Arabidopsis proteome and the mass spectral assay library

Huoming Zhang\(^1,6\), Pei Liu\(^2,6\), Tiannan Guo\(^3,6\), Huayan Zhao\(^2\), Dalila Bensaddek\(^1\), Ruedi Aebersold\(^3,4\) & Liming Xiong\(^2,5\)

Arabidopsis is an important model organism and the first plant with its genome completely sequenced. Knowledge from studying this species has either direct or indirect applications for agriculture and human health. Quantitative proteomics by data-independent acquisition mass spectrometry (SWATH/DIA-MS) was recently developed and is considered as a high-throughput, massively parallel targeted approach for accurate proteome quantification. In this approach, a high-quality and comprehensive spectral library is a prerequisite. Here, we generated an expression atlas of 10 organs of Arabidopsis and created a library consisting of 15,514 protein groups, 187,265 unique peptide sequences, and 278,278 precursors. The identified protein groups correspond to ~56.5% of the predicted proteome. Further proteogenomics analysis identified 28 novel proteins. We applied DIA-MS using this library to quantify the effect of abscisic acid on Arabidopsis. We were able to recover 8,793 protein groups of which 1,787 were differentially expressed. MS data are available via ProteomeXchange with identifier PXD012708 and PXD012710 for data-dependent acquisition and PXD014032 for DIA analyses.

Background & Summary

Arabidopsis thaliana is a flowering plant with a short but complex life cycle. It has a relatively small genome size with low repetitive content (10%). These features make it an ideal organism for laboratory research. Knowledge from studying this species has either direct or indirect applications to agriculture and human health. Arabidopsis therefore became the first plant to have its genome completely sequenced and annotated under "The Arabidopsis Genome Initiative 2000"\(^2\), which greatly promoted whole genome sequencing and global transcriptome analysis using next generation sequencing technology\(^3\). Proteins bridge genetic information and phenotypes. However, the protein abundances generally exhibit poor correlation with genetic variations\(^4\), necessitating direct study of proteins under different biological conditions.

Proteomics, defined as the study of all proteins in any given sample, has advanced at a fast rate in the last decade, especially in quantitative proteomics that has been widely used for both discovery and targeted analyses. Three commonly used discovery quantitative proteomics strategies are chemical labeling such as isobaric tags for relative and absolute quantitation (iTRAQ\(^5\)) and tandem mass tags (TMT\(^6\)), metabolic labeling (SILAC\(^7\), \(^15\)N labeling\(^8\) etc.) and label-free\(^9,10\) approaches. These methods are generally high-throughput and provide in-depth coverage suitable for system-wide analyses. However, in order to maximize proteome coverage, it is necessary to incorporate a sample prefractionation step prior to liquid chromatography-mass spectrometry (LC-MS) analysis. In addition to substantially increasing both data acquisition and analysis times, this leads to a reduction in the reproducibility of measurements and the quantitative accuracy especially in label-free experiments of large sample cohorts. Multiplexed analyses such as TMT/iTRAQ reduce the analysis time, but suffer from ratio compressions which in turn impact protein quantification\(^11,12\). On the other hand, the targeted proteomics (S/MRM\(^13\); selected/multiple reaction monitoring; PRM\(^14\); parallel reaction monitoring) provides higher sensitivity and reproducibility but it has limited use for proteome-wide survey.

Recently, a relatively new technique termed SWATH/DIA-MS (SWATH\(^15\): sequential window acquisition of all theoretical mass spectra; DIA: data-independent acquisition) was developed to complement discovery and targeted proteomics. In this approach, the precursors across the mass range of interest (e.g. 400–1200 Da) are...
Arabidopsis except rosette leaves which were grown for 3 weeks. Seeds were harvested when siliques turned yellow or brown, greenhouse room under the same growth conditions as in the growth chamber for an additional 5-week period, open flowers) and siliques (green siliques), seedlings grown for one week on the plate were transferred to soil in a and 20 days, respectively. For preparation of leaves (rosette leaves and cauline leaves), stems, flowers (buds and open flowers) and siliques (green siliques), seedlings grown for one week on the plate were transferred to soil in a greenhouse room under the same growth conditions as in the growth chamber for an additional 5-week period, except rosette leaves which were grown for 3 weeks. Seeds were harvested when siliques turned yellow or brown. For the study of abscisic acid effect on Arabidopsis, rosette leaves were sprayed with or without 100 μM of abscisic acid and harvested at 2 h, 24 h and 72 h post treatment. In each condition, three leaves per plant from three independent plants were harvested and combined into a single sample per condition. All collected materials were immediately frozen in liquid nitrogen and stored at −80 °C until use.

Arabidopsis thaliana root cell suspension culture. Cells isolated from roots of Arabidopsis thaliana were grown in Gamborg’s B5 basal salt mixture (Sigma-Aldrich) with 2,4-dichlorophenoxyacetic acid (2,4-D; 1 mg mL−1) and kinetin (0.05 μg mL−1) in sterile flask as described18. Briefly, cells were grown in a growth chamber (Innova® 43, New Brunswick Scientific Co., NJ) with shaking at 120 rpm, and subcultured every 10 days. Photosynthetic light of the growth chamber was set for 16 h light/8 h dark cycles at 21 °C. The cells were harvested by draining off the media using Stericup® filter unit (Millipore, Billerica, MA), immediately flash frozen in liquid nitrogen and stored at −80 °C until use.

