our understanding of biology, after candidate variants have been identified, investigators are faced with the challenge of functionally characterizing the variants. Therefore, there is a significant need to characterize the impact of these variants at the protein level. In conjunction with a known and defined genome sequence, shotgun proteomics is a remarkable high-throughput technology for identifying and quantifying variant proteins. A fundamental challenge for shotgun proteomics, however, is that the technique ultimately relies on the completeness of the genome sequence. Although several

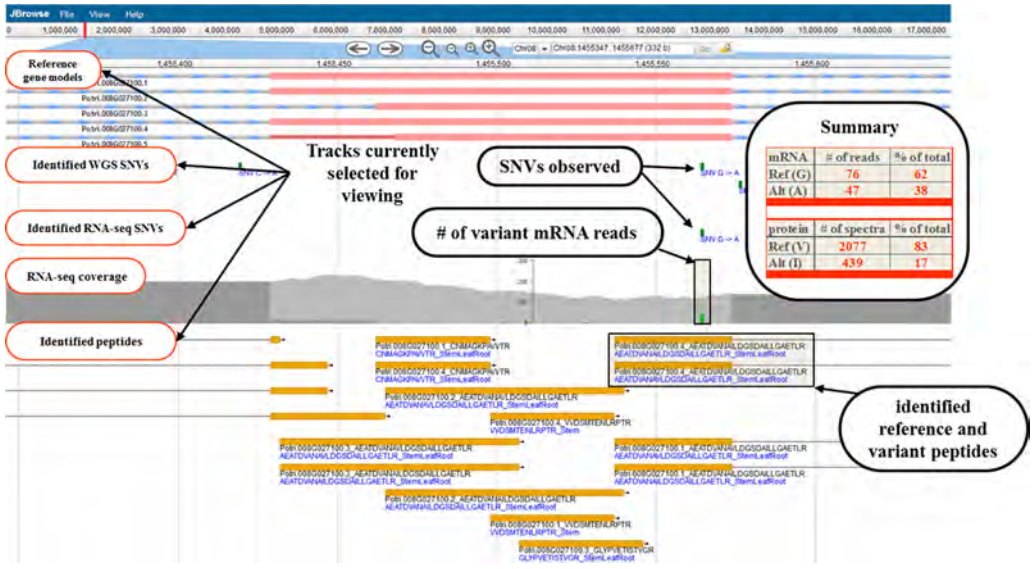


Figure 5. Viewing sequence variance across all -omic resources. A screenshot of the JBrowse resource (http://besc-portal.ornl.gov/jbrowse/JBrowse-1.11.4/?data=Populus) illustrates the available whole-genome resequencing (WGS), RNA-sequencing (RNA-seq), and peptide sequencing data for a specific gene for DENA-17-3. In addition, the mRNA and peptide counts were provided for an example of identifying both alleles of a heterozygous locus.

DOI: 10.1021/acs.jproteome.5b00823
J. Proteome Res. 2015, 14, 5318–5326

bioinformatic solutions for this issue have been presented, we propose that leveraging RNA sequencing information to build sample-specific protein databases will become commonplace when investigating protein sequence variants. Therefore, in this study, we evaluated a conceptual, analytical, and quantitative framework to profile natural sequence variants at the protein level in *Populus*.

Using RNA sequencing data to construct sample-specific protein reference databases, we were able to quantify a variety of natural sequence variants in two *Populus trichocarpa* genotypes. We detected well-known observations related to natural variants; e.g., most single amino acid polymorphisms were conservative changes to chemically and physically similar amino acids, which minimize effects to protein function, verifying the accuracy of our overall approach. In addition, we show how natural amino acid compositions heavily influence the frequency of the types of substitutions. Most importantly, in addition to expected, neutral single amino acid polymorphisms, we identified polymorphisms predicted to be non-neutral and located in regions of the genome predicted to have undergone recent positive and/or divergent selection and therefore represent a candidate list of protein variants relevant to plant adaptability. Additionally, this integrated -omics effort further improves the characterization of genotypes by enabling the reliable detection and quantitation of relevant biallelic combinations of protein isoforms. Overall, this approach afforded the detection of peptide sequence variants spanning 792 unique genome coding positions belonging to 659 loci. Because these variants would not have been detected had we used the reference genome as a proxy for the protein database, profiling genotype-specific proteomes derived from RNA sequencing data better defines the link between genotypes and phenotypes, which will enable future studies to detect and quantitatively profile non-neutral variants underpinning plant adaptation.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.5b00823.

