Received: 7 June 2022 | Revised: 16 August 2022 | Accepted: 23 August 2022

DOI: 10.1002/pmic.202100377

Proteomics
Proteomics and Systems Biology

RESEARCH ARTICLE

Large-scale top-down proteomics of the *Arabidopsis thaliana* leaf and chloroplast proteomes

Qianjie Wang[1,2,3] | Liangliang Sun[3] | Peter Knut Lundquist[1,2]

[1]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan, USA

[2]Plant Resilience Institute, Michigan State University, East Lansing, Michigan, USA

[3]Department of Chemistry, Michigan State University, East Lansing, Michigan, USA

Correspondence
Peter Knut Lundquist, Department of Biochemistry & Molecular Biology & the Plant Resilience Institute, Michigan State University, Plant & Soil Science Building, Room A498E, East Lansing, MI 48864, USA.
Email: pklundqu@msu.edu

Funding information
National Institute of General Medical Sciences, Grant/Award Numbers: R01GM118470, R01GM125991; National Cancer Institute, Grant/Award Number: R01CA247863; National Science Foundation, Grant/Award Numbers: MCB-2034631, DBI-1846913

Abstract

We present a large-scale top-down proteomics (TDP) study of plant leaf and chloroplast proteins, achieving the identification of over 4700 unique proteoforms. Using capillary zone electrophoresis coupled with tandem mass spectrometry analysis of offline size-exclusion chromatography fractions, we identify 3198 proteoforms for total leaf and 1836 proteoforms for chloroplast, with 1024 and 363 proteoforms having post-translational modifications, respectively. The electrophoretic mobility prediction of capillary zone electrophoresis allowed us to validate post-translational modifications that impact the charge state such as acetylation and phosphorylation. Identified modifications included Trp (di)oxidation events on six chloroplast proteins that may represent novel targets of singlet oxygen sensing. Furthermore, our TDP data provides direct experimental evidence of the N- and C-terminal residues of numerous mature proteoforms from chloroplast, mitochondria, endoplasmic reticulum, and other sub-cellular localizations. With this information, we suggest true transit peptide cleavage sites and correct sub-cellular localization signal predictions. This large-scale analysis illustrates the power of top-down proteoform identification of post-translational modifications and intact sequences that can benefit our understanding of both the structure and function of hundreds of plant proteins.

KEYWORDS
Arabidopsis thaliana, chloroplast, mass spectrometry, post-translational modifications, proteoform, singlet oxygen, top-down proteomics, transit peptide

1 | INTRODUCTION

Proteins often undergo modifications following their translation, such as proteolytic processing or addition of covalent linkages that are critical for proper function. Because of such modifications, a bewildering array of potential proteoforms, each with a distinct chemical structure, can arise from a single genetic locus [1]. The combination of such modifications, ultimately resulting in a mature proteoform, can profoundly influence protein function, stability, interaction, structure, localization, or activity by altering physico–chemical properties of the protein. For example, a subset of the *Arabidopsis thaliana* light-harvesting complex II subunit pool is dynamically phosphorylated in response to light quality shifts, promoting their migration within the thylakoid membrane [2]. Similarly, the N-terminal residue of a sequence influences protein function and dictates protein stability through the so-called N-end rule [3, 4].

Abbreviations: cTP, chloroplast transit peptide; CZE, capillary zone electrophoresis; luTP, lumenal transit peptide; MS, mass spectrometry; mTP, mitochondrial transit peptide; SP, signal peptide.

Proteolytic processing of proteins often serves to remove N-terminal sequence tags used to target a protein to its proper sub-cellular localization. These N-terminal extensions can range from 15 to 162 amino acids in length, and are proteolytically removed after import [5, 6] and, at least in some cases, are likely further processed at the N-terminus by unidentified peptidases [7]. The chloroplast alone is estimated to harbor approximately 3000 nuclear-encoded proteins, targeted using an obligatory N-terminal chloroplast transit peptide (cTP). A subset of such chloroplast-targeted proteins are subsequently further targeted to the lumen of the internal thylakoid membrane, requiring a second cleavable protein sequence called the lumenal targeting peptide (luTP), immediately downstream of the cTP. Likewise, signal peptides (SPs) are necessary for proper targeting to the secretory pathway including the endoplasmic reticulum and Golgi, while mitochondrial transit peptides (mTPs) are necessary for targeting to the mitochondria. The different sub-cellular localization signals have broadly distinct characteristics that can facilitate their prediction from primary protein sequences. Multiple algorithms have been developed to predict sub-cellular localization sequences and propose cleavage sites of the localization signal [5, 8–10, 11]. However, sequence conservation among targeting signals is almost wholly non-existent, and exceptions to the general sequence patterns abound. This has made the prediction of targeting signals difficult [5].

Numerous bottom-up proteomics (BUP) studies have been undertaken for the large-scale determination of N- and C-termini of mature protein sequences. A comprehensive proteomics analysis of the chloroplast was performed in part to establish the N-termini of chloroplast proteins and thereby propose cTP cleavage sites [12]. However, information on N-termini was limited to the subset of N-terminally acetylated tryptic peptides. Subsequent efforts employed a covalent tagging approach that greatly expanded the coverage of identified N-termini of chloroplast proteoforms [7]. This work identified a clear enrichment for N-terminal residues of Ala, Val, Thr, and Ser. However, the bottom-up nature of the study limited the ability to characterize N-termini in the context of full-length sequence or possible covalent modifications.

In contrast, top-down proteomics (TDP) directly characterizes the primary, intact sequence of different proteoforms. In 2002, Whitelegge et al. applied intact mass measurements to the chloroplast grana proteome, in which one of the first single-pass membrane proteoforms was defined [13]. Since then, the subunits of the cytochrome b_6f complex [14, 15], the photosystem II complex (PSII) [16], and the 26S proteasome [17] have been investigated using TDP. Novel insights, such as the presence of palmitoylation, phosphorylation and distinct lipid modifications have been gleaned [18], expanding our understanding of the composition and assembly of large protein complexes of the plant cell. TDP also provides an effective strategy to determine the mature (i.e., post-transit peptide cleavage) proteoform identities of a proteome while avoiding extra sample handling steps and artificial covalent modifications [19]. Smith et al. have established a five-level classification system that assesses the ambiguity a given proteoform identification concerning the PTM localization, PTM identification, amino acid sequence, and gene, ranging from no ambiguity (Level 1) to ambiguity among all four categories (Level 5) [20]. Among the first applications of TDP to chloroplast samples was the use of three-dimensional Fourier transform MS [21]. Of the 22 molecular weight values found (from 9 to 26 kDa), seven proteins were fully characterized, in comparison to 97 identified by BUP. The application of TDP could delineate similar proteins differing only by 12 residues, differentiate proteins with and without N-methylation, and correct the cleavage site of transit peptides. While the TDP applications from these early studies represent pioneering efforts that provided significant biological insights, characterization of sequence tags was performed manually, and protein separation was performed offline with direct infusion [21], or by reverse-phase liquid chromatography (RPLC) separation [13].

Compared with RPLC, capillary-zone electrophoresis (CZE) is known for its high separation efficiency for large biomolecules and high sensitivity for intact protein characterization [22, 23]. The advanced CZE-MS/MS interface [24, 25], capillary coating [26], and online stacking methods enabled identification of nearly 600 intact proteoforms from an *Escherichia coli* cell lysate in a single shot CZE- MS/MS [27]. Furthermore, 5700 proteoforms were identified from *E. coli* lysate by combining size exclusion chromatography (SEC) with RPLC pre-fractionation [28]. Orthogonal to SEC and RPLC, CZE separates proteoforms according to their different electrophoretic mobility (μ_{ef}), which is directly related to the size and charge of the proteoform. Thus, charge-modified PTMs should alter mobility of proteoforms in a predictable manner, unlike migration in RPLC [29].

SIGNIFICANCE OF THE STUDY

- Top-down proteomics is uniquely capable of characterizing the mature chemical structures (i.e., proteoforms) of proteins that result from post-translational modifications. However, top-down proteomics has been applied relatively rarely in the field of plant biology. In this study, we sought to demonstrate the capability of top-down proteomics in the context of plant leaf tissue, with focus on the photosynthetically-active chloroplast organelle. Using capillary-zone electrophoresis coupled with tandem mass spectrometry, we identified over 4700 unique proteoforms and determined the mature N-termini of over 200 proteins localized to multiple sub-cellular compartments. We suggested corrected cleavage sites for 35 sub-cellular localization signals. Seven proteins were identified with Trp (di)oxidation, six of which are chloroplast-localized, that may represent novel targets of singlet oxygen sensing. Finally, we demonstrate the capability of capillary zone electrophoresis to validate (and in some cases correct) post-translational modification identifications based on predictable electrophoretic migration patterns. Our results reveal novel insight into the mature protein structure of hundreds of plant proteins and demonstrate the great potential of top-down proteomics in plant biology.

With the substantial advancements in informatics tools for proteoform identification, such as ProSight [30, 31] and TopPIC suite [32], and the advancement of multi-dimensional separations, numerous TDP experiments have been successfully applied to human and animal samples [33]. In contrast, large-scale TDP studies in plants have been broadly lacking since the initial studies of the early 2000s. Given that, we performed a large-scale TDP analysis of *A. thaliana* leaf and chloroplast samples using two-dimensional orthogonal separations of SEC followed by CZE-MS/MS. *A. thaliana* was selected for our study as it represents the foremost model plant species with a high-quality annotated genome. We identified 3198 and 1836 proteoforms from the total leaf and the chloroplast sample, respectively, and a total of 4782 unique proteoforms across the two samples. We identified numerous PTMs, established protein N-termini, and corrected predicted sub-cellular localization signals. New chloroplast protein targets of Trp oxidation, indicative of singlet oxygen retrograde signaling, were found. This work fills a significant gap in plant proteoform characterization, demonstrates the advancement of TDP methods, and provides the foundation for future developments in the characterization of intact protein species of plant proteomes.