Protein extraction and digestion. All plant materials were ground in liquid nitrogen with a prechilled mortar and a pestle. The fine powder was resuspended with the extraction buffer (50 mM Tris, pH 8, 8 M urea, and 0.5% SDS) supplemented with protease inhibitor (Roche Diagnostics GmbH, Mannheim, Germany), and homogenized with a Dounce homogenizer. In order to extract more proteins, the crude homogenate was further subjected to 30 cyclic high/low pressurization (50 s of 35,000 PSI and 10 s of ambient pressure) using a pressure cycling technology (Barocycle, PressureBioSciences, MA). The extracts were then centrifuged at 10,000 g for 5 min at 4 °C. The proteins in the supernatant were purified using methanol/chloroform precipitation and dried under vacuum. The dried pellets were resuspended into the extraction buffer (50 mM Tris, pH 8, 8 M urea, and 0.5% SDS) and sonicated. The protein content was determined using a microBCA kit (Thermo Scientific). For library generation, approximately 200 μg of proteins were reduced, alkylated and digested with trypsin as described19. The digests were desalted with microcolumns packed with C18 and Poros oligo R3 materials prior to a shallow-gradient Strong Cation Exchange (SCX) fractionation. For protein quantitation of ABA-treated sample, approximately 10 μg of proteins were digested using FASP method20 prior to DIA-MS analysis.

SCX peptide fractionation. The peptides were reconstituted in 90 μL SCX buffer A (10 mM KH2PO4, 25% acetonitrile [ACN], pH 3.0) and loaded into the polySULFOETHYL A column (200 × 4.6 mm, 5 μm, 200 Å) (PolyLC, Columbia, MD) for SCX fractionation on Accela HPLC (Thermo Scientific). An increasing gradient of buffer B (10 mM KH2PO4, 500 mM KCl and 25% ACN, pH 3.0) was applied in a shallow-gradient elution protocol of total 100 min. The gradient consists of 100% buffer A for the initial 5 min, 0%-15% buffer B for 80 min, 30%-100% buffer B for 5 min, 100% buffer B for 5 min and 100% A for 5 min at a flow rate of 1 mL/min. The chromatography was monitored at 214 nm using diode array detector. After pooling of some fractions based on the absorption intensity, a total of 30 fractions were obtained, desalted as described above and dried in the SpeedVac (Thermo Scientific).

MS analysis using TripleTOF 5600+. The dried peptide mixture was redissolved into 0.1% formic acid (FA) and 3% ACN in water supplemented with indexed retention time (iRT) peptide standards according to the manual (Biognosys, Switzerland). They were then analyzed using a TripleTOF 5600 Plus MS (Sciex, USA) coupled with an UltiMate™ 3000 UHPLC (Thermo Scientific). Briefly, approximately 1.5 μg of peptide mixture was injected into a precolumn (Acclain PepMap100, C18, 300 μm × 5 mm, 5 μm particle size) and desalted for 15 min with 3% ACN and 0.1% FA in water at a flow rate of 5 μL/min. The peptides were eluted into an Acclain
MS analysis using Orbitrap Fusion Lumos. In both data-dependent acquisition (DDA) and DIA analysis, an Orbitrap Fusion Lumos mass spectrometer (Thermo Scientific) was coupled with an UltiMate™ 3000 UHPLC (Thermo Scientific). The peptide injection and elution gradient were essentially the same as described above. The peptides were separated and introduced into the Orbitrap MS through an integrated Easy-Spray LC column (50 cm × 75 μm ID, PepMap C18, 2 μm particles, 100 Å pore size, Thermo Scientific) with an electrospray potential of 1.9 kV. The ion source transfer tube temperature was set at 270 °C. The MS parameters included application mode as standard for peptide, default charge state of 3 and the use of EASY-IC as internal mass calibration in both precursor ions (MS1) and fragment ions (MS2).

In DDA mode, a full MS scan (375–1400 m/z range) was acquired in the Orbitrap at a resolution of 120,000 (at 200 m/z) in a profile mode. The cycle time was set to 3 s between master scans, whereas the RF lens was set to 30%. A maximum ion accumulation time was 100 milliseconds and a target value was 4e5. MIPs (monoisotopic peak determination of peptide) was activated. The isolation window for ions was 1.6 m/z. The ions above an intensity threshold of 5 e4 and carrying charges from 2+ to 4+ were selected for collision-induced dissociation. A rolling collision energy option was applied. The maximum cycle time was fixed to 2 s and a maximum accumulation time for individual ions was set for 250 ms. Dynamic exclusion was set to 15 s with a 50 mDa mass tolerance.