Supplementary Table 1, list of DENA-17-3 identified single nucleotide variants and insertions/deletions and related information; Supplementary Table 2, list of VNDL-27-4 identified single nucleotide variants and insertions/deletions and related information; Supplementary Table 3,A list of DENA-17-3 identified variant peptides and related information; Supplementary Table 4, list of VNDL-27-4 identified variant peptides and related information; Supplementary Table 5, list of identified single amino acid polymorphism and their frequencies for both genotypes; Supplementary Table 6, list of DENA-17-3 variants that are found in regions of the genome with strong polymorphism patterns consistent with recent positive and/or divergent selectionl Supplementary Table 7, list of VNDL-27-4 variants that are found in regions of the genome with strong polymorphism patterns consistent with recent positive and/or divergent selection (ZIP)

DENA-17-3-LEAF, DENA-17-3-ROOT, and DENA-17-3-STEM peptide identifications, and VNDL-27-4-LEAF, VNDL-27-4-ROOT, and VNDL-27-4-STEM Peptide Identifications, including DENA-customized.fasta and VNDL-customized.fasta (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

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

■ ABBREVIATIONS

BCA, bicinchoninic acid assay; GWA, genome-wide association; INDEL, insertion and deletion; LTQ, linear ion trap mass spectrometer; LPI, leaf plastochron index; RPKM, reads per kilobase of transcript per million reads mapped; RNA-seq, RNA sequencing; SAAP, single amino acid polymorphism; TCA, trichloroacetic acid; 2D LC−MS/MS, two-dimensional liquid chromatography−tandem mass spectrometry; SNV, single nucleotide variant

■ REFERENCES

(1) Hirschhorn, J. N.; Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **2005**, *6* (2), 95−108.

(2) Watt, F. M. Mammalian skin cell biology: at the interface between laboratory and clinic. *Science* **2014**, *346* (6212), 937−40.

(3) He, J.; Zhao, X.; Laroche, A.; Lu, Z. X.; Liu, H.; Li, Z. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front. Plant Sci.* **2014**, *5*, 484.

(4) Stranger, B. E.; Stahl, E. A.; Raj, T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* **2011**, *187* (2), 367−83.

(5) Hindorff, L. A.; Sethupathy, P.; Junkins, H. A.; Ramos, E. M.; Mehta, J. P.; et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (23), 9362−7.

(6) Yang, W.; Guo, Z.; Huang, C.; Duan, L.; Chen, G.; et al. Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nat. Commun.* **2014**, *5*, 5087.

(7) Morris, G. P.; Ramu, P.; Deshpande, S. P.; Hash, C. T.; Shah, T.; et al. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (2), 453−458.

(8) Evans, L. M.; Slavov, G. T.; Rodgers-Melnick, E.; Martin, J.; Ranjan, P.; et al. Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations. *Nat. Genet.* **2014**, *46* (10), 1089−1096.

(9) Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10* (1), 57−63.

(10) Lundby, A.; Rossin, E. J.; Steffensen, A. B.; Acha, M. R.; Newton-Cheh, C.; et al. Annotation of loci from genome-wide association studies using tissue-specific quantitative interaction proteomics. *Nat. Methods* **2014**, *11* (8), 868−874.

(11) Zhang, Y. Y.; Fonslow, B. R.; Shan, B.; Baek, M. C.; Yates, J. R. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chem. Rev.* **2013**, *113* (4), 2343−2394.