2 | MATERIALS AND METHODS

2.1 | Materials and reagents

Acrylamide was purchased from Acros Organics (NJ, USA). Ammonium bicarbonate (NH_4HCO_3), urea, dithiothreitol (DTT), iodoacetamide (IAA) and 3-(Trimethoxysilyl) propyl methacrylate, Tris-HCl, Hepes, $MgCl_2$, phosphatase inhibitors: NaF, β-Glycerophosphate 2Na·5H_2O, Na-Orthovanadate, Na-Pyrophosphate·10H_2O and protease inhibitors: Antipain·2HCl, Bestatin, Chymostatin, E-64, Leupeptin (hemisulfate), P-ramidon·2Na, AEBSF, Aprotinin were purchased from Sigma-Aldrich (St. Louis, MO). LC/MS grade water, acetonitrile (ACN), methanol, and formic acid were purchased from Fisher Scientific (Pittsburgh, PA). Aqueous mixtures were filtered with Nalgene Rapid-Flow Filter units (Thermo Scientific) with 0.2 μm CN membrane and 50 mm diameter. Fused silica capillaries (50 μm i.d./360 μm o.d.) were obtained from Polymicro Technologies (Phoenix, AZ). Complete, mini protease inhibitor cocktail (provided in EASYpacks) was bought from Roche (Indianapolis, IN).

2.2 | Sample collection

Total leaf samples: Two trays of 8-week-old *A. thaliana* (ecotype Columbia-0) were grown at 16/8 light/dark photoperiod, 20°C. Leaves were cut from 64 plants, pooled and flash frozen in a mortar with liquid nitrogen, and thoroughly ground. The powder was then mixed with a lysis buffer containing 50 mM Tris-HCl (pH 8.0), 2% SDS and protease inhibitor cocktail by pipetting, and then vortexed for 25 s. After centrifugation (13,000 rpm for 2.5 min), the supernatant containing the extracted proteins was collected and stored in –80°C. Whole chloroplast samples: 4 trays (128 plants) of 39-day-old *A. thaliana* Col-0 were grown at 10/14 light/dark photoperiod. Leaves were cut and washed in pre-cold water, and ground in isolation buffer (330 mM sorbitol, 20 mM Hepes, 13 mM Tris-HCl, 3 mM $MgCl_2$, 0.1% fat-free BSA, 5 mM ascorbic acid, and 5 mM reduced cysteine, and phosphatase inhibitors) using a Waring blender with medium intensity for 10 s. Lysate was filtered through one layer of gauze, and then centrifuged for 5 min at 1500 g. The supernatant was discarded, and the pellet was resuspended in the wash buffer (330 mM sorbitol and 50 mM Hepes with phosphatase inhibitors). Pellet was washed and re-collected with 5 min centrifugation at 1500 g, after which the pellet was resuspended in 6 ml osmotic shocking buffer (0.6 M sucrose, 1 mM EDTA, 10 mM Tricine, protease inhibitors and phosphatase inhibitors), sitting on ice for 30 min. Samples were collected in 15 ml falcon tubes, lyophilized, and stored in –80°C. To prepare samples for MS/MS, they were thawed on ice, sonicated with a Branson Sonifier 250 (VWR Scientific, Batavia, IL) on ice for 10 min, and then resuspended in 2% SDS with protease inhibitor cocktail. The protease inhibitors with final concentrations are listed below: 50 μg/ml Antipain· 2HCl, 40 μg/ml Bestatin, 10 μg/ml Chymostatin, 10 μg/ml E-64, 5 μg/ml Leupeptin (hemisulfate), 10 μg/ml P-ramidon· 2Na, 50 μg/ml AEBSF, 2 μg/ml Aprotinin. The concentration of phosphatase inhibitors is: 50 mM NaF, 25 mM β-Glycerophosphate·2Na·5H_2O, 1 mM Na-Orthovanadate, 10 mM Na-Pyrophosphate·10H_2O.

2.3 | Sample preparation

A 4:1 (v/v) ratio of acetone was added to solubilized protein samples (both chloroplast and total leaf) with overnight precipitation. A 10,000 × *g* centrifugation removed the supernatant, and the protein pellet was resuspended in 8 M urea and 100 mM ammonium bicarbonate (pH 8.0), denatured at 37°C for 30 min, reduced with dithiothreitol (DTT) at 37°C for 30 min and alkylated with iodoacetamide (IAA) at room temperature without light for 20 min. Then, samples were desalted by a 30 kDa molecular weight cut off centrifugal filter (Millipore Sigma, Inc.) washed with 100 mM NH_4HCO_3 (pH 8.0). Finally, sample was diluted into 50 mM NH_4HCO_3 (pH 8.0).

2.4 | Size exclusion chromatography (SEC) separation

Samples were fractionated by SEC in preparation for CZE-MS/MS analysis. For total leaf sample, the SEC column was 4.6 × 300 mm, 3 μm particles, 300 Å pores from Agilent, the mobile phase was 0.1% (v/v) FA, and the flow rate was 0.25 ml/min. The column temperature was kept at 40°C. We collected six fractions from 10–22 min (2 min for each fraction) from 120 μl of 1 mg/ml total leaf sample input. For chloroplast sample, the Bio SEC-5 column (4.6 × 300 mm, 3 μm particles, 500 Å pores) from Agilent was used. The mobile phase was 0.4% (v/v) FA, and the flow rate was 0.25 ml/min. The column temperature was kept at

40°C. We collected nine fractions from 10 to 36 min (4 min for each fraction except for 2 min for the 3rd, 4th, 5th, and 6th fractions) from 60 μl of 2 mg/ml chloroplast protein sample input.

2.5 | CZE-MS/MS

An automated ECE-001 CE autosampler and a commercialized electrokinetically pumped sheath flow CE—MS interface from CMP Scientific (Brooklyn, NY) [24, 25] was coupled to a Q-Exactive HF mass spectrometer (Thermo Fisher Scientific). The protocol of a 100-min CZE-MS/MS and parameters of QE-HF were according to McCool, et al [28]. Briefly, a fused silica capillary (50 μm i.d., 360 μm o.d., 1 m) was coated with linear polyacrylamide (LPA) and etched with hydrofluoric acid at the end near the CE-MS interface to reduce the outer diameter of the capillary. Acetic acid, 10% (v/v), was used as the background electrolyte (BGE). The sheath buffer was 0.2% (v/v) formic acid containing 10% (v/v) methanol. Sample injection was carried out by applying pressure (5 psi) at the sample injection end, and the injection periods were calculated based on Poiseuille's law for different sample loading volumes [34]. For the total leaf sample, the 500 nl injection volume was adopted with ~500 ng of protein in each sample assuming equal distribution of protein among each SEC fraction. For chloroplast samples which have a relatively smaller proteome, the injection volume was 250 nl (~160 ng of protein assuming equal distribution among SEC fractions) except 200 nl for fraction 5 and 6 due to the higher protein abundance. A dynamic pH junction was used to concentrate sample in acidic BGE to accommodate the large injection volumes [35, 36, 37, 23]. A 30 kV high voltage was applied at the injection end of the capillary and around 2.0–2.2 kV was applied for electrospray. MS parameters are listed as follows. Microscan is 3 for both full MS and MS/MS. For full MS, the resolution was 240,000 at m/z 200, and AGC target value was 1E6 with 50 ms maximum injection time. The scan range was 600–2000 m/z and the top 5 ions of the highest intensity in full MS were isolated with a 4 m/z isolation window and fragmented with a 20% normalized collision energy. The resolution for MS/MS is 120,000 at m/z 200, and the AGC target was 1E5 with 200 ms maximum injection time. Intact protein mode and exclude isotopes settings were on. Proteins with 1–5 charge state were excluded and the dynamic exclusion was set for 30 s.

2.6 | Data analysis

All .raw files were converted to mzML by MSconvert tool and then analyzed with the TopFD and TopPIC pipeline [32] with an estimated 1% FDR at the spectrum level and 5% FDR at the proteoform level. Cysteine carbamidomethylation was set as a fixed modification. The maximum number of unexpected modifications was 2. The precursor and fragment mass error tolerances were 15 ppm. The maximum mass shift of unknown modifications was 500 Da.

3 | RESULTS

3.1 | TDP workflow of leaf and chloroplast samples

The general workflow for capillaryzone electrophoresis separation coupled with electrospray ionization-tandem mass spectrometry (CZE-ESI-MS/MS) is shown in Figure 1, along with a representative electropherogram. A sample of total leaf tissue was prepared from wild-type *A. thaliana* leaves late in the vegetative growth stage using 2% SDS in the presence of a protease inhibitor cocktail. After sample collection and purification, the sample was separated by SEC into six fractions, followed by a 100-minute-CZE-ESI-MS/MS online separation of each fraction. Using an estimated 5% FDR at the proteoform level and 1% FDR at the MS/MS spectrum level by the TopPIC suite [32], we identified 3198 unique proteoforms from 458 proteins across the six SEC fractions of the total leaf sample with an average of 20.8 matched fragment ions per proteoform (Table S1 and Figure S1A). According to the 5-level classification system established by Smith et al. [20] that describes the level of ambiguity within a proteoform identification, 47.9% of proteoform identifications are categorized as Level 1, indicating no ambiguity at all, in which PTMs are both characterized and well assigned (Table 1).

To provide a targeted survey of chloroplast proteoforms, we also investigated whole chloroplast samples isolated from *A. thaliana* leaf tissue during the middle of vegetative growth. The same pipeline as for total leaf was used for the subsequent proteomics analysis using nine SEC fractions with a 120-minute CZE-ESI-MS/MS online separation for each fraction. In total, 1836 proteoforms from 200 proteins were identified in the chloroplast sample, with 40.7% of proteoform identifications categorized at Level 1 (Table S2 and Figure S1B). The proteoforms were identified with an average of 20.7 fragment ions per proteoform. We illustrate MS/MS spectra and electropherograms of five representative proteoforms in Figures S3–S7. Comparing the total leaf and chloroplast samples, identifications of 242 proteoforms from 99 proteins are present in both experiments (Figure S2). As an example output, the fragmentation pattern of a double phosphorylated proteoform of Thylakoid Soluble Phosphoprotein 9 (at3g47070) is shown in Figure 1. Residue-level assignment of the two phosphorylation sites is possible due to the fragmentation within the consecutive Thr residues.