Protein identification. Mass spectrometry data (.raw file from Orbitrap Fusion and .wiff file from TripleTOF 5600+) were processed using the Maxquant software (version 1.5.3.30) 15. The TAIR10 proteome database (35,386) and 262 common contaminant sequences were combined and used for the database search. The “revert” protein sequence was chosen in the decoy mode option for all Maxquant searches. Carbamidomethylation at cysteine residues was set as a fixed modification. The variable modifications included oxidation at methionine residues and N-terminal protein acetylation. The enzyme limits were set at full trypsin cleavage with a maximum of two missed cleavages allowed. A positive peptide identification was required to contain a minimum of seven amino acids and a maximum of five modifications. The mass tolerances of the precursor ion of Orbitrap Fusion data were set to 20 and 4.5 ppm for the first and main searches, respectively, whereas they were 0.07 and 0.006 Da for the TripleTOF 5600 data. The mass tolerances of the fragment ions were set 20 ppm and 4 ppm for Orbitrap Fusion and TripleTOF 5600 data, respectively. The false discovery rates (FDRs) of peptide-spectral match (PSM), protein identification and site decoy fraction were all set to 0.01.

To identify novel proteins, the Maxquant results were loaded into Scaffold software (version 4.4, Proteome software Inc., Portland, OR). The unmatched spectra from Maxquant searches were exported. They were then searched against a proteome database constructed using six-frame translation of the TAIR9 genome. The FDRs for PSM and protein were 0.01 to filter all identification. To minimize potentially random matches due to the use of only unmatched spectra data and the newly constructed large proteome database from TAIR9 genome, we extracted the novel identified sequences and merged them with the TAIR10 proteome database. This combined proteome database was considered as “clean database”. All MS raw files were searched against this “clean proteome database” using Maxquant. The search results were imported into Scaffold software for validation and verification. Proteins and peptides were filtered based on the possibilities of 99% and 95% respectively. For novel identification, two minimum spectra were required to match exclusively to the novel proteins. The spectral matches were required to have a minimum of 5 continuous y or b ion series, containing parent ion and match of major (high intensity) fragment ions to sequence. Each protein was matched with the Arabidopsis 11 transcript (DNA) using blast program (tblastn) to extract detailed annotation information in TAIR webpage (https://www.arabidopsis.org) such as genomics locus and gene model type.

MS spectral library generation and DIA data analysis. All the DDA MS data files were loaded into Spectronaut Pulsar X (version 12, Biognosys, Switzerland) for the library generation. The protein database was the combination of TAIR10 proteome sequence and the aforementioned novel identifications. The default settings for database match include: full trypsin cleavage, peptide length of between 7 and 52 amino acids and maximum missed cleavage of 2. Besides, lysine and arginine (KR) were used as special amino acids for decoy generation, and N-terminal methionine was removed during pre-processing of the protein database. Fixed modification was
carbamidomethylation at cysteine and variable modifications were acetylation at protein N-terminal and oxidation at methionine. All FDRs were set as 0.01 for the peptide-spectrum match (PSM), peptide and protein. To be included into the final ion library, fragment ions were required to contain a minimum of three amino acids in length, with a mass range between 300 and 1800 m/z, and a minimum relative intensity of 5%. The best 3–6 fragments per peptide were included in the library. The IRT calibration was required with minimum R-Square of 0.8.

DIA data were analyzed using Spectronaut software against the spectral libraries to identify and quantify peptides and proteins. The Biognoys default settings were applied for identification: excluding duplicate assay; generation decay based on mutated method at 10% of library size; and estimation of FDRs using Q value as 0.01 for both precursors and proteins. The p-value was calculated by kernel-density estimator. Interference correction was activated and a minimum of 3 fragment ions and 2 precursor ions were kept for the quantitation. The area of extracted ion chromatogram (XIC) at MS/MS level were used for quantitation. Peptide (striped sequence) quantity was measured by the mean of 1–3 best precursors, and protein quantity was calculated accordingly by the mean of 1–3 best peptides. Local normalization strategy and q-value sparse selection were used for cross run normalization. As the control and treated plants were grown in a same condition, and leaves were collected from three individual plants and pooled in each condition, the initial abundances of proteins (if without treatment) for all conditions at the same growth stage were considered as identical. Thus, a paired Student’s t-test (one sample, null hypothesis, no change, mean μ = 0) was performed to uncover differential expression between control and ABA treated sample for the same growth stage of plants. The t-test was performed based on the log2 ratios of the peptide intensities of the individual peptides of a protein. The resulting p values were corrected for multiple testing using the q-value approach to control the overall FDR22. Proteins with a fold-change of higher than 1.5 and a q-value of less than 0.01 were considered as differentially expressed proteins. The candidate proteins were submitted to the web-based platform of the Database for Annotation, Visualization and Integrated Discovery (DAVID; http://david.abcc.ncifcrf.gov) for Gene Ontology (GO) enrichment and pathway analysis23.

Data Records
The mass spectrometry DDA proteomics data acquired using Orbitrap Fusion Lumos and TripleTOF 5600 plus have been deposited to the ProteomeXchange Consortium via the PRIDE24 partner repository with the dataset identifier PXD01270825 and PXD01271026 respectively, whereas the mass spectrometry DIA proteomics data acquired using Orbitrap Fusion Lumos have been deposited at the same server with the dataset identifier PXD01403227. The MS spectral libraries in tab delimited text file format can be accessed at Figshare28.