DOI: 10.1021/acs.jproteome.5b00823
J. Proteome Res. 2015, 14, 5318−5326

(12) Abraham, P.; Adams, R.; Giannone, R. J.; Kalluri, U.; Ranjan, P.; et al. 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.
(13) Abraham, P.; Giannone, R. J.; Adams, R. M.; Kalluri, U.; Tuskan, G. A.; et al. 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.
(14) Baerenfaller, K.; Grossmann, J.; Grobei, M. A.; Hull, R.; Hirsch-Hoffmann, M.; et al. Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science* **2008**, *320* (5878), 938−941.
(15) Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; et al. A draft map of the human proteome. *Nature* **2014**, *509* (7502), 575.
(16) Frank, A.; Pevzner, P. PepNovo: De novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **2005**, *77* (4), 964−973.
(17) Ma, B.; Johnson, R., De Novo Sequencing and Homology Searching. *Mol. Cell. Proteomics* **2012**, *11*, (2).O111.01490210.1074/mcp.O111.014902
(18) Dasari, S.; Chambers, M. C.; Slebos, R. J.; Zimmerman, L. J.; Ham, A. J. L.; et al. TagRecon: High-Throughput Mutation Identification through Sequence Tagging. *J. Proteome Res.* **2010**, *9* (4), 1716−1726.
(19) Abraham, P.; Adams, R. M.; Tuskan, G. A.; Hettich, R. L. Moving Away from the Reference Genome: Evaluating a Peptide Sequencing Tagging Approach for Single Amino Acid Polymorphism Identifications in the Genus Populus. *J. Proteome Res.* **2013**, *12* (8), 3642−3651.
(20) Su, Z. D.; Sheng, Q. H.; Li, Q. R.; Chi, H.; Jiang, X.; et al. De novo identification and quantification of single amino-acid variants in human brain. *J. Mol. Cell Biol.* **2014**, *6* (5), 421−433.
(21) Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **2014**, *11* (11), 1114−25.
(22) Woo, S.; Cha, S. W.; Merrihew, G.; He, Y.; Castellana, N.; et al. Proteogenomic database construction driven from large scale RNA-seq data. *J. Proteome Res.* **2014**, *13* (1), 21−8.
(23) Wang, X.; Slebos, R. J.; Wang, D.; Halvey, P. J.; Tabb, D. L.; et al. Protein identification using customized protein sequence databases derived from RNA-Seq data. *J. Proteome Res.* **2012**, *11* (2), 1009−17.
(24) Wang, X. J.; Zhang, B. Integrating Genomic, Transcriptomic, and Interactome Data to Improve Peptide and Protein Identification in Shotgun Proteomics. *J. Proteome Res.* **2014**, *13* (6), 2715−2723.
(25) Evans, V. C.; Barker, G.; Heesom, K. J.; Fan, J.; Bessant, C.; et al. De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nat. Methods* **2012**, *9* (12), 1207−11.
(26) Tuskan, G. A.; DiFazio, S.; Jansson, S.; Bohlmann, J.; Grigoriev, I.; et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **2006**, *313* (5793), 1596−1604.
(27) Geraldes, A.; Difazio, S. P.; Slavov, G. T.; Ranjan, P.; Muchero, W.; et al. A 34K SNP genotyping array for Populus trichocarpa: Design, application to the study of natural populations and transferability to other Populus species. *Mol. Ecol. Resour.* **2013**, *13* (2), 306−323.
(28) Slavov, G. T.; DiFazio, S. P.; Martin, J.; Schackwitz, W.; Muchero, W.; et al. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree Populus trichocarpa. *New Phytol.* **2012**, *196* (3), 713−25.
(29) 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.
(30) 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.
(31) Ma, Z. Q.; Dasari, S.; Chambers, M. C.; Litton, M. D.; Sobecki, S. M.; et al. IDPicker 2.0: Improved Protein Assembly with High Discrimination Peptide Identification Filtering. *J. Proteome Res.* **2009**, *8* (8), 3872−3881.
(32) Kapp, E.; Schutz, F. Overview of tandem mass spectrometry (MS/MS) database search algorithms. *Current Protocols in Protein Science*; Wiley: New York, 2007; Chapter 25, Unit 25.2.
(33) Geraldes, A.; Pang, J.; Thiessen, N.; Cezard, T.; Moore, R.; et al. SNP discovery in black cottonwood (Populus trichocarpa) by population transcriptome resequencing. *Mol. Ecol. Resour.* **2011**, *11*, 81−92.
(34) Wang, X. J.; Zhang, B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **2013**, *29* (24), 3235−3237.
(35) Kim, D.; Pertea, G.; Trapnell, C.; Pimentel, H.; Kelley, R.SalzbergS. L., , TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **2013**, *14*, (4).R3610.1186/gb-2013-14-4-r36
(36) Nagaraj, N.; Kulak, N. A.; Cox, J.; Neuhauser, N.; Mayr, K.; et al. System-wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-shot Ultra HPLC Runs on a Bench Top Orbitrap. *Mol. Cell. Proteomics* **2012**, *11* (3), M111.013722.
(37) Nowicka, A.; Mackiewicz, P.; Dudkiewicz, M.; Mackiewicz, D.; Kowalczuk, M., et al., Correlation between mutation pressure, selection pressure, and occurrence of amino acids. Computational Science - Iccs 2003, Pt Ii, Proceedings 2003, *2658*, 650−657.
(38) Hormoz, S., Amino acid composition of proteins reduces deleterious impact of mutations. *Sci. Rep.* **2013**, 3.291910.1038/srep02919
(39) Choi, Y.; Sims, G. E.; Murphy, S.; Miller, J. R.; Chan, A. P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS One* **2012**, *7*, e46688.
(40) Tan, Y. F.; O'Toole, N.; Taylor, N. L.; Millar, A. H. Divalent metal ions in plant mitochondria and their role in interactions with proteins and oxidative stress-induced damage to respiratory function. *Plant Physiol.* **2010**, *152* (2), 747.
(41) Hanikenne, M.; Kroymann, J.; Trampczynska, A.; Bernal, M.; Motte, P.; et al. Hard selective sweep and ectopic gene conversion in a gene cluster affording environmental adaptation. *PLoS Genet.* **2013**, *9* (8), e1003707.
(42) Smith, J. M.; Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **2007**, *89* (5−6), 391−403.
(43) Hoehe, M. R.; Church, G. M.; Lehrach, H.; Kroslak, T.; Palczewski, S., et al., Multiple haplotype-resolved genomes reveal population patterns of gene and protein diplotypes. *Nat. Commun.* **2014**, *5*.556910.1038/ncomms6569
(44) Skinner, M. E.; Uzilov, A. V.; Stein, L. D.; Mungall, C. J.; Holmes, I. H. JBrowse: A next-generation genome browser. *Genome Res.* **2009**, *19* (9), 1630−1638.
(45) Pang, C. N. I.; Tay, A. P.; Aya, C.; Twine, N. A.; Harkness, L.; et al. Tools to Covisualize and Coanalyze Proteomic Data with Genomes and Transcriptomes: Validation of Genes and Alternative mRNA Splicing. *J. Proteome Res.* **2014**, *13* (1), 84−98.

DOI: 10.1021/acs.jproteome.5b00823
J. Proteome Res. 2015, 14, 5318−5326