3.2 | Proteoform mass shifts and post-translational modifications

Across the total leaf samples, a total of 2390 mass shifts were identified. In fact, over 61% of identified proteoforms in our datasets contained at least one mass shift. We generated a histogram of these shifts to identify those most frequently represented within our datasets (Figure 2A). The most prevalent mass shifts match with common PTMs, such as acetylation (460 proteoforms), N-terminal Met excision (357 proteoforms), oxidation (179 proteoforms), and methylation (28 proteoforms) (Figure S8A). Acetylation was the most abundant

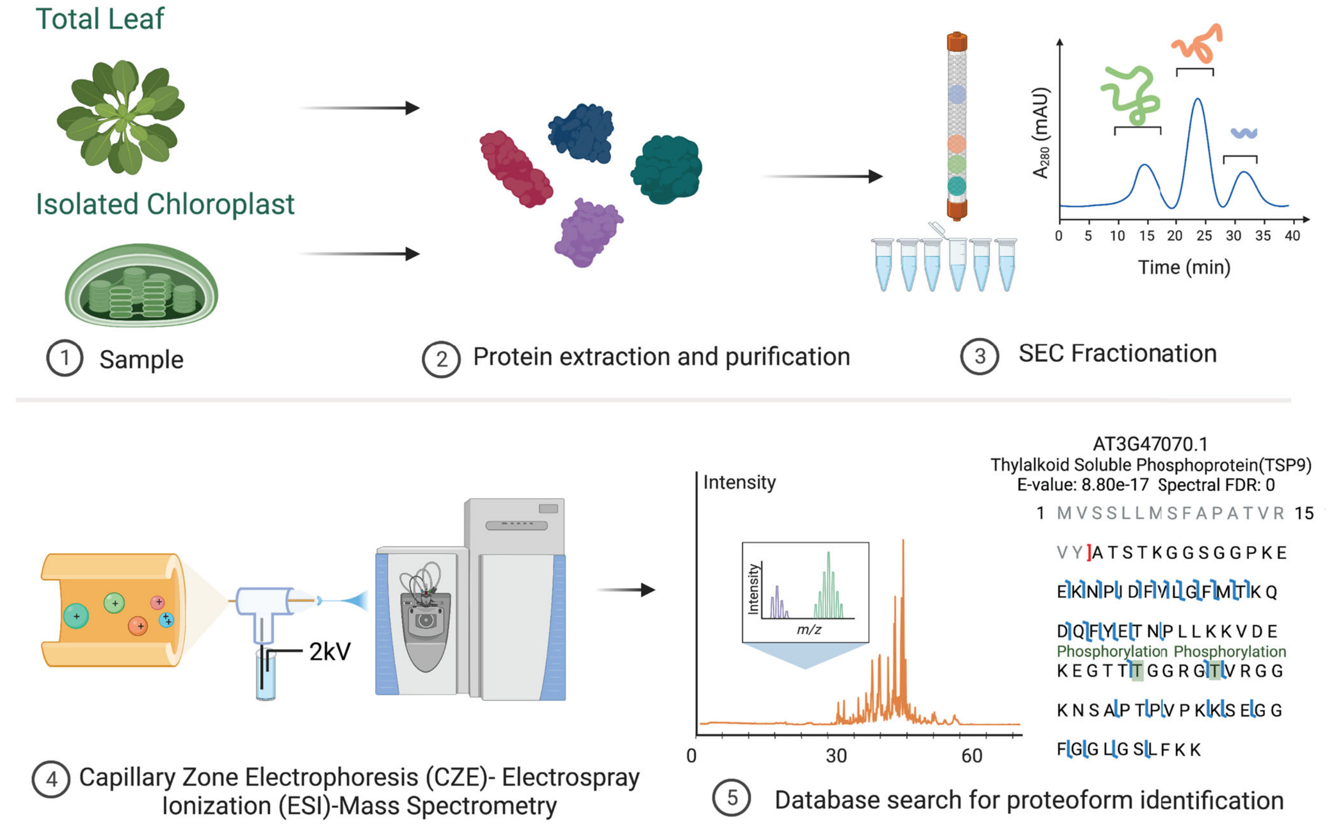

Total Leaf
Isolated Chloroplast
1 Sample
2 Protein extraction and purification
3 SEC Fractionation
A280 (mAU)
0 5 10 15 20 25 30 35 40
Time (min)
2kV
4 Capillary Zone Electrophoresis (CZE)- Electrospray Ionization (ESI)-Mass Spectrometry
Intensity
Intensity
m/z
0 30 60
AT3G47070.1
Thylalkoid Soluble Phosphoprotein(TSP9)
E-value: 8.80e-17 Spectral FDR: 0
1 MVSSLLMSFAPATVR 15
VYATSTKGGSGGPKE
EKNPIDFVLGFMTKQ
DQFYETNPLLKKVDE
Phosphorylation Phosphorylation
KEGTTTGGRGTVRGG
KNSAPTPVPKKSEGG
FGGLGSLFKK
5 Database search for proteoform identification

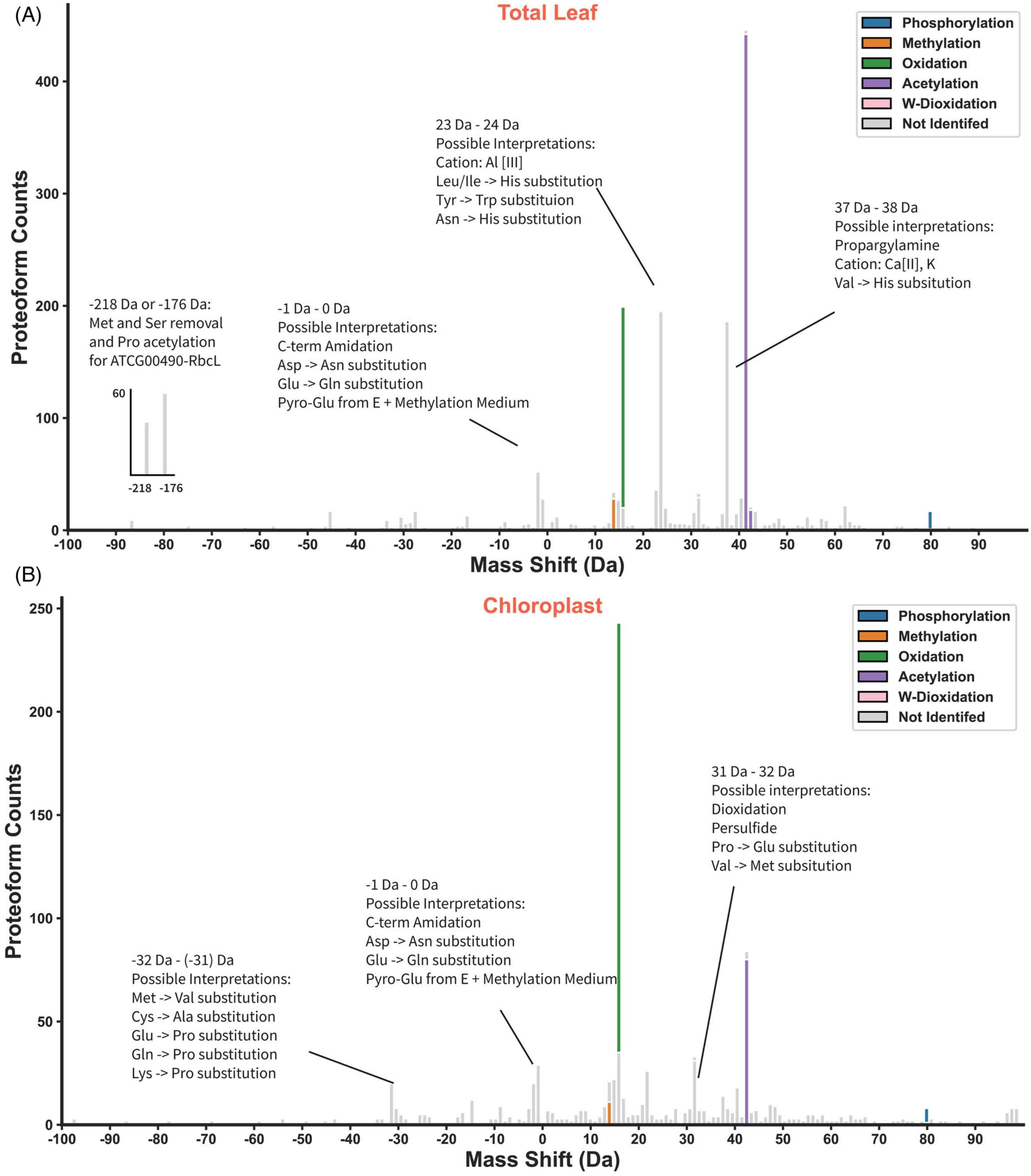
FIGURE 1 The top-down proteomics workflow of total leaf and isolated chloroplast samples. After isolation of total leaf and chloroplast samples, proteins were extracted in 2% (w/v) sodium dodecyl sulfate detergent, precipitated by acetone, resuspended with 8 M Urea in 100 mM ammonium bicarbonate (ABC), and buffer exchanged into 100 mM ABC. After separation into six size-exclusion chromatography (SEC) fractions, they were characterized by CZE-ESI-MS/MS and identified with the TopPIC MS/MS analysis suite using the TAIR10 *A. thaliana* protein sequence database. An example of an identified double phosphorylation on the Thylakoid Soluble Protein 9 (TSP9) is shown with its fragmentation pattern. Figure Made in BioRender.com.

