Direct data-independent acquisition (direct DIA) enables substantially improved label-free quantitative proteomics in Arabidopsis

Devang Mehta[1], Sabine Scandola[1] and R. Glen Uhrig[1 #]

[1] Department of Biological Sciences, University of Alberta, Edmonton T6G 2E9, Alberta, Canada

[#]Correspondence
Dr. R. Glen Uhrig
ruhrig@ualberta.ca

Keywords
Arabidopsis thaliana, cell culture, proteome, quantitative proteomics, data dependent acquisition, data independent acquisition, mass spectrometry

Funding
This work was funded by the National Science and Engineering Research Council of Canada (NSERC) and the Canadian Foundation for Innovation (CFI).

Abstract

The last decade has seen significant advances in the application of quantitative mass spectrometry-based proteomics technologies to tackle important questions in plant biology. This has included the use of both labelled and label-free quantitative liquid-chromatography mass spectrometry (LC-MS) strategies in model[1,2] and non-model plants[3]. While chemical labelling-based workflows (e.g. iTRAQ and TMT) are generally considered to possess high quantitative accuracy, they nonetheless suffer from ratio distortion and sample interference issues[4,5], while being less cost-effective and offering less throughput than label-free approaches. Consequently, label free quantification (LFQ) has been widely used in comparative quantitative experiments profiling the native[6] and post-translationally modified (PTM-ome)[7,8] proteomes of plants. However, LFQ shotgun proteomics studies in plants have so far, almost universally, used data-dependent acquisition (DDA) for tandem MS (MS/MS) analysis. Here, we systematically compare and benchmark a state-of-the-art DDA LFQ workflow for plants against a new direct data-independent acquisition (direct DIA) method[9]. Our study demonstrates several advantages of direct DIA and establishes it as the method of choice for quantitative proteomics on plant tissue. We also applied direct DIA to perform a quantitative proteomic comparison of dark and light grown Arabidopsis cell cultures, providing a critical resource for future plant interactome studies using this well-established biochemistry platform.

Introduction

In a typical DDA workflow, elution groups of digested peptide ions (precursor ions) are first analysed at the MS^1 level using a high-resolution mass analyser (such as modern Orbitrap devices). Subsequently, selected precursor ions are isolated and fragmented, generating MS^2 spectra that deduce the sequence of the precursor peptide. For each MS1 scan usually around 10-12 MS^2 scans are performed after which the instrument cycles to the next MS^1 scan and the cycle repeats. While this "TopN" approach enables identification of precursors spanning the entire mass range, the fragmentation of semi-stochastically selected precursor ions (generally,