Technical Validation
Experimental design. In SWATH/DIA-MS, the high quality and coverage of an assay library is the key for accurate quantitation of high number of proteins. Arabidopsis is a complex organism in that they have specialized organs at different developmental stages, and its proteome undergoes dynamic changes in different tissues during development.

To build a comprehensive assay library, we collected 10 samples from four different organs of Arabidopsis (Fig. 1a) including leaf, stem, flower and root. We improved the protein and peptide preparation protocols (Fig. 1b): use of cryogenic grinding of tissues, Dounce homogenization and pressure cyclic treatment (PCT technology) sequentially to get better yields in protein amount and species; purifying proteins sample by methanol/ chloroform which helped to remove majority of non-proteins (such as lipids and pigments), and deconvoluting peptides using a cartridge containing both C18 and R3 material to minimize loss of hydrophilic peptides. Besides we fractionated sample extensively using an optimal 100-min shallow-gradient SCX and combined into final 30 fractions for LC-MS analysis by two MS platforms (Fig. 1c).

Protein identification. Using the optimal workflow, we identified a high number of proteins from each organ (Table 1). The newer mass spectrometry platform Orbitrap Fusion generally identified ~30% higher number of proteins compared to the TripleTOF 5600 plus. In total, we identified more than 180,000 unique peptides from ~15,400 distinct protein groups, which were approximately 30% higher compared with two earlier genome-wide proteome analyses of Arabidopsis29,30, and accounted for ~56.5% of the total predicted proteome of Arabidopsis.

Next, we constructed a new proteome database using six-frame translation of TAIR9 genome. The unmatched spectra from earlier searches against TAIR10 proteome database were searched again with the new database to identify potential novel genes or gene models as described30. As a result, we were able to identify a total of 42 previously unannotated proteins. Most novel findings were previously annotated as “transposable element gene”, “novel transcribed region” and “long noncoding RNA” as well as novel alternative translation and splicing. Of these, 28 proteins were not documented even in the latest TAIR11 proteome database (Table 2). The sequences of novel identifications and their matched peptides as well as the matched spectra if the spectral count is smaller than 4 are available on Figshare28. Interestingly, half of these novel proteins (n = 14) were annotated as transposable element genes in TAIR webpage. There were 2 novel proteins annotated as novel transcribed region, 2 as long noncoding RNA, and 1 antisense long RNA. Four proteins (Table 2, No. 20–24) were probably alternatively transcribed and translated proteins since there were other proteins from the same genomic locus. There were also 5 proteins (Table 2, No. 24–28) with slightly different amino acids from the proteins in the predicted TAIR proteome.

14 out of the 42 novel identifications were found in the latest TAIR11 proteome database, confirming the validity of our proteogenomics approach (Table 3). The sequences of these novel identifications and their matched peptides as well as the matched spectra if the spectral count is smaller than 4 are available on Figshare28. These include 2 new isoforms/entries (Table 3, No. 1–2) that were added in TAIR11 in that they share most part of sequences with other isoforms/entries in TAIR10 but at least 1 unique peptide sequence were identified in this study. Six
proteins (Table 3, No. 3–8) were only identified in this study and in TAIR11 but not in TAIR10 proteome database; and that included two earlier annotated as "transposable element gene". One protein (AT5G13590.1) was identified with difference in two amino acids (in TAIR10: T\text{R}GAFLNSNR, and D\text{E}EEPTELNLSLSK; this study, they were: T\text{S}GAFLNSNR and N\text{E}EEPTELNLSLSK). There were also 5 proteins (Table 3, No. 10–14) containing additional sequences that are not present in TAIR10 protein database. Taken together, these data provide direct experimental evidence confirming the new revision of protein sequences.

In the subsequent DIA-MS analysis of ABA-treated leave sample, seven novel proteins (Table 2: No. 1–5, No. 7 and Table 3: No. 3) with their relatively high spectral counts were also identified and quantified. Of these, the protein AtChr1@30100671@30103941 was observed to be down-regulated (fold change, 0.57, q < 0.0027) at 2 h post ABA-treatment, suggesting that this "transposable element gene" protein is not only expressed at relatively high abundance but may be also functionally active.

Fig. 1 Schematic diagram of experimental workflow. (a) Detailed sample information of the 10 Arabidopsis organs used for generation of spectral library; (b) Optimal protein extraction and peptide purification procedures for in-depth coverage of Arabidopsis spectral library; (c) The samples were analyzed using Orbitrap Fusion and TripleTOF mass spectrometer platforms for data dependent acquisition to construct of the spectral library. The DDA data were analyzed following by protein identification by Maxquant and generation of comprehensive library using Spectronaut Pulsar.