TABLE 1 5-Level classification of identified proteoforms

Level:		1	2A	2B	3	4	5	Total
Total Leaf	No	1233	0	0	0	0	0	1233
	1 Mod	272	42	177	1101	0	0	1592
	2 Mods	26	0	22	267	0	0	315
	3 Mods	1	0	0	57	0	0	56
	Total	1532	42	199	1425	0	0	3198
Chloroplast	No	636	0	0	0	0	0	636
	1 Mod	101	105	119	741	0	0	1272
	2 Mods	10	7	6	111	0	0	134
	3 Mods	0	0	0	0	0	0	0
	Total	747	112	125	852	0	0	1836

PTM identified in the total leaf proteoforms, 87% of which was found to occur on the N-terminus, generally accompanied by Met excision. Oxidation was found on multiple amino acids, most frequently on Lys. Twenty-eight proteoforms were methylated which localized on Lys, Gln, Asp, and Glu. Sixteen proteoforms were found to be phosphorylated despite the absence of special enrichment of phosphoproteins or the use of phosphatase inhibitors. Numerous other mass shifts were found which could not be readily assigned to a specific PTM. Possible interpretations of these unassigned mass shifts, based on the Unimod database [38], are indicated in Figure 2A.

While almost all mass shifts fell within the −100 to +100 Da range, we did identify two peaks of −176.1 Da and −218.1 Da found on 59 and 49 proteoforms, respectively, of the Rubisco large subunit (RbcL, tcg00490). Manual interpretation indicated that software incorrectly predicted the N-terminal residue to be the initiating Met1. The mass shift of −176.1 Da was found on proteoforms lacking a predicted acetylation, and was consistent with removal of Met1 and Ser2 and inclusion of an acetylation. Meanwhile, the -218.1 Da mass shift was found on proteoforms with a predicted acetylation and was consistent with removal of Met1 and Ser2. Similarly, 14 RbcL proteoforms, all with a predicted removal of Met1, were found with a mass shift of −45.1 Da, consistent with removal of Ser2 and an acetylation. Thus, we manually corrected 122 RbcL proteoforms, all resulting in Pro3 as the first residue and with the presence of an acetylation (Table S1). In total, 144 proteoforms of RbcL were found with Pro3 as first residue and an acetyl PTM. This assignment is consistent with previous studies [39]



Total Leaf

-218 Da or -176 Da:
Met and Ser removal
and Pro acetylation
for ATCG00490-RbcL

-1 Da - 0 Da
Possible Interpretations:
C-term Amidation
Asp -> Asn substitution
Glu -> Gln substitution
Pyro-Glu from E + Methylation Medium

23 Da - 24 Da
Possible Interpretations:
Cation: Al [III]
Leu/Ile -> His substitution
Tyr -> Trp substituion
Asn -> His substitution

37 Da - 38 Da
Possible interpretations:
Propargylamine
Cation: Ca[II], K
Val -> His subsitution

Chloroplast

-32 Da - (-31) Da
Possible Interpretations:
Met -> Val substitution
Cys -> Ala substitution
Glu -> Pro substitution
Gln -> Pro substitution
Lys -> Pro substitution

-1 Da - 0 Da
Possible Interpretations:
C-term Amidation
Asp -> Asn substitution
Glu -> Gln substitution
Pyro-Glu from E + Methylation Medium

31 Da - 32 Da
Possible interpretations:
Dioxidation
Persulfide
Pro -> Glu substitution
Val -> Met subsitution

FIGURE 2 Proteoform mass shift distribution from −100 Da to +100 Da. Common PTMs are marked according to the color code indicated in the legend. (A) A histogram of mass shifts among proteoforms identified in the total leaf sample. Two bins containing prevalent mass shifts which are out of the indicated range are included as subsets in the bottom left, both of which are specific to RbcL proteoforms. Because the underlying predicted N-terminus of the proteoforms are different, both mass shifts result in RbcL proteoforms that initiate with an N-terminally acetylated Pro3 residue, as described in the text. (B) A histogram of mass shifts among proteoforms identified in the chloroplast sample. The bin size of all histograms is 1 Da. Possible interpretations of the three most prevalent unidentified mass shifts are proposed based on the Unimod database.

and with their electrophoretic mobility in CZE, as described below in Section 3.3. It is notable that very few proteoforms were found to start with Ser2, which amount to less than 4% of total RbcL proteoform feature intensity. Although the mature form of RbcL is recognized to begin with an N-acetylated Pro3, it remains unclear whether processing of the initiating Met1 and Ser2 occurs in a stepwise fashion or as a single cleavage event between Ser2 and Pro3. Our proteoform identifications uncover little evidence of an intermediate state in which only the Met1 is removed. This strongly suggests that N-terminal processing of RbcL occurs in a single step from an unknown dipeptidase, as suggested previously [40].

We identified 16 proteoforms with phosphorylation in the total leaf sample, including five different phosphorylated proteoforms of Plastocyanin-1 and -2 (at1g76100, at1g20340). Although twelve of the sixteen proteoforms were chloroplast-localized, a non-overlapping set of phosphorylated proteoforms were identified in the chloroplast sample. This included three proteoforms of TSP9 (Table S2). As noted above and observed in previous BUP experiments [41, 42], double phosphorylation of Thr66 and Thr71 was observed. Phosphorylation on Thr64 was also observed in two other proteoforms, however never shared with phosphorylation of Thr66 or Thr71. This suggests that phosphorylation of the Thr66/Thr71 pair and phosphorylation of Thr64 may be mutually exclusive. Curiously, oxidation of Met42 on TSP9 was also observed in high abundance in the chloroplast sample, although never in combination with phosphorylation of Thr64 or Thr66/Thr71.

Across the chloroplast sample, we found a total of 1540 mass shifts (Table S2). Over 76.6% of identified proteoforms include at least one mass shift. As with total leaf sample, the most prevalent mass shifts match with common PTMs such as oxidation (208 proteoforms), acetylation (80 proteoforms), Met excision (64 proteoforms), and methylation (11 proteoforms) (Figures 2B and S8B). In addition, we also found that the relative ratio of oxidation is substantially higher in the chloroplast sample (5.6% of proteoforms vs. 13.5% of proteoforms, respectively), which we suggest to be physiologically relevant as a reflection of the high oxidative pressure found in the chloroplast [43, 44, 45] (Tables S1 and S2). In support of physiologically relevant oxidation in the chloroplast, we identified (di)oxidation of tryptophan (+15.99 and +31.99 Da) on six different chloroplast proteins: Photosystem I reaction center subunit N (PsaN; at5g64040), CP12 (at2g47400), CP12-like (at3g62410), Photosystem II light harvesting complex protein (LHCII-1.5; at2g34420), Chlorophylll a/b binding protein 3 (CAB3; at1g29910), and RbcL (Tables S1 and S2). Trp dioxidation of Executer 1 and Executer 2 of the chloroplast thylakoid has previously been shown to function as a specific sensor of oxidative stress through reaction with singlet oxygen, triggering retrograde signaling [46]. Our identification of Trp dioxidation modifications on several additional proteins may indicate a broader suite of singlet oxygen sensors than was previously recognized. Curiously, in addition to the six chloroplast proteins, a single non-chloroplast protein, Pathogenesis-related 5 (PR5; at1g75040), also had (di)oxidation found on Trp 37 in the total leaf sample.

3.3 | Predicting electrophoretic mobility with CZE-MS/MS

As an open tubular configuration, CZE outperforms RPLC in the accurate prediction of proteoform separation times based on electrophoretic mobility (μ_{ef}) [47]. The semi-empirical prediction model of protein μ_{ef} has been modified and evaluated for the large-scale CZE-MS/MS-based proteomics, and has been discussed at length previously [29].

Accurate prediction of retention/migration times can assist in correctly identifying proteoforms and corresponding mass shifts. To explore the prediction of proteoform mobility within the context of our total leaf and chloroplast samples, we applied our prediction model to proteoforms without modifications, or with only one acetylation or phosphorylation, using proteoform identifications from the second run of total leaf fraction 6, which has the highest number of proteoforms. While predicted and experimentally-determined μ_{ef} values of unmodified proteoforms aligned excellently, the phosphorylated and single N-terminal/lysine-acetylated proteoforms deviated from expectation, as seen in Figure 3A. Acetylation removes the positive charge on the N-terminus or on the lysine side chain, while phosphorylation adds a single negative charge. After accounting for the (–1) charge reduction of these PTMs, we found that most corrected proteoforms aligned well with the trend line (Figure 3B), and the R^2 increased to 0.92 from 0.89. The R^2 value for non-modified proteoforms alone is 0.91, showing that the modified charge proteoforms match the linear correlation well. The several remaining outliers may represent incomplete unfolding in the 10% acetic acid or incorrect proteoform IDs. The well improved linear correlation between experimental and predicted μ_{ef} of proteoforms after charge correction highlights the value of CZE-MS/MS for confident proteoform identification and accurate characterization.