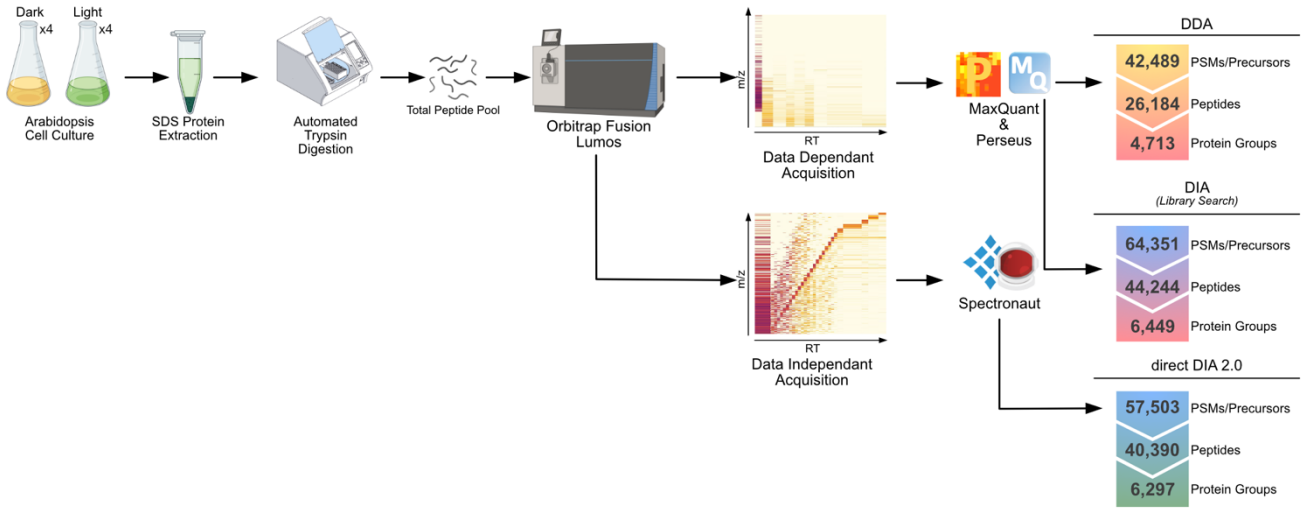

Figure 1: Experimental workflow and summary results.
Total protein was isolated using denaturing extraction from light and dark grown Arabidopsis cells and digested using trypsin. Peptides were desalted and subject to LC-MS/MS using two different acquisition modes. Ion maps showing a single MS1 scan and subsequent MS2 scans are presented to illustrate differences in acquisition schemes. Raw data was analyzed by MaxQuant & Perseus for data-dependent acquisition (DDA) analysis and by Spectronaut for data-independent acquisition (DIA) analysis using spectral libraries created from both acquisitions, and for direct DIA analysis. Counts of FDR-filtered (0.01) peptide spectrum matches (PSMs)/precursors, peptides, and protein groups for each analysis type are shown.

more intense ions) limits the reproducibility of individual DDA runs, results in missing values between replicate runs, and biases quantitation toward more abundant peptides[10]. These limitations are also true for faster, hybrid Orbitrap devices which are still only able to use 1% of the full ion current for mass analysis, hence missing out on lower abundant species[11]. This is particularly disadvantageous for label-free workflows and samples with a high protein dynamic range, such as human plasma and photosynthetic tissue.

In order to address these limitations, several data-independent acquisition (DIA) workflows have been pioneered, famously exemplified by Sequential Window Acquisition of All Theoretical Mass Spectra (SWATH-MS)[12,13]. In DIA workflows, specific, often overlapping, m/z windows spanning a defined mass range are used to sub-select groups of precursors for fragmentation and MS^2 analysis. As a result, complete fragmentation of all precursors in that window follows MS^1 scans resulting in a more reproducible and complete analysis. A major disadvantage of DIA workflows however, is that each MS^2 scan contains multiplexed spectra from several precursor ions making accurate identification of peptides difficult. Traditionally, this has been addressed through the use of global or project-specific spectral-libraries obtained from a fractionated, high-resolution DDA survey of all samples—adding to experimental labour and instrumentation analysis

time. More recently, alternative approaches have been developed that avoid the use of spectral libraries and instead use "pseudo-spectra" derived from DIA runs that are then searched in a spectrum-centric approach analogous to conventional DDA searches[14–16]. Improvements in such library-free DIA approaches have included the incorporation of high precision indexed Retention Time (iRT) prediction[17] and the use of deep-learning approaches[18–20]. Direct DIA (an implementation of a library-free DIA method; Biognosys AG) and a hybrid (direct DIA in combination with library-based DIA) approach has been recently used to quantify more than 10,000 proteins in human tissue[21] and reproducibly identify >10,000 phosphosites across hundreds of human retinal pigment epithelial-1 cell line samples[9].

Direct DIA combines the advantages of DIA for reproducible quantification of proteins in complex mixtures with high dynamic range, with the ease of use of earlier DDA methodologies. Hence, a systematic comparison of these two technologies for LFQ proteomics is essential to define best practice in plant proteomics. In order to execute this analysis, we compared the proteomes of light- and dark-grown Arabidopsis suspension cells. Arabidopsis suspension cells are a long-established platform for plant biochemistry and have recently seen a resurgence in popularity due to their utility in facilitating protein interactomic experimentation using technologies such as tandem affinity purification-mass spectrometry [22–26], nucleic acid crosslinking[27], and proximity labelling (e.g. TurboID)[28]. Despite this, no existing resource profiling the basal differences in proteomes of Arabidopsis cells grown in light or dark exists—a fundamental requirement to determine the choice of growth conditions for protein interactomics experiments and targeted proteomic assays in this system.