Table 1. Number of proteins identified from each organ using either TripleTOF 5600+ or Orbitrap Fusion Lumos.
Table 2. The list of novel proteins identified using proteogenomics approach. These proteins were not present in TAIR proteome database.

| NO. | 6ORF ID | Genomics Locus sequence (DNA) | TAIR annotation | Total Spectrum Count |
|-----|---------|-----------------------------|-----------------|---------------------|
| 1   | AtChr1@30100671@30103941 | AT1G80020.1 | Transposable_element_gene | at1 38 25 22 24 72 51 31 46 79 13 |
| 2   | AtChr3@4967609@4969613 | AT3G14800.2 | Transposable_element_gene | at2 14 2 9 4 28 29 12 12 19 |
| 3   | AtChr2@2859228@2863482 | AT2G69950.1 | Transposable_element_gene | at3 6 5 2 25 21 6 10 1 |
| 4   | AtChr3@1247686@1249978 | AT3G04605.1 | Transposable_element_gene | at4 5 1 3 1 14 11 6 1 23 |
| 5   | AtChr3@22515146@22515965 | AT3G60930.1 | Transposable_element_gene | at5 4 1 1 1 9 8 1 2 5 4 |
| 6   | AtChr5@19854556@19856566 | AT5G49865.1 | Transposable_element_gene | at6 2 2 8 8 3 3 17 |
| 7   | AtChr5@7082385@7083282 | AT5G20880.1 | Transposable_element_gene | at7 3 1 3 2 5 9 2 4 3 1 |
| 8   | AtChr3@22515868@22516741 | AT3G60930.2 | Transposable_element_gene | at8 1 1 3 5 3 1 7 |
| 9   | AtChr1@26344612@26348515 | AT1G69950.1 | Transposable_element_gene | at9 4 5 |
| 10  | AtChr5@15403874@15404801 | AT5G43510.1 | Transposable_element_gene | at10 1 5 |
| 11  | AtChr5@13106362@13104208 | AT5G43510.3 | Transposable_element_gene | at11 2 1 |
| 12  | AtChr5@13107087 | AT5G43510.1 | Transposable_element_gene | at12 3 2 |
| 13  | AtChr5@19854597@19854666 | AT5G49865.1 | Transposable_element_gene | at13 1 2 |
| 14  | AtChr3@18791524@18795901 | AT3G50625.1 | Transposable_element_gene | at14 4 |
| 15  | AtChr4@7689781@7690192@0 | AT4G06020.1 | Novel_transcribed_region | at15 5 9 |
| 16  | AtChr5@10277623@10277968 | AT5G00495.1 | Novel_transcribed_region | at16 4 3 |
| 17  | AtChr1@10428741@10429146 | AT1G29785.1 | Antisense_long_noncoding_rna | at17 2 2 1 3 5 1 1 |
| 18  | AtChr4@9188288@9188534 | AT4G06385.1 | Long_noncoding_rna | at18 5 1 |
| 19  | AtChr3@325541@325595 | AT3G02345.1 | Long_noncoding_rna | at19 3 |
| 20  | AtChr5@26518786@26519188 | AT5G66400.1 | Arabidopsis thaliana drought-induced 8 | at20 8 |
| 21  | AtChr1@1714192@17153152 | AT2G42560.1 | Late embryogenesis abundant 25 | at21 4 |
| 22  | AtChr2@2056425@2056639 | AT2G05580.1 | Glycine-rich protein family | at22 2 |
| 23  | AtChr4@693384@6936978 | AT4G10895.1 | Plant self-incompatibility protein S1 family | at23 0 0 0 0 6 7 0 0 0 0 |
| 24  | AtChr5@734553@734787 | AT5G22650.1 | Arabidopsis histone deacetylase 2 | at24 5 2 1 1 |
| 25  | AtChr1@20399024@203900324 | AT1G54630.1 | Acyl carrier protein 3 | at25 2 |
| 26  | AtChr1@18512902@18512548 | AT1G49980.9 | DNA/RNA polymerases superfamily protein | at26 5 5 6 |
| 27  | AtChr2@15278911@15279358 | AT2G36410.1 | Transcriptional activator (DUF662) | at27 1 2 1 3 1 |
| 28  | AtChr5@21376666@21377527 | AT5G52710.1 | DNA/RNA polymerases superfamily protein | at28 7 |

**Arabidopsis spectral libraries.** In order to quantify proteome dynamics in *Arabidopsis* by DIA-MS, we constructed a combined spectral library as well as platform-specific libraries from individual LC-MS platform (Table 4) using Spectronaut Pulsar. The library from Orbitrap Fusion analysis was comprised of 15,514 protein groups, 187,265 unique peptide sequences, and 278,278 precursors, while the library from the TripleTOF analysis contains 10,915 protein groups, 80,492 peptides, and 118,675 precursors. The combined library was comprised of a similar number of proteins (15,485) and slightly higher number of 284,418 precursors, suggesting that Orbitrap Fusion apparently recovered nearly all proteins and peptides that from TripleTOF 5600 mass spectrometry in this study.

The peptide distribution of the combined library was shown in Fig. 2a. Most peptides ranged from 8 to 15 amino acids in length, consistent with the properties of full tryptic peptides. Approximately 79% of precursors fall in a mass range between 400–850 m/z (Fig. 2b). In SWATH/DIA-MS analysis, most studies acquired over a mass window from 400 to 1200 Da (m/z) to obtain full coverage of peptides. The large scan window compromises the resolution and increases the complexity of mass spectra, leading to fewer peptide identifications. Our data could provide a valuable reference for designing DIA-MS methods with optimal mass windows. Approximately 50% of precursors have a charge of 2 (Fig. 2c), and 15% are cysteine modified peptides (Fig. 2d). The large number of cysteine-containing peptides may serve as a useful resource for redox proteomics quantitation.