We further looked specifically at the identified RbcL proteoforms. The predicted and experimental μ_{ef} of all 43 proteoforms of RbcL with mass shifts are shown in Figure S9A. There are five non-modified proteoforms, 22 proteoforms with single acetylation, 13 proteoforms with a –176 Da mass shift, one proteoform with a –45 Da mass shift, one proteoform with a +37.9 Da mass shift, and one proteoforms with a –2 Da mass shift. The linear correlation between experimental and predicted μ_{ef} is poor ($R^2 = 0.69$) due to the PTMs of proteoforms, which relate to the mass shifts. The mass shifts (i.e., –176 Da and –45 Da) are difficult to explain. However, after –1 and –2 charge corrections for RbcL proteoforms with mass shifts, as highlighted in Figure S9, the linear correlation was drastically improved ($R^2 = 0.99$). The data suggest that those mass shifts reduced the positive charges of proteoforms significantly during CZE separation. Considering the positive charge reduction from the –176 Da mass shift, we attributed the mass shift to the loss of Met1 (–131 Da) and Ser2 (–87 Da) amino acid residues plus an N-terminal acetylation on Pro3 (+42 Da). Similarly, we speculated that the –45 Da mass shift was due to the removal of Ser1 residue and N-terminal acetylation on Pro2 residue. Four proteoforms with –2 charge reduction most likely had a combination of multiple PTMs that

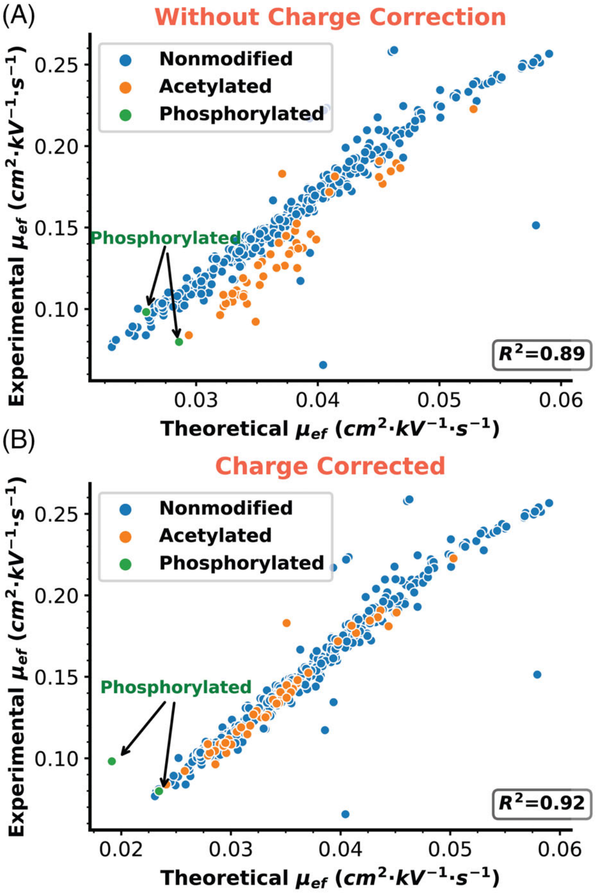


FIGURE 3 Linear correlation between predicted and experimental electrophoretic mobility (μ_{ef}). Comparison is made using μ_{ef} values from the size-exclusion chromatography fraction 6 of the total leaf sample before charge correction (A) and after charge correction (B). Proteoforms with at least one unidentified mass shift were removed. Unmodified proteoforms with no mass shift were labeled in blue, proteoforms with a single acetylation were labeled in orange, and proteoforms with a single phosphorylation are labeled in green. The theoretical electrophoretic mobility is calculated from the number of positive charges (counts of positively charged residue K, R, H, and N-term) and the theoretical mass. The modified predictions are corrected by subtracting 1 from the charge due to acetylation or phosphorylation.

reduced the charge of proteoforms. Three of these four proteoforms had a –175 Da mass shift and one N-terminal acetylation. We expect that those proteoforms have loss of the first two amino acid residues as mentioned before (Met1, –131 Da, and Ser2, –87 Da), and two acetylation sites including the identified N-terminal acetylation. The results further demonstrate that CZE-MS/MS has the capability for accurate characterization of proteoforms with PTMs.

3.4 | Identification of processed protein sequences and truncation pattern

The identification of intact proteoforms offers a prime opportunity to establish the processed N-termini of mature protein sequences, including putative cleavage sites of sub-cellular localization signals such as cTPs or mTPs. Taking advantage of the proteoform data generated from our total leaf and chloroplast samples, we proposed mature N-termini for 343 proteins (proteoforms of an additional 216 proteins were clearly limited to internal fragments of the full protein and hence N-termini could not be determined). Determination was performed manually, relying on frequency of N-termini among proteoforms, relative abundance, and coincident N-terminal acetylation (Table S3). Of the 343 proteins for which mature N-termini were proposed, 253 were consistent with the prediction from TargetP 2.0. including predicted cleavage sites of 65 cTPs, 9 mTPs, 20 SPs, and 16 luTPs. Of those that were inconsistent, most were cTPs (40), or SPs (30). Significantly, the confidence values of TargetP predictions that were inconsistent with the experimental evidence were, on average, almost as high as those that matched with experimental results (88% vs. 94%, respectively). This indicates that the measure of confidence of TargetP 2.0 may not provide a reliable indication of incorrect predictions. Below, we consider results from each class of sub-cellular localization signal separately, highlighting representative proteins in each case.

3.4.1 | Chloroplast transit peptides

Our proteoform identifications are rich in nuclear-encoded chloroplast proteins and allow us to propose cTP cleavage sites for 105 proteins (Table S3). Localization of experimental and predicted cTP sites were consistent in a majority of cases, with 45.4% of proteoforms precisely matched with the predicted cleavage site of cTPs (Figure S10). In one case, that of the cold-regulated protein 15a (AtCor15a; at2g42540), we observed proteoforms starting sequentially from residue 38 to residue 43 in both total leaf and chloroplast samples (Figure S11). While the N-termini of the most abundant proteoform was consistent with the predicted cleavage site (i.e., residue 38), the sequential coverage of five residues likely indicates imprecise cleavage of this cTP.

In 11 cases we suggest a corrected cTP cleavage site (Tables S3 and S4). For example, LHCII-CP26 (at4g10340) was predicted to have a 25% possibility of cleaving between 50K-51A and a 21% possibility between 36V-37A by TargetP 2.0 (Figure S12). In contrast, we identified 11 out of 15 proteoforms starting from residue 38L in total leaf sample (CP26 was not identified in chloroplast), comprising over 90% of total proteoform intensity. Based on these experimental results we suggest the cTP cleavage site of CP26 is, in fact, 37A-38L. Likewise, a cTP cleavage site is predicted for Rubredoxin A (RubA; at1g54500) at residue 59. However, proteoform evidence from both total leaf and chloroplast samples suggest the mature protein sequence begins at residue 55. This is supported by the identification of multiple abundant proteoforms beginning at residue 55 in both total leaf (with N-acetylation) and chloroplast samples, as well as the absence of any proteoform beginning at residue 59.

Several of our corrected cTP cleavage sites are consistent with results from other studies that rely on orthogonal (i.e., non-TDP-based) experimental methods. Based on proteoform identifications we determined cleavage sites for CP29 (at5g01530) and PsbS (at1g44575) at 31T-32A and 53L–54F, respectively (Table S3). These two results are consistent with a previous TDP study [30]. Similarly, Heat Shock Protein of 70 kDa (HSP70; at4g24280) was predicted to start at 93A in TargetP 1.0 and updated to 69T in TargetP2.0. In total leaf, we identified three proteoforms, all of which began at residue 78E, indicating the cTP cleavage site lies at 77N-78E, as previously reported from the TAILS experiment [7].

Proteoform patterns of the two chloroplast ferredoxin isoforms (Ferredoxin-1; at1g10960, and Ferrredoxin-2; at1g60950) represent unusual cases (Figure 4, Table S3). Both proteins hold a predicted cTP cleavage site at residue 52 M-53A (the same residue and position in both isoforms). Consistent with this, proteoforms identified from the chloroplast sample routinely begin at residue 53A. But surprisingly, these proteoforms represent a miniscule proportion of all proteoforms identified from the total leaf sample, even though the chloroplast-localized, and hence processed, proteins should be highly represented in these samples. Instead, a majority of proteoforms, both in prevalence and in relative abundance, start at residue 78D (the same residue and position in both isoforms). This places the starting residue well within the 2Fe-2S Ferredoxin-type iron-sulfur binding domain, indicating that a substantial proportion of the domain is not present within the proteoforms and that they are not functional. As the chloroplast sample is dominated with proteoforms initiating at 53A, consistent with the TargetP 2.0 prediction, we conclude that the true cTP cleavage site of both Ferredoxin isoforms is 53A, however it remains unclear why a majority of both isoforms are processed precisely to the 78D residue specifically in the total leaf sample.

3.4.2 | Lumenal transit peptides

Remarkably, among predicted luTPs, all were consistent with our experimental determinations (Table S3). For example, proteoforms of PsbP-1 predominantly began at 78A, consistent with the predicted luTP cleavage site at 77A-78A (Figure S13). Interestingly, three lower abundant proteoforms of PsbP-1 were identified that start from 34T, which may represent proteins that have had their cTP processed but still await transport into the lumen and subsequent removal of the luTP.

3.4.3 | Signal peptides

Predicted SP cleavage sites were consistent with our experimental determinations for 40.0% of proteins. Remarkably, 14 SP predicted proteins appeared not to have processed sub-cellular localization signals of any sort, instead accumulating proteoforms starting at Met1 or residue 2. In fact, eight of the 14 proteins were found acetylated on the N-terminus. For example, the fragmentation pattern of a single proteoform of Transmembrane Protein (at1g11850) was found with almost full sequence, and initiating from the Met1 residue (Figure 5A).

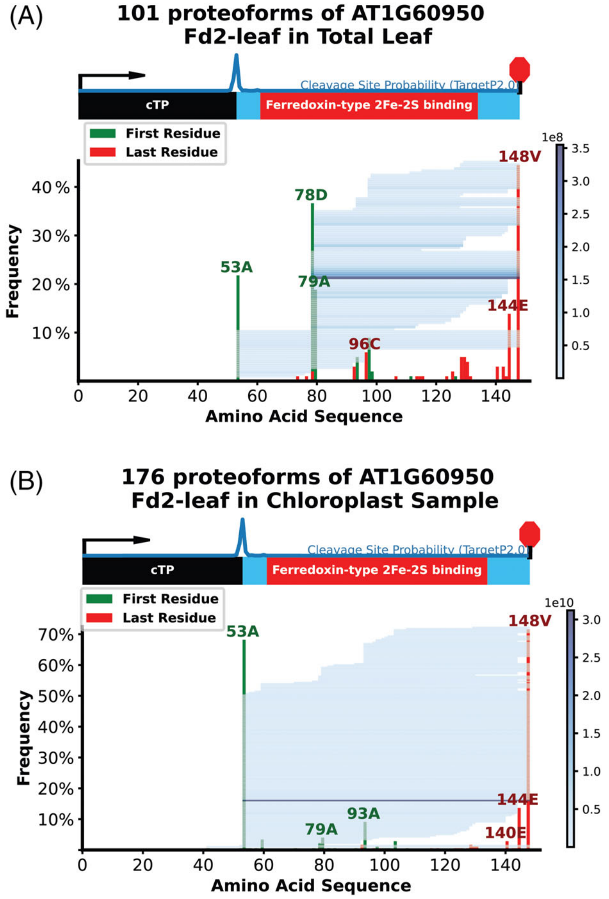


FIGURE 4 Proteoform identifications of Ferredoxin-2 in total leaf (A) and chloroplast (B) samples. Proteoforms identified in the chloroplast sample support cTP cleavage at residues 52–53, while an additional, frequent cleavage site at residues 77–78 is seen specifically in the total leaf sample. A cartoon diagram of the protein sequence domains overlays, and is aligned to, the proteoform identifications indicated as blue horizontal lines shaded according to their estimated abundance (based on ion intensity). The blue trace above the cartoon represents the TargetP 2.0 predicted probability of the cTP cleavage site for each peptide bond. A histogram of green and red bins, overlaying the identified proteoforms, indicates the frequency with which each residue represents either the N-terminal residue (green) or the C-terminal residue (red) among all identified proteoforms. In total leaf sample, the dominant proteoform is 79A-148S, with a mass shift of +23.9 Da. This proteoform is not identified in the chloroplast sample. In the chloroplast sample, the dominant proteoform is 53A-148S with a mass shift of −39.4 Da. This proteoform is identified in the total leaf sample as well, but is only 0.28% the intensity of the highest abundant proteoform in the total leaf sample.