Results & Discussion

We performed total protein extraction under denaturing conditions from Arabidopsis (cv. Ler) suspension cells grown for five days in either constant light or dark. Trypsin digestion of the extracted proteome was performed using an automated sample preparation protocol and 1ug of digested peptide was analysed using an Orbitrap Fusion Lumos mass spectrometer operated in either DDA or DIA acquisition mode over 120-minute gradients. A total of 8 injections (4 light & 4 dark) per analysis type were carried out. DDA data processing was performed using MaxQuant, while DIA data processing was performed using Spectronaut v14 (Biognosys AG.). For DIA analysis, both hybrid (library + direct DIA) and direct DIA analysis was performed. The hybrid analysis was performed by first creating a spectral library from DDA raw files using the Pulsar search engine implemented in Spectronaut, followed by a peptide-centric DIA analysis with DIA raw output files. Direct DIA was performed directly on raw DIA files as implemented in Spectronaut. The entire workflow is depicted in **Figure 1**. Both hybrid DIA and direct DIA

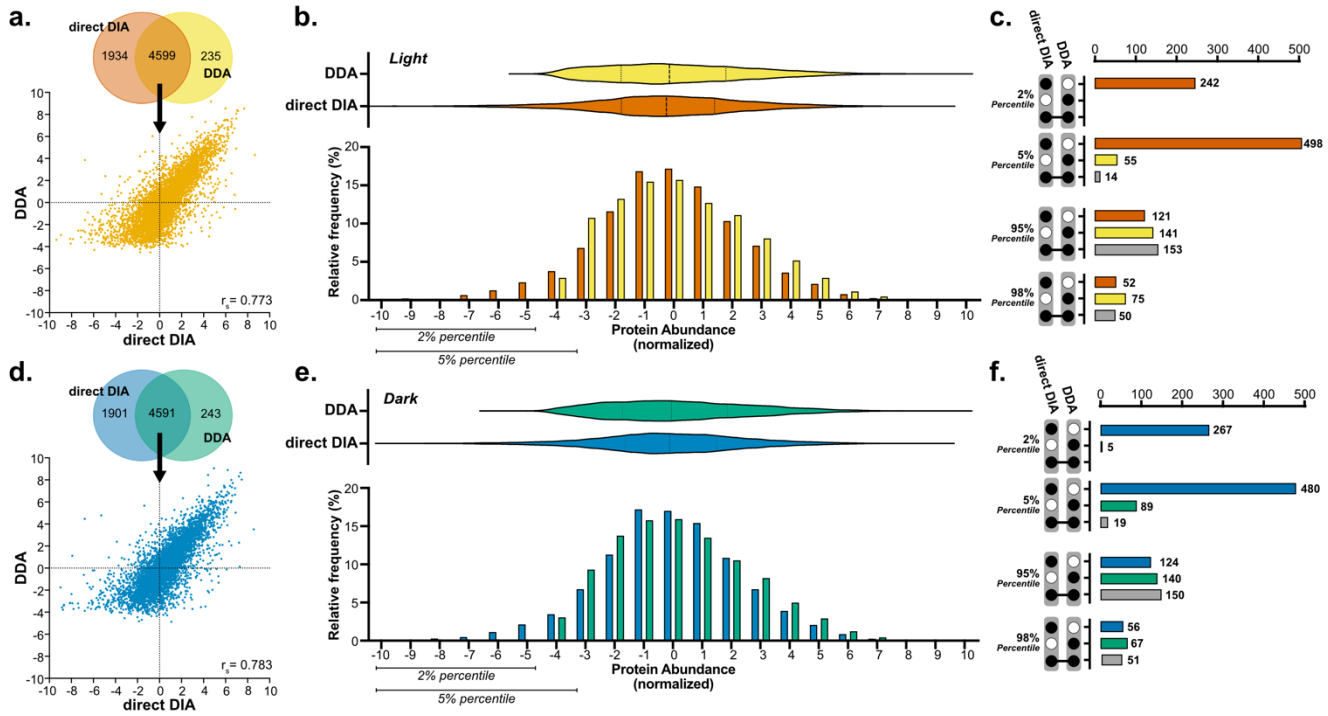

Figure 2: Comparison of protein quantification results using DDA and direct DIA analysis for (a.-c.) light grown and (d.-f.) dark grown Arabidopsis cells.