**DIA-MS analysis.** To demonstrate the usefulness of our assay libraries, we performed DIA analysis of ABA-treated *Arabidopsis* leaf sample. A total of 8,793 protein groups were quantified (PXD014032) with low number of missing values from replicates (Fig. 3) using Fusion library. The number of the identification represents 56.7% recovery of the *Arabidopsis* library. The median coefficient of variation (CV) for protein quantities in the experiment was below 10% (figure available on Figshare25), indicating high reproducibility and high quantitation accuracy. We also analyzed these data against TripleTOF 5600 library, it gave similar quantitation of proteins and low CVs, but with slightly lower number of proteins (n = 7670) and about 34% fewer peptides. Surprisingly, when we analyzed these data using the combined library, it quantified approximately 8% lower number of both proteins and peptides than those from using Fusion library alone. Nevertheless, these clearly showed the advantages of DIA-analysis for the high-throughput quantitation analysis for the* Arabidopsis*.
Functional analysis of the abscisic acid (ABA)-regulated proteins. The abundance of proteins in ABA-treated sample was directly compared to their controls at the same developmental stage and with the same treatment time (e.g. ABA 2 h versus control 2 h; ABA 24 h versus control 24 h) to eliminate growth effect. Of the

Table 3. The list of novel proteins identified using proteogenomics approach. These proteins/sequences were either not present in TAIR10 or incomplete, but present in the latest TAIR11 proteome database.

| No | 6ORF ID         | Genomics Locus sequence (DNA) | TAIR annotation                                      | Notes               | Total Spectral Count |
|----|-----------------|-------------------------------|-----------------------------------------------------|---------------------|----------------------|
|    |                 |                               |                                                     |                     | at1 at2 at3 at4 at5 at6 at7 at8 at9 at10 |
| 1  | Atchr1@1925697@1925662 | AT1G51850.2                   | Leucine-rich repeat protein kinase family protein    | New isoform/entry in TaIR11 | 2                     |
| 2  | Atchr5@18590218@18590728 | AT3G4830.4                   | Delay of germination 1                               | New isoform/entry in TaIR11 | 3 31                  |
| 3  | Atchr3@174539@1745543 | AT3G05850.1                   | MuDR family transposase                              | New entry in TaIR11   | 7 1 2 1 14 13 15 6 29 1 |
| 4  | Atchr3@174505@1745966 | AT3G05850.1                   | MuDR family transposase                              | New entry in TaIR11   | 1 2 1 2                |
| 5  | Atchr5@2095254@20953303 | AT5G51858.1                   | Transmembrane protein                                | New entry in TaIR11   | 2                     |
| 6  | Atchr3@617445@6174928 | AT3G11804.4                   | MAP kinase 9                                          | New entry in TaIR11   | 1 1 1                  |
| 7  | Atchr4@9187917@9188217 | AT4G16233.1                   | GDSL-like lipase/acylhydrolase superfamily protein   | New entry in TaIR11   | 4 4                    |
| 8  | Atchr3@3252605@3253004 | AT3G10455.1                   | Plant self-incompatibility protein S1 family protein  | New entry in TaIR11   | 3 6                    |
| 9  | Atchr5@4376549@4378589 | AT5G13590.1                   | Unknown protein                                       | Point correction      | 2 1 2 1 1 1           |
| 10 | Atchr4@15093731@15095657 | AT4G30990.3                   | ARM repeat superfamily protein                       | Sequence addition in TaIR11 | 2                     |
| 11 | Atchr4@13558438@13559653 | AT4G27010.2                   | Ribosome 60S biogenesis amino-terminal protein       | Sequence addition in TaIR11 | 2                     |
| 12 | Atchr4@7455224@7455494 | AT4G12610.1                   | Transcription initiation factor IIF subunit alpha Rap74 | Sequence addition in TaIR11 | 1 2 5 4 3 6 4   |
| 13 | Atchr4@1344112@1344550 | AT4G03050.1                   | 2-oxoglutarate-dependent dioxygenase                 | Sequence addition in TaIR11 | 40 19                  |
| 14 | Atchr5@15331919@15332135 | AT5G38360.1                   | Alpha/beta-Hydrolases superfamily protein            | Sequence addition in TaIR11 | 3 1 3 1 1 1           |

Table 4. Arabidopsis spectral libraries constructed from Orbitrap Fusion and TripleTOF 5600+.

| Platform       | Precursors | Modified peptides | Peptides | ProteinPepsiPeptides | Protein Groups | Proteins | Single hits | Fragments | Serch Engine | Created by |
|----------------|------------|-------------------|----------|----------------------|----------------|----------|-------------|-----------|--------------|------------|
| Fusion Lumos   | 278,278    | 219,883           | 187,265  | 120,750              | 15,514         | 19,375   | 1,352       | 1,635,725 | Pulsar       | Spectronaut v12.0 |
| TripleTOF 5600 | 118,675    | 89,149            | 80,492   | 50,366               | 10,915         | 14,420   | 1,352       | 705,063   |              |            |