In four cases we could propose a corrected SP cleavage site. The SP of a protein of unknown function (at3g07470) is predicted to cleave at 24A-25I. However, a single proteoform, beginning with residue 14 V and continuing to the final encoded residue, was identified. Similarly, the SP of the TSK-associating protein (at1g52410) is predicted to cleave at 29C-30Q. In contrast, the most abundant proteoforms were found to begin at 21L. Furthermore, only a single proteoform at much lower abundance was found to begin at 30Q.

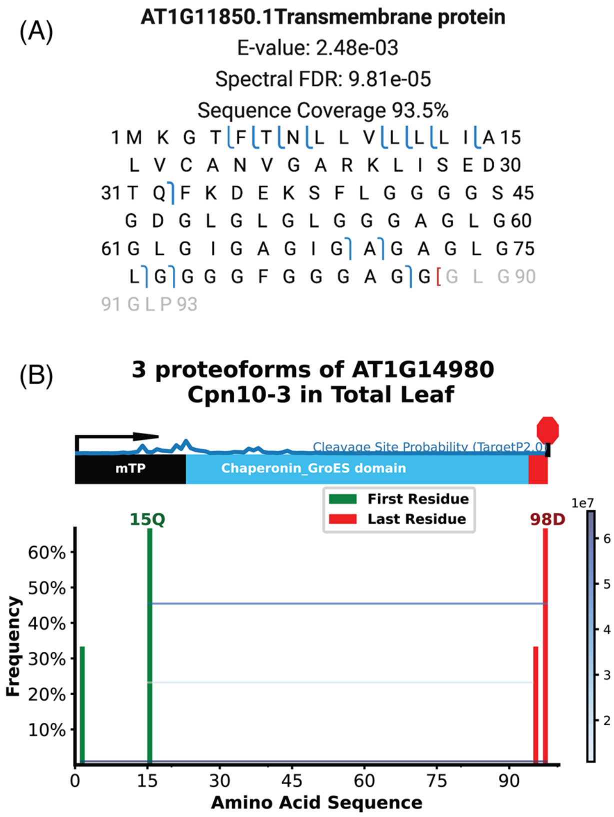


FIGURE 5 Suggested corrections to the predicted sub-cellular localization signals of two proteins. (A) Transmembrane Protein (at1g11850) is predicted to use a Signal Peptide cleaved at 23A-24R. However, the only proteoform identified from this protein reveals a sequence initiating with Met1, indicating the protein does not harbor a cleavable Signal Peptide. The fragmentation pattern from the identified proteoform, along with e-value and FDR are indicated. (B) Chaperonin 10-3 (Cpn10-3; at1g14980) is predicted to use a mitochondrial targeting peptide cleaved at 22K-23T. However, proteoform identifications indicate the protein cleaves at 14V-15Q. A cartoon diagram of the protein sequence domains overlays, and is aligned to, the proteoform identifications indicated as blue horizontal lines shaded according to their estimated abundance (based on ion intensity). The blue trace above the cartoon represents the TargetP 2.0 predicted probability of the cTP cleavage site for each peptide bond. A histogram of green and red bins, overlaying the identified proteoforms, indicates the frequency with which each residue represents either the N-terminal residue (green) or the C-terminal residue (red) among all identified proteoforms.

3.4.4 | Mitochondrial transit peptides

Our experimentally concluded mTP cleavage sites coincided with prediction in 9 out of 18 (50%) proteins (Tables S3 and Table S4). We identified three proteoforms of Cpn10-3 (at1g14980) in the total leaf, which all conflicted with the predicted cleavage site at 22K-23T (Figure 5B). Our results suggest that the mTP cleavage site for this protein is 14V-15Q. While the highest abundant proteoform begins at residue Met1 (and is evidently not imported into the mitochondria), the two remaining proteoforms both initiate at residue 15Q.

Significantly, we identified likely mTP cleavage sites on three other proteins that are not currently predicted to contain mTPs (or any other sub-cellular localization signal) in TargetP 2.0. To conclude a mitochondrial localization for these proteins we relied on the SUBA4 database, which compiles a consensus localization based on disparate experimental and predictive datasets, including MS-based proteomics, fluorescent protein tagging experiments, co-expression data, and 22 computational prediction algorithms [48]. According to SUBA4, the three proteins (Voltage Dependent Ion Channel 3 [VDAC3, at5g15090], Caspase 6 [CASP6, at2g15000], and D-Tyr-tRNA Deacylase family protein [YtDA, at4g18460]) are all strongly expected to localize in the mitochondria. Consistent with this notion, a single proteoform was identified from each of the proteins, each consistent with a cleavage site ranging from residue 33 to 65. Significantly, no proteoform was identified initiating at residue Met1, as would be expected based on the TargetP 2.0 prediction. The identified proteoforms from VDAC3, CASP6, and YtDA began at residues 35S, 65P, and 65D, respectively, directly presenting putative mTP cleavage sites.

3.5 | Residue frequency of cleavage sites

We plotted residue site occupancy around the updated cleavage sites for cTP, luTP, mTP, and SP sequences. WebLogos representing the absolute frequency of residues at each position relative to the cleavage site reveal a weak preference for Ala in the −1 position (relative to the cleavage site) in cTP, SP, and luTP sequences (Figure S14A–D). Conversely, mTPs displayed a somewhat stronger preference for Phe in the −1 position, as well as Ser in the +1 position and a clear preference for Arg in the −3 position. These patterns were drawn out more clearly when presented as an iceLogo which calculates a residue probability in each position by normalizing the absolute frequency of a residue by its frequency throughout all *A. thaliana* protein sequences (Figure S14E–H). Site occupancies of the sub-cellular localization signals are consistent with those reported previously [49, 7].

4 | DISCUSSION

Methodological and technical advancements in the past 10 years have greatly expanded the capabilities of TDP, including more powerful MS/MS search algorithms and increased resolution of mass analyzers [30, 32, 50]. However, these advancements have not, thus far, been applied to large-scale studies in plants. It was the objective of this study to exploit and exhibit the capabilities of TDP in characterizing the proteoforms of *A. thaliana* leaf tissue, with a particular focus on the chloroplast. Using CZE-MS/MS analysis and offline pre-fractionation by SEC, we identified over 4700 unique proteoforms across total leaf and chloroplast samples. This included a substantial number of proteoforms in each sample that contain mass shifts, arising from PTMs or sequence differences relative to the reference protein database. While most mass shifts could not be confidently associated with common

PTMs, 683 and 306 mass shifts in total leaf and chloroplast samples were assigned to common PTMs, such as phosphorylation, oxidation, and acetylation (Figure 2). Some assignments were based on manual curation of the data, comparing mass shift values with monoisotopic masses and literature reports. Identification of these common PTMs was based on the MIscore [32, 51] as well as electrophoretic mobility, providing robust confidence in the identifications.

Among the identified PTMs was Trp oxidation (+15.99 Da) or dioxidation (+ 31.99 Da) on seven proteins, six of which are chloroplast-localized. This somewhat lesser known PTM has recently been found to arise through reaction of singlet oxygen with the Trp side chain, and is a crucial component of singlet oxygen retrograde signaling [52, 46]. Trp oxidation of multiple proteins in ROS-producing mitochondria has similarly been reported to function in retrograde signaling [53]. The functional role (if any) of the Trp (di)oxidation identified in this study is not clear, however the oxidation disrupts the indole ring of the Trp side chain affecting the physico-chemical properties.

A primary goal of this study was to identify mature N-termini of proteins and, by extension, propose cleavage sites of sub-cellular localization signals. In 35 cases, we could confidently propose cleavage sites inconsistent with prediction from the TargetP 2.0 algorithm (Table S3). Alignment of cleavage sites, as determined from our datasets, produced residue frequency plots (i.e., WebLogo and iceLogo) that were largely consistent with those reported in other studies and with other experimental methods [49, 7]. Significantly, we did not observe free cTPs. This is consistent with observations from others that turnover of cTPs occurs rapidly following their cleavage after chloroplast import [54, 7].

Among the interesting observations we report from our study, we found that many proteins accumulated mature sequences with multiple N-termini (Table S4), often varying by a single residue. It is unclear whether this holds functional significance for a given protein, though it seems likely that it would influence stability of at least some proteins. The multiple N-termini may arise due to imprecise cleavage of processing peptidases that recognize sub-cellular localization signals. Alternatively, and not mutually exclusive, the multiple N-termini may represent evidence of additional processing following cleavage of a sub-cellular localization signal. Indeed, Rowland, et al. conclude from their chloroplast N-terminome study that additional, and yet to be identified, peptidases further process the N-terminus of cTP-cleaved proteins to arrive at a limited set of N-terminal residues in mature protein sequences [7].