(a.) & **(d.)** Venn diagram of protein groups quantified with direct DIA and DDA and scatter plot of protein groups quantified by both methods. r_s: Spearman's correlation coefficient. **(b.)** & **(e.)** Frequency distribution of normalized protein abundances for DDA and direct DIA analysis and corresponding violin plots with median and quartile lines marked. **(c.)** & **(f.)** Upset plots depicting intersections in protein groups quantified by DDA and direct DIA at either extreme of the abundance distribution.

analysis greatly outperformed DDA analysis with an average of 65,351; 57,503; and 42,489 peptide-spectrum matches (precursors) quantified across all 8 samples for each analysis, respectively. Hybrid DIA and direct DIA also displayed similar gains over DDA in terms of quantified peptides and protein groups (**Figure 1**). While hybrid DIA analysis performed marginally better than direct DIA, further analysis was performed with the results of only direct DIA and DDA analyses in order to compare methods that use an equivalent number of MS raw input files, comparable instrumentation time and relatively comparable data analysis workflows.

Next, we undertook a series of data analyses to compare the completeness, quality, and distribution of protein group-level quantification of the DDA and direct DIA analyses. In order to compare quantification results across the different analysis types, raw intensity values for each sample were $\log_2$ transformed, median-normalized (per sample), and then averaged for each condition to produce a normalized protein abundance value. For DDA analysis, the number of proteins quantified was determined by first filtering for proteins with valid quantification values in at least 3 of 4 replicates in

either condition (light or dark) and then imputing missing values using MaxQuant with standard parameters[29,30]. For direct DIA analysis, quantified proteins were defined as those passing standard Q-value filtering in Spectronaut. In total, DDA analysis resulted in the quantification of 4,837 proteins (both conditions) and DIA analysis quantified 6,526 proteins (light) and 6,454 proteins (dark) (**Supplementary Tables 1-3**). Upon comparing the quantified proteins between both methods, we found that 4,599 proteins were quantified by both techniques, 1,934 were quantified only by direct DIA and 235 proteins were exclusively DDA-quantified, for light-grown cells (**Figure 2a**). A correlation plot of normalized quantification values for the 4,599 common proteins showed a moderate correlation between DDA and direct DIA quantification (Spearman's R = 0.773) (**Figure 2a**). Examining the frequency distribution of proteins quantified in light-grown cells, by both methods, revealed that the DDA results were substantially skewed towards higher abundant proteins compared to direct DIA (**Figure 2b**). In order to investigate the overlap of quantified proteins between direct DIA and DDA at extreme protein abundances, we sub-selected the 2%, 5%, 95% and 98% percentile of the combined quantification distribution and constructed UpSet plots[31] for these datasets. This analysis revealed that direct DIA quantifies hundreds of more proteins at the lower extremes but is only marginally less effective than DDA at the upper extremes of the protein abundance distribution (**Figure 2c**). These results were similarly replicated for dark-grown cells, suggesting that this is a universal feature of the two acquisition methods, irrespective of sample treatment or type (**Figure 2 d-f**). In order to assess if this difference in quantification ability is specific to plant cells (that have a high dynamic range of protein levels), we further analysed a commercial HeLa cell digest standard using the same mass spectrometry and chromatography settings, with quadruplicate injections per analysis type. Analysing the HeLa quantification results (**Supplementary Tables 4 & 5**) showed a similarly uniform quantification across a wide range by direct DIA and a slightly better, but still skewed, performance by DDA compared to Arabidopsis cells (**Figure S1 a & b**). Comparing the quantification values for HeLa proteins acquired by direct DIA and DDA showed a stronger correlation than Arabidopsis (Spearman's R=0.886). Indeed, correlations between quantification values for lower abundant proteins (defined here as proteins below the median quant value), were much lower than for the overall dataset in both species, and yet slightly stronger in the case of HeLa proteins (**Figure S1 c-e**).

In order to deduce the underlying factors limiting the ability of DDA to quantify low abundant proteins, especially in Arabidopsis cells, we next investigated quantification distributions for both DDA and direct DIA derived data after various data-filtering steps (**Figure S2**; **Supplementary Tables 6-15**). We found that DDA was indeed able to identify a similar number of proteins as direct DIA for both Arabidopsis cells and HeLa digests.

Predictably these numbers dropped dramatically upon filtering proteins for only those with valid quantification values across 3 of 4 replicates, with only mild gains realized due to imputation of missing values. In contrast, even upon filtering for valid values across 4 of 4 replicates, direct DIA resulted in the quantification of more than 5,400 proteins compared to 3,600 complete quantifications for DDA. Strikingly, quantification distributions remained unchanged regardless of various types of data-filtering for direct DIA but were greatly skewed towards high abundance upon filtering for valid values in 3 of 4 replicates in DDA outputs (**Figure S2**). This suggests that the poor quantification of low abundant proteins is related to the presence of missing values in DDA analysis.