Fig. 2 Peptide properties in the Arabidopsis spectral library: peptide length distribution (a), precursor mass-to-charge ratio distribution (b), peptide charge (c) and peptide modification (d).
8,793 protein groups, 1,787 were found to be regulated by ABA treatment at least at one of the three measured time points. To gain insights into the ABA-regulated proteome, we used the DAVID functional annotation tools to perform GO enrichment analysis (biological processing) of all ABA-responding proteins. The enriched GO terms were plotted versus their enrichment $p$-values (logarithm transformation of Benjamini-Hochberg corrected $p$-value) of the GO terms biological process (Fig. 4). The enriched biological process (BP) term of down-regulated protein groups were shown in Fig. 4a–c, whereas those upregulated were showed in Fig. 4d–f. There were slightly fewer GO terms enriched at 24 h treatment compared with 2 h and 72 h treatment.

Not surprisingly, the “response to abscisic acid” was enriched in all conditions. The other highly enriched BP terms were “oxidative-reduction process”, “hydrogen peroxide catabolic process” and “response to oxidative stress”. These oxidative-stress related processes were enriched among the up- and down-regulated proteins at all measured time points, indicating ABA treatment induced active reactive oxygen species (ROS) production. Indeed, it has been well documented that ABA can cause oxidative stress in *Arabidopsis*.[31–33] Several metabolic processes including carbohydrate metabolic process, sucrose biosynthetic/metabolic process, chitin catabolic process and macromolecule catabolic process were downregulated at the 2 h post-treatment (Fig. 4a), whereas they were enriched from the upregulation group of proteins (Fig. 4f) at 72 h post-treatment, indicating the ABA treatment initially reduces metabolism followed by gradually increasing the metabolism to the highest level at 72 h post-treatment. The other highly significant enrichments at 2 h included RNA secondary unwinding, translation, rRNA processing and ribosome biogenesis (Fig. 4d), all of which are related to gene transcription and translation, suggesting that protein synthesis was immediately activated in response to ABA stimuli.

**Pathway alteration in response to ABA treatment.** Based on KEGG database, pathway ath03010: Ribosome was the most significant enrichment (with Benjamini Hochberg adjusted $p$-value of 8.41E-14) from the upregulated proteins at 2 h post-treatment. This significance attenuates at both 24 h (a $p$-value of 0.022) and at 72 h post-treatment ($p$-value > 0.05). Other significant enrichments were related to biosynthesis and metabolic pathways. Most of these pathways were down-regulated at 2 h (ath00940: Phenylpropanoid biosynthesis; ath01100: Metabolic pathways; ath01110: Biosynthesis of secondary metabolites) and 24 h (ath00940: Phenylpropanoid biosynthesis; Biosynthesis of secondary metabolites; Biosynthesis of secondary metabolites; ath01100: Metabolic pathways). In contrast, they were enriched from the upregulated proteins at 72 h post-treatment (ath00940: Phenylpropanoid biosynthesis; ath01110: Biosynthesis of secondary metabolites; ath01100: Metabolic pathways). These results are consistent with the GO BP enrichment analysis, suggesting that both transcription and translation were more active upon ABA treatment and they gradually returned to normal after days, whereas the metabolic pathways were suppressed at earlier stages and subsequently became more active.
In this study, we generated the largest set of tissue-specific DDA data using two high resolution of mass spectrometry platforms. This in-depth proteomics spectral information enabled us to identify and validate novel proteins using proteogenomics analysis by a 6-frame translation approach. These data are made available as a resource to the research community. Researchers can use these data to further annotate Arabidopsis genome data and correlate with RNAseq data.

This study provided the first comprehensive Arabidopsis spectral libraries, either instrument-specific or a combined library. In addition to the submission to PRIDE server, all these libraries in tab delimited text file format were shared on Figshare which are ready for input for Spectronaut, Skyline and Peakview software. We used these libraries in DIA-MS and quantified a higher number of proteins in study of Arabidopsis proteome dynamics upon abscisic acid treatment, suggesting that rich information can be obtained in a high-throughput approach. It is worth to note that we have tested DIA data acquired from Orbitrap Fusion against the platform-specific and the combined libraries. All the analyses returned similar quantitation and low CVs, but use of Fusion library quantified highest number of proteins; it is a 8% higher than that from combined library analysis and a 13% higher than that of 5600 library analysis. This suggests that Fusion library should be applied for DIA analysis conducted using Orbitrap instrument. Although the optimal choice of libraries for SWATH data from TripleTOF or other MS platforms remains to be determined, we anticipate use of either of these libraries would be able to achieve in-depth proteome quantitation. Since this library contained approximately 30,000 cysteine-modified peptides, it can be also used for redox study of reversible cysteine modification. In addition, targeted proteomics using SRM and PRM can be designed based on this reference spectral library.