Importantly, our proteoform identifications are limited to those less than ca. 30 kD. Identification of larger proteoforms is a well-known challenge of TDP studies that is attributed to the negative effects of larger molecular species on the signal/noise ratio [55, 56, 57]. As a molecular species gets larger, its number of possible charge states increases, leading to a dilution of ion intensity across an increasingly larger number of charge state molecules. Identification of larger proteoforms is however possible, and has been accomplished, but generally requires simpler protein mixtures [55]. The development of strategies to handle the mass problem represents one of the greatest opportunities for future improvements in TDP analysis.

ACKNOWLEDGEMENTS

This work was supported by a grant from the National Science Foundation to P.K.L (MCB-2034631). L.S. thanks the support from the National Institutes of Health through Grants R01GM125991, R01GM118470, and R01CA247863, and from the National Science Foundation through Grant DBI-1846913 (CAREER Award). We thank Dr. Daoyang Chen (Merck) for assistance with analysis of the electrophoretic mobility shifts, Dr. Xiaojing Shen (Protein Simple) and Tian Xu (MSU) for advice with chromatography methods, and Dr. Carrie Hiser (MSU) for assistance with sample preparation.

CONFLICT OF INTEREST

Authors have no real or perceived conflict of interest to report.

DATA AVAILABILITY STATEMENT

Proteomics raw datasets have been deposited to ProteomeXchange via the PRIDE partner repository according to MIAPE standards and can be publicly accessed with the identifier PXD034368.

ORCID

Qianjie Wang https://orcid.org/0000-0002-1824-7111
Liangliang Sun https://orcid.org/0000-0001-8939-5042
Peter Knut Lundquist https://orcid.org/0000-0001-8390-8089

REFERENCES

1. Smith, L. M., & Kelleher, N. L. (2013). Proteoform: A single term describing protein complexity. *Nature Methods*, *10*(3), 186–187. https://doi.org/10.1038/nmeth.2369
2. Longoni, P., Douchi, D., Cariti, F., Fucile, G., & Goldschmidt-Clermont, M. (2015). Phosphorylation of the light-harvesting complex II isoform Lhcb2 is central to state transitions. *Plant Physiology*, *169*(4), 2874–2883. https://doi.org/10.1104/pp.15.01498
3. Bouchnak, I., & Van Wijk, K. J. (2019). N-degron pathways in plastids. *Trends in Plant Science*, *24*(10), 917–926. https://doi.org/10.1016/j.tplants.2019.06.013
4. Tasaki, T., Sriram, S. M., Park, K. S., & Kwon, Y. T. (2012). The N-end rule pathway. *Annual Review of Biochemistry*, *81*, 261–289. https://doi.org/10.1146/annurev-biochem-051710-093308
5. Almagro Armenteros, J. J., Salvatore, M., Emanuelsson, O., Winther, O., Von Heijne, G., Elofsson, A., & Nielsen, H. (2019). Detecting sequence signals in targeting peptides using deep learning. *Life Science Alliance*, *2*(5), e201900429. https://doi.org/10.26508/lsa.201900429
6. Christian, R. W., Hewitt, S. L., Nelson, G., Roalson, E. H., & Dhingra, A. (2020). Plastid transit peptides-where do they come from and where do they all belong? Multi-genome and pan-genomic assessment of chloroplast transit peptide evolution. *Peer Journal*, *8*, e9772. https://doi.org/10.7717/peerj.9772
7. Rowland, E., Kim, J., Bhuiyan, N. H., & Van Wijk, K. J. (2015). The arabidopsis chloroplast stromal N-terminome: Complexities of amino-terminal protein maturation and stability. *Plant Physiology*, *169*(3), 1881–1896. https://doi.org/10.1104/pp.15.01214
8. Emanuelsson, O., Brunak, S., Von Heijne, G., & Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols*, *2*(4), 953–971. https://doi.org/10.1038/nprot.2007.131

9. Emanuelsson, O., Nielsen, H., Brunak, S., & Von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology*, *300*(4), 1005–1016. https://doi.org/10.1006/jmbi.2000.3903
10. Emanuelsson, O., & von Heijne, G. (2001). Prediction of organellar targeting signals. *Biochimica Et Biophysica Acta*, *1541*(1–2), 114–119. https://doi.org/10.1016/s0167-4889(01)00145-8
11. Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., & Nakai, K. (2007). WoLF PSORT: Protein localization predictor. *Nucleic Acids Research*, *35*(Web Server issue), W585–W587. https://doi.org/10.1093/nar/gkm259
12. Zybailov, B., Rutschow, H., Friso, G., Rudella, A., Emanuelsson, O., Sun, Q., & Van Wijk, K. J. (2008). Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. *Plos One*, *3*(4), e1994. https://doi.org/10.1371/journal.pone.0001994
13. Gómez, S. M., Nishio, J. N., Faull, K. F., & Whitelegge, J. P. (2002). The chloroplast grana proteome defined by intact mass measurements from liquid chromatography mass spectrometry. *Molecular & Cellular Proteomics*, *1*(1), 46–59. https://doi.org/10.1074/mcp.m100007-mcp200
14. Ryan, C. M., Souda, P., Bassilian, S., Ujwal, R., Zhang, J., Abramson, J., Ping, P., Durazo, A., Bowie, J. U., Hasan, S. S., Baniulis, D., Cramer, W. A., Faull, K. F., & Whitelegge, J. P. (2010). Post-translational modifications of integral membrane proteins resolved by top-down Fourier transform mass spectrometry with collisionally activated dissociation. *Molecular & Cellular Proteomics*, *9*(5), 791–803. https://doi.org/10.1074/mcp.M900516-MCP200
15. Whitelegge, J. P., Zhang, H., Aguilera, R., Taylor, R. M., & Cramer, W. A. (2002). Full subunit coverage liquid chromatography electrospray ionization mass spectrometry (LCMS+) of an oligomeric membrane protein: Cytochrome b(6)f complex from spinach and the cyanobacterium Mastigocladus laminosus. *Molecular & Cellular Proteomics*, *1*(10), 816–827. https://doi.org/10.1074/mcp.m200045-mcp200
16. Granvogl, B., Zoryan, M., Plöscher, M., & Eichacker, L. A. (2008). Localization of 13 one-helix integral membrane proteins in photosystem II subcomplexes. *Analytical Biochemistry*, *383*(2), 279–288. https://doi.org/10.1016/j.ab.2008.08.038
17. Russell, J. D., Scalf, M., Book, A. J., Ladror, D. T., Vierstra, R. D., Smith, L. M., & Coon, J. J. (2013). Characterization and quantification of intact 26S proteasome proteins by real-time measurement of intrinsic fluorescence prior to top-down mass spectrometry. *Plos One*, *8*(3), e58157. https://doi.org/10.1371/journal.pone.0058157
18. Lambertz, J., Liauw, P., Whitelegge, J. P., & Nowaczyk, M. M. (2021). Mass spectrometry analysis of the photosystem II assembly factor Psb27 revealed variations in its lipid modification. *Photosynthesis Research*, https://doi.org/10.1007/s11120-021-00891-7
19. Gómez, S. M., Bil', K. Y., Aguilera, R., Nishio, J. N., Faull, K. F., & Whitelegge, J. P. (2003). Transit peptide cleavage sites of integral thylakoid membrane proteins. *Molecular & Cellular Proteomics*, *2*(10), 1068–1085. https://doi.org/10.1074/mcp.M300062-MCP200
20. Smith, L. M., Thomas, P. M., Shortreed, M. R., Schaffer, L. V., Fellers, R. T., Leduc, R. D., Tucholski, T., Ge, Y., Agar, J. N., Anderson, L. C., Chamot-Rooke, J., Gault, J., Loo, J. A., Paša-Tolić, L., Robinson, C. V., Schlüter, H., Tsybin, Y. O., Vilaseca, M., Vizcaíno, J. A., … Danis, P. O. (2019). A five-level classification system for proteoform identifications. *Nature Methods*, *16*(10), 939–940. https://doi.org/10.1038/s41592-019–0573-x
21. Zabrouskov, V., Giacomelli, L., Van Wijk, K. J., & Mclafferty, F. W. (2003). A new approach for plant proteomics: Characterization of chloroplast proteins of Arabidopsis thaliana by top-down mass spectrometry. *Molecular & Cellular Proteomics*, *2*(12), 1253–1260. https://doi.org/10.1074/mcp.M300069-MCP200
22. Chen, D., Mccool, E. N., Yang, Z., Shen, X., Lubeckyj, R. A., Xu, T., Wang, Q., & Sun, L. (2021). Recent advances (2019-2021) of capillary electrophoresis-mass spectrometry for multilevel proteomics. *Mass Spectrometry Reviews*, mas.21714-mas.21714. https://doi.org/10.1002/mas.21714
23. Shen, X., Yang, Z., Mccool, E. N., Lubeckyj, R. A., Chen, D., & Sun, L. (2019). Capillary zone electrophoresis-mass spectrometry for top-down proteomics. *Trends in Analytical Chemistry*, *120*, 115644–115644. https://doi.org/10.1016/j.trac.2019.115644
24. Sun, L., Zhu, G., Zhang, Z., Mou, S., & Dovichi, N. J. (2015). Third-generation electrokinetically pumped sheath-flow nanospray interface with improved stability and sensitivity for automated capillary zone electrophoresis-mass spectrometry analysis of complex proteome digests. *Journal of Proteome Research*, *14*(5), 2312–2321. https://doi.org/10.1021/acs.jproteome.5b00100
25. Wojcik, R., Dada, O. O., Sadilek, M., & Dovichi, N. J. (2010). Simplified capillary electrophoresis nanospray sheath-flow interface for high efficiency and sensitive peptide analysis. *Rapid Communications in Mass Spectrometry*, *24*(17), 2554–2560. https://doi.org/10.1002/rcm.4672
26. Zhu, G., Sun, L., & Dovichi, N. J. (2016). Thermally-initiated free radical polymerization for reproducible production of stable linear polyacrylamide coated capillaries, and their application to proteomic analysis using capillary zone electrophoresis-mass spectrometry. *Talanta*, *146*, 839–843. https://doi.org/10.1016/j.talanta.2015.06.003
27. Lubeckyj, R. A., Mccool, E. N., Shen, X., Kou, Q., Liu, X., & Sun, L. (2017). Single-shot top-down proteomics with capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry for identification of nearly 600 escherichia coli proteoforms. *Analytical Chemistry*, *89*(22), 12059–12067. https://doi.org/10.1021/acs.analchem.7b02532
28. Mccool, E. N., Lubeckyj, R. A., Shen, X., Chen, D., Kou, Q., Liu, X., & Sun, L. (2018). Deep top-down proteomics using capillary zone electrophoresis-tandem mass spectrometry: Identification of 5700 proteoforms from the *Escherichia coli* proteome. *Analytical Chemistry*, *90*(9), 5529–5533. https://doi.org/10.1021/acs.analchem.8b00693
29. Chen, D., Lubeckyj, R. A., Yang, Z., Mccool, E. N., Shen, X., Wang, Q., Xu, T., & Sun, L. (2020). Predicting electrophoretic mobility of proteoforms for large-scale top-down proteomics. *Analytical Chemistry*, *92*(5), 3503–3507. https://doi.org/10.1021/acs.analchem.9b05578
30. Fellers, R. T., Greer, J. B., Early, B. P., Yu, X., LeDuc, R. D., Kelleher, N. L., & Thomas, P. M. (2015). ProSight lite: Graphical software to analyze top-down mass spectrometry data. *Proteomics*, *15*(7), 1235–1238. https://doi.org/10.1002/pmic.201570050
31. Greer, J. B., Early, B. P., Durbin, K. R., Patrie, S. M., Thomas, P. M., Kelleher, N. L., Leduc, R. D., & Fellers, R. T. (2022). ProSight annotator: Complete control and customization of protein entries in UniProt XML files. *Proteomics*, e2100209. https://doi.org/10.1002/pmic.202100209
32. Kou, Q., Xun, L., & Liu, X. (2016). TopPIC: A software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics*, *32*(22), 3495–3497. https://doi.org/10.1093/bioinformatics/btw398
33. Smith, L. M., Agar, J. N., Chamot-Rooke, J., Danis, P. O., Ge, Y., Loo, J. A., Paša-Tolić, L., Tsybin, Y. O., & Kelleher, N. L. (2021). The human proteoform project: Defining the human proteome. *Science Advances*, *7*(46), eabk0734. https://doi.org/10.1126/sciadv.abk0734
34. Pfitzner, J. (1976). Poiseuille and his law. *Anaesthesia*, *31*(2), 273–275. https://doi.org/10.1111/j.1365–2044.1976.tb11804.x
35. Britz-Mckibbin, P., & Chen, D. D. Y. (2000). Selective focusing of catecholamines and weakly acidic compounds by capillary electrophoresis using a dynamic pH junction. *Analytical Chemistry*, *72*(6), 1242–1252. https://doi.org/10.1021/ac990898e
36. Zhu, G., Sun, L., Yan, X., & Dovichi, N. J. (2014). Bottom-up proteomics of *Escherichia coli* using dynamic pH junction preconcentration and capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry. *Analytical Chemistry*, *86*(13), 6331–6336. https://doi.org/10.1021/ac5004486