This hypothesis was reinforced when we distributed the protein quantification data for direct DIA and DDA based on the number of replicates with valid quantification values for each protein (**Figure S3**). Here we found that the overwhelming majority (>95%) of proteins quantified by direct DIA had valid values in at least 3 of 4 biological replicates for Arabidopsis cells grown in the light or dark (**Figure S3 a & b**). Unsurprisingly, in the HeLa digest, more than 98% of direct DIA quantified proteins were accurately quantified in 4 of 4 technical injections. In contrast, only 68% and 74% of proteins were accurately quantified by DDA in 4 of 4 replicates of light and dark grown Arabidopsis cells, respectively. In fact, the distribution of protein quantification was bimodal, with as many as 17% of proteins accurately quantified in only 1 of 4 replicates by DDA in light-grown cells (10.9% in dark-grown cells). Nearly a quarter of proteins were accurately quantified by DDA in only 1 of 4 technical injections of the same HeLa cell digest, suggesting an inherent disadvantage in reproducibility across replicate runs. In contrast, using direct DIA, no proteins were quantified in only 1 of 4 technical injections of the same HeLa digest and 99.4% were quantified in 4 of 4 technical replicates. When these distributions were further plotted against the normalized protein quantification values, it became clear that proteins found in a lower number of replicates trended lower in abundance in DDA, while this trend did not hold true for direct DIA (**Figure S3 d-i**). In the case of direct DIA, a greater number of quantified proteins were found in 1 of 4 or 2 of 4 biological replicates in Arabidopsis cells compared the technical replicates of HeLa digests. This suggests that the inconsistent quantification of some low abundant proteins using direct DIA is a reflection of real biological variance rather than a methodological artefact. This is contrary to DDA where similar proportions of low abundant proteins were inconsistently identified across biological replicates of Arabidopsis cells and technical replicates of HeLa digests. Overall, this further reinforces that DDA acquisition results in inconsistent quantification between injections, and that this may in fact obscure real biological variance between samples; especially with regards to lower abundant proteins.