Received: 19 June 2019; Accepted: 27 September 2019;
Published: 22 November 2019

References
1. Jones, A. M. et al. The impact of Arabidopsis on human health: diversifying our portfolio. Cell 133, 939–943 (2008).
2. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815 (2000).
3. Klepikova, A. V., Kasinov, A. S., Gerasimov, E. S., Logacheva, M. D. & Penin, A. A. A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. Plant J. 88, 1058–1070 (2016).
4. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell 165, 535–550 (2016).
5. Unwin, R. D., Pierce, A., Watson, R. B., Sternberg, D. W. & Whetton, A. D. Quantitative proteomic analysis using isobaric protein tags enables rapid comparison of changes in transcript and protein levels in transformed cells. Mol. Cell. Proteomics 4, 924–935 (2005).

Fig. 4 Gene ontology enrichment analysis (biological process: BP) of differential expressed proteins in response to ABA treatment. Blue bar indicates down regulation of biological processes at 2 h (a), 24 h (b), and 72 h post-treatment (c) respectively, whereas the orange bar indicates upregulation of biological process at 2 h (d), 24 h (e), and 72 h post-treatment (f) respectively. Only BP terms with a Benjamini-Hochberg corrected p-value of less than 0.05 were included.
6. Thompson, A. et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal. Chem. 75, 1895–1904 (2003).
7. Ong, S. E. et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol. Cell. Proteomics 1, 376–386 (2002).
8. Oda, Y., Huang, K., Cross, F. R., Cowburn, D. & Chait, B. T. Accurate quantitation of protein expression and site-specific phosphorylation. Proc. Natl. Acad. Sci. U.S.A. 96, 6591–6596 (1999).
9. Washburn, M. P., Wolters, D. & Yates, J. R. 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat. Biotechnol. 19, 242–247 (2001).
10. Chellius, D. & Bondarenko, P. V. Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. J. Proteome Res. 1, 317–323 (2002).
11. Ow, S. Y. et al. iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly". J. Proteome Res. 8, 5347–5355 (2009).
12. Rauniar, N. & Yates, J. R. 3rd. Isobaric labeling-based relative quantification in shotgun proteomics. J. Proteome Res. 13, 5293–5309 (2014).
13. Kuhn, E. et al. Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and 13C-labeled peptide standards. Proteomics 4, 1175–1186 (2004).
14. Gallien, S. et al. Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. Mol. Cell. Proteomics 11, 1709–1723 (2012).
15. Gillet, L. C. et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol. Cell. Proteomics 11, O111 016717 (2012).
16. Guo, T. et al. Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. Nat. Med. 21, 407–413 (2015).
17. Picotti, P. et al. A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. Nature 494, 266–270 (2013).
18. Ordóñez, N. M. et al. Cyclic mononucleotides modulate potassium and calcium flux responses to H2O2 in Arabidopsis roots. FEBS Lett 588, 1008–1015 (2014).
19. Zhang, H., Qian, P. Y. & Ravasi, T. Selective phosphorylation during early macrophage differentiation. Proteomics 15, 3731–3743 (2015).
20. Wisniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. Nat. Methods 6, 359–362 (2009).
21. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372 (2008).
22. Bruderer, R. et al. Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. Mol. Cell. Proteomics 16, 2296–2309 (2017).
23. Dennis, G. J. et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 4, P3 (2003).
24. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res 47, D442–D450 (2019).
25. Comprehensive Arabidopsis proteome spectral library. PRIDE, https://identifiers.org/pride.project:PXD012708 (2019).
26. Arabidopsis proteome spectral library. PRIDE, https://identifiers.org/pride.project:PXD012710 (2019).
27. Tissue specific proteome quantitation and temporal proteome profiles of Arabidopsis upon ABA exposure using DIA-MS. PRIDE, https://identifiers.org/pride.project:PXD014632 (2019).
28. Zhang, H. et al. Arabidopsis proteome and the mass spectral assay library. figshare, https://doi.org/10.6084/m9.figshare.c.4647293 (2019).
29. Baerenfaller, K. et al. Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. Science 320, 938–941 (2008).
30. Castellana, N. E. et al. Discovery and revision of Arabidopsis genes by proteogenomics. Proc. Natl. Acad. Sci. USA. 105, 21034–21038 (2008).
31. Bohmer, M. & Schroeder, J. I. Quantitative transcriptomic analysis of abscisic acid-induced and reactive oxygen species-dependent expression changes and proteomic profiling in Arabidopsis suspension cells. Plant J 67, 105–118 (2011).
32. Ghassemian, M. et al. Abscisic acid-induced modulation of metabolic and redox control pathways in Arabidopsis thaliana. Phytochemistry 69, 2899–2911 (2008).
33. Watkins, J. M., Chapman, J. M. & Muday, G. K. Abscisic Acid-Induced Reactive Oxygen Species Are Modulated by Flavonols to Control Stomata Aperture. Plant Physiol. 175, 1807–1825 (2017).

Acknowledgements
We thank the facilities director of bioscience and analytical core labs, Stine Buechmann-Moeller, for her endorsement and support in this project.

Author contributions
Study design: Z.H., G.T., R.A. and X.L. Experiment: Z.H., L.P. and Z.H.Y. Data analysis: Z.H., L.P. and G.T. Manuscript preparation: Z.H., L.P., D.B. and X.L.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to H. Zhang.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