37. Lubeckyj, R. A., Basharat, A. R., Shen, X., Liu, X., & Sun, L. (2019). Large-scale qualitative and quantitative top-down proteomics using capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry with nanograms of proteome samples. *Journal of the American Society for Mass Spectrometry*, *30*(8), 1435–1445. https://doi.org/10.1007/s13361-019-02167-w
38. Creasy, D. M., & Cottrell, J. S. (2004). Unimod: Protein modifications for mass spectrometry. *Proteomics*, *4*(6), 1534–1536. https://doi.org/10.1002/pmic.200300744
39. Houtz, R. L., Stults, J. T., Mulligan, R. M., & Tolbert, N. E. (1989). Post-translational modifications in the large subunit of ribulose bisphosphate carboxylase/oxygenase. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(6), 1855–1859. https://doi.org/10.1073/pnas.86.6.1855
40. Houtz, R. L., Magnani, R., Nayak, N. R., & Dirk, L. M. A. (2008). Co- and post-translational modifications in Rubisco: Unanswered questions. *Journal of Experimental Botany*, *59*(7), 1635–1645. https://doi.org/10.1093/jxb/erm360
41. Al-Momani, S., Qi, D., Ren, Z., & Jones, A. R. (2018). Comparative qualitative phosphoproteomics analysis identifies shared phosphorylation motifs and associated biological processes in evolutionary divergent plants. *Journal of Proteomics*, *181*, 152–159. https://doi.org/10.1016/j.jprot.2018.04.011
42. Carlberg, I., Hansson, M., Kieselbach, T., Schröder, W. P., Andersson, B., & Vener, A. V. (2003). A novel plant protein undergoing light-induced phosphorylation and release from the photosynthetic thylakoid membranes. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(2), 757–762. https://doi.org/10.1073/pnas.0235452100
43. Li, Z., Wakao, S., Fischer, B. B., & Niyogi, K. K. (2009). Sensing and responding to excess light. *Annual Review of Plant Biology*, *60*(1), 239–260. https://doi.org/10.1146/annurev.arplant.58.032806.103844
44. Niyogi, K. K. (2000). Safety valves for photosynthesis. *Current Opinion in Plant Biology*, *3*(6), 455–460. https://doi.org/10.1016/s1369-5266(00)00113-8
45. Pinnola, A., & Bassi, R. (2018). Molecular mechanisms involved in plant photoprotection. *Biochemical Society Transactions*, *46*(2), 467–482. https://doi.org/10.1042/BST20170307
46. Dogra, V., Li, M., Singh, S., Li, M., & Kim, C. (2019). Oxidative post-translational modification of EXECUTER1 is required for singlet oxygen sensing in plastids. *Nature Communications*, *10*(1), 2834. https://doi.org/10.1038/s41467-019-10760-6
47. Cifuentes, A., & Poppe, H. (1994). Simulation and optimization of peptide separation by capillary electrophoresis. *Journal of Chromatography A*, *680*(1), 321–340. https://doi.org/10.1016/0021–9673(94)80083-9
48. Heazlewood, J. L., Verboom, R. E., Tonti-Filippini, J., Small, I., & Millar, A. H. (2007). SUBA: The *Arabidopsis* subcellular database. *Nucleic Acids Research*, *35*(Database issue), D213–D218. https://doi.org/10.1093/nar/gkl863
49. Huang, S., Taylor, N. L., Whelan, J., & Millar, A. H. (2009). Refining the definition of plant mitochondrial presequences through analysis of sorting signals, N-terminal modifications, and cleavage motifs. *Plant Physiology*, *150*(3), 1272–1285. https://doi.org/10.1104/pp.109.137885
50. Wu, Z., Roberts, D. S., Melby, J. A., Wenger, K., Wetzel, M., Gu, Y., Ramanathan, S. G., Bayne, E. F., Liu, X., Sun, R., Ong, I. M., Mcilwain, S. J., & Ge, Y. (2020). MASH explorer: A universal software environment for top-down proteomics. *Journal of Proteome Research*, *19*(9), 3867–3876. https://doi.org/10.1021/acs.jproteome.0c00469
51. Kou, Q., Zhu, B., Wu, S., Ansong, C., Tolić, N., Paša-Tolić, L., & Liu, X. (2016). Characterization of proteoforms with unknown post-translational modifications using the MIScore. *Journal of Proteome Research*, *15*(8), 2422–2432. https://doi.org/10.1021/acs.jproteome.5b01098
52. Dogra, V., & Kim, C. (2019). Chloroplast protein homeostasis is coupled with retrograde signaling. *Plant Signaling & Behavior*, *14*(11), 1656037. https://doi.org/10.1080/15592324.2019.1656037
53. Taylor, S. W., Fahy, E., Murray, J., Capaldi, R. A., & Ghosh, S. S. (2003). Oxidative post-translational modification of tryptophan residues in cardiac mitochondrial proteins. *Journal of Biological Chemistry*, *278*(22), 19587–19590. https://doi.org/10.1074/jbc.C300135200
54. Richter, S., & Lamppa, G. K. (1999). Stromal processing peptidase binds transit peptides and initiates their ATP-dependent turnover in chloroplasts. *Journal of Cell Biology*, *147*(1), 33–44. https://doi.org/10.1083/jcb.147.1.33
55. Cai, W., Tucholski, T., Chen, B., Alpert, A. J., Mcilwain, S., Kohmoto, T., Jin, S., & Ge, Y. (2017). Top-down proteomics of large proteins up to 223 kDa enabled by serial size exclusion chromatography strategy. *Analytical Chemistry*, *89*(10), 5467–5475. https://doi.org/10.1021/acs.analchem.7b00380
56. Chen, B., Brown, K. A., Lin, Z., & Ge, Y. (2018). Top-down proteomics: Ready for prime time? *Analytical Chemistry*, *90*(1), 110–127. https://doi.org/10.1021/acs.analchem.7b04747
57. Melby, J. A., Roberts, D. S., Larson, E. J., Brown, K. A., Bayne, E. F., Jin, S., & Ge, Y. (2021). Novel strategies to address the challenges in top-down proteomics. *Journal of the American Society for Mass Spectrometry*, *32*(6), 1278–1294. https://doi.org/10.1021/jasms.1c00099

SUPPORTING INFORMATION

Additional supporting information may be found online https://doi.org/10.1002/pmic.202100377 in the Supporting Information section at the end of the article.

How to cite this article: Wang, Q., Sun, L., & Lundquist, P. K. (2022). Large-scale top-down proteomics of the *Arabidopsis thaliana* leaf and chloroplast proteomes. *Proteomics*, e2100377. https://doi.org/10.1002/pmic.202100377