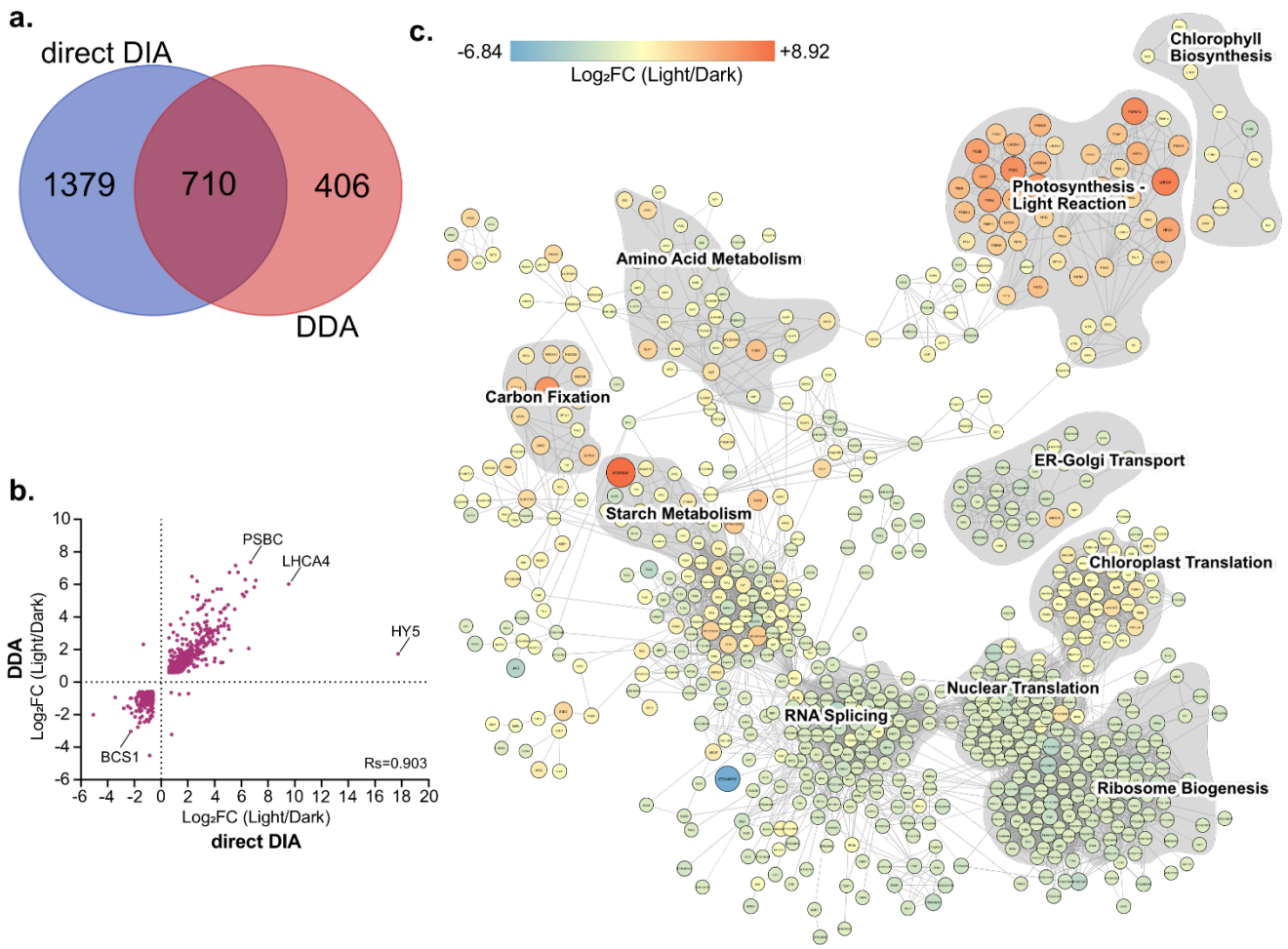

Figure 3: Differential protein abundance analysis for light- and dark-grown Arabidopsis cells.
(a.) Venn diagram of protein groups with significantly changing protein abundances (q<0.01; Abs Log_2FC>1.5) as measured by direct DIA and DDA. **(b.)** Scatter plot of significant changes in protein abundance changes based on DDA and direct DIA analysis with selected proteins labeled. **(c.)** Association network of significantly changing proteins detected with either direct DIA or DDA analysis. Network was constructed based on StringDB database and experiment datasets with a probability cut-off of 0.8. Only nodes with >3 edges are depicted. Clusters were manually annotated based on GO-terms and KEGG/Reactome pathway membership. Node sizes and color are scaled based on the average Log_2FC (Light/Dark) from DDA and DIA analysis.
PSBC: PHOTOSYSTEM II REACTION CENTER PROTEIN C; LHCA4: LIGHT-HARVESTING CHLOROPHYLL-PROTEIN COMPLEX I SUBUNIT A4; HY5: ELONGATED HYPOCOTYL 5; BCS1: CYTOCHROME BC1 SYNTHESIS

Having systematically investigated the advantages and limitations of direct DIA and DDA acquisition for LFQ proteomics, we next performed a differential abundance analysis comparing the proteomes of light- and dark-grown cell cultures quantified with both techniques. We found 2,089 proteins changing significantly in their abundance (Absolute Log2FC > 0.58; q-value <0.05) in our direct DIA analysis and 1,116 proteins changing significantly (Absolute Log2FC > 0.58; q-value <0.05) in DDA analysis. Of these, 710 proteins were found to change significantly in both analyses

(**Figure 3a**). The Log_2 Fold-Change values of these 710 proteins were found to correlate to a high degree between the two analyses (Spearman's R=0.9003), with proteins that were up-regulated in light- vs. dark-grown cells in direct DIA analysis also up-regulated in DDA, and vice-versa (**Figure 3b**). This complete dataset of 2,495 proteins changing significantly in abundance in light- vs dark-grown Arabidopsis cells is a valuable resource for future biochemical studies aiming to use these cell culture systems for protein interactomics experiments and other targeted proteomics analyses (**Supplementary Table 16**). We also created a functional association network of these proteins by probing previously characterized databases and experiments compiled by StringDB[32]. This network validates our analysis, showing that clusters of proteins involved in photosynthesis, carbon-fixation, starch metabolism and amino-acid metabolism have increased abundance in light- vs. dark-grown cells, as expected (**Figure 4c**). Interestingly, clusters representing RNA splicing, ER-Golgi transport, ribosome biogenesis, and nuclear translation are all downregulated, while chloroplast translation is upregulated, in light- vs. dark-grown cells (**Figure 3c**).

Conclusions

Until recently, DDA proteomics (using both label-based and label-free approaches) was clearly the method of choice for functional genomics studies in plants, due to the disadvantages of conventional DIA analysis, such as the requirement for project-specific spectral libraries. Here, we conclusively demonstrate that the newly developed direct DIA proteomics approach is a vastly superior technique for plant proteomics as compared to currently used DDA methodologies. The advantages offered by direct DIA include a greater number of protein identifications and more robust protein quantification than DDA, with no change in instrumentation or increase in instrument analysis time. Our DDA results, even using an advanced Tribrid Orbitrap-linear ion trap device, show that DDA acquisition is particularly inconsistent in its quantification of low-abundant proteins across samples. Similar results have been reported when comparing the abilities of direct DIA and DDA to profile the phosphoproteome (a protein fraction with high dynamic range) of human tissue and cells [9]. Our finding that more than 20% of identified proteins in a DDA experiment are found in only 1 of 4 replicate injections of the same digest, and that these poorly quantified proteins tend to reside in the lower quartile of protein abundance, suggests an inherent drawback in DDA that likely plagues previous studies using this approach, in both label-free and label-based incarnations. Further, the greater proportion of missing values in light- vs. dark-grown cells with both DDA and DIA analysis suggests that the effect of ion suppression is greater in photosynthetically active tissue (**Figure S3 a & b**).

The direct DIA acquisition method is compatible with a wide range of modern mass spectrometers, including older Orbitrap (e.g. QExactive Orbitrap mass spectrometers; ThermoScientific) and Triple TOF devices (Sciex). It has been suggested that the limitations of DDA stem from the limited ion capacity of the C-traps in Orbitrap mass-spectrometers[10]. As a result alternate acquisition schemes have been developed (e.g. BoxCar acquisition) that, inspired by DIA methods, segment precursor ions in order to fragment ions across a larger dynamic range[11,33,34]. In fact, a new method called diaPASEF combining precursor ion separation and DIA acquisition using a trapped ion mobility spectrometer (timsTOF; Bruker) has further developed these ideas to combine the high reproducibility of DIA with reduced MS^2 spectral complexity[35]. However, it will take several years for these, and other methods currently in development, to become standardized and applicable on a wide range of instruments and sample types. The various data analyses undertaken in our plant proteomic study provide a useful template for benchmarking these future quantitative mass-spectrometry proteomics technologies from an end-user perspective. In the meantime, our results argue persuasively for the widespread adoption of direct DIA for quantitative LFQ proteomics in plants. It should be noted that while we utilized proprietary software for direct DIA analysis (Spectronaut v.14, Biognosys AG), multiple free open-source alternatives exist[14,15,19,20] and proprietary software are often available to scientists via professional mass-spectrometry facilities. The demonstrated benefits in reproducibility and dynamic range of direct DIA could be especially powerful for plant biology studies such as the proteomic analysis of multiple treatments (e.g. plant nutrition or herbicide studies), genotypes (e.g. breeding and selection trials), or timepoints (e.g. chronobiology studies).

Methods

Arabidopsis cell culture

Heterotrophic *Arabidopsis thaliana*, cv. Ler suspension cells were obtained from the Arabidopsis Biological Resource Center (ABRC) and maintained in standard Murashige-Skoog media basal salt mixture (M524; PhytoTech Laboratories) at 21 °C as previously described[36] under constant light (100 µmol $m^{-2}s^{-1}$) or constant dark. For the generation of experimental samples, 10 mL aliquots of each cell suspension (7 days old) were used to inoculate 8 separate 500 mL flasks that each contained 100 mL of fresh media. Experimental samples were grown for an additional 5 days prior to harvesting. Cells were harvested by vacuum filtration and stored at −80 °C.

Sample Preparation, Digestion and Phosphopeptide Enrichment

Quick-frozen cells were ground to a fine powder under liquid N_2 using a mortar and pestle. Ground samples were aliquoted into 400 mg fractions. Aliquoted samples were then extracted at a 1:2 (w/v) ratio with a solution of

50 mM HEPES-KOH pH 8.0, 50 mM NaCl, and 4% (w/v) SDS. Samples were then vortexed and placed in a 95°C tabletop shaking incubator (Eppendorf) at 1100 RPM for 15 mins, followed by an additional 15 mins shaking at room temperature. All samples were then spun at 20,000 x g for 5 min to clarify extractions, with the supernatant retained in fresh 1.5 mL Eppendorf tubes. Sample protein concentrations were measured by bicinchoninic acid (BCA) assay (23225; ThermoScientific). Samples were then reduced with 10 mM dithiothreitol (DTT) at 95°C for 5 mins, cooled, then alkylated with 30 mM iodoacetamide (IA) for 30 min in the dark without shaking at room temperature. Subsequently, 10 mM DTT was added to each sample, followed by a quick vortex, and incubation for 10 min at room temperature without shaking.

Total proteome peptide pools were generated using a KingFisher Duo (ThermoScientific) automated sample preparation device as outlined by Leutert et al. (2019)[37] without deviation. Sample digestion was performed using sequencing grade trypsin (V5113; Promega), with generated peptide pools quantified by Nanodrop, acidified with formic acid to a final concentration of 5% (v/v) and then dried by speedbv. Peptides were then dissolved in 3% ACN/0.1% TFA, desalted using ZipTip C18 pipette tips (ZTC18S960; Millipore) as previously described[7], then dried and dissolved in 3.0% ACN/0.1% FA prior to MS analysis.

HeLa proteome analysis was carried out using a HeLa Protein Digest Standard (88329; Pierce). Four replicate injections of this digest per analysis type were carried out with the same methods as for Arabidopsis cell samples.

Nanoflow LC-MS/MS analysis

Peptide samples were analyzed using a Fusion Lumos Tribrid Orbitrap mass spectrometer (ThermoScientific) in data dependent acquisition (DDA) and data independent acquisition (DIA) modes. Dissolved peptides were injected using an Easy-nLC 1200 system (LC140; ThermoScientific) and separated on a 50 cm Easy-Spray PepMap C18 Column (ES803A; ThermoScientific). The column was equilibrated with 100% solvent A (0.1% formic acid (FA) in water). Common MS settings between DDA and DIA runs included a spray voltage of 2.2 kV, funnel RF level of 40 and heated capillary at 300°C. All data were acquired in profile mode using positive polarity with peptide match off and isotope exclusion selected. All gradients were run at 300 nL/min with analytical column temperature set to 50°C. DDA acquisition - Peptide were eluted with a solvent B gradient (0.1% (v/v) FA in 80% (v/v) ACN): 4% - 41% B (0 – 120 min); 41% - 98% B (120-125 min).

DDA acquisition was performed using the Universal Method (ThermoScientific). Full scan MS^1 spectra (350 - 2000 m/z) were acquired with a resolution of 120,000 at 200m/z with a normalized AGC Target of 125% and a maximum injection time of 50 ms. DDA MS^2 were acquired in the linear ion trap using quadrupole isolation in a window of 2.5 m/z. Selected

339 ions were HCD fragmented with 35% fragmentation energy, with the ion
340 trap run in rapid scan mode with an AGC target of 200% and a maximum
341 injection time of 100 ms. Precursor ions with a charge state of +2 - +7 and a
342 signal intensity of at least 5.0e3 were selected for fragmentation. All
343 precursor signals selected for MS/MS were dynamically excluded for 30s.

344 DIA acquisition - included peptides being eluted using a sigmoidal solvent B
345 gradient of 0.1% (v/v) FA in 80% (v/v) ACN from 4% - 41% B (0 - 107 min;
346 **Supplementary Table 17**). DIA acquisition was performed as per Bekker-
347 Jensen et al. (2020)[9] and Biognosys AG. Full scan MS^1 spectra (350 - 1400
348 m/z) were acquired with a resolution of 120,000 at 200 m/z with a
349 normalized AGC Target of 250% and a maximum injection time of 45 ms.
350 ACG target value for fragment spectra was set to 2000%. Twenty-eight 38.5
351 m/z windows were used with an overlap of 1 m/z. Resolution was set to
352 30,000 using a dynamic maximum injection time and a minimum number
353 of desired points across each peak set to 6.

354 ## Raw data processing

355 DDA files were processed using MaxQuant software version 1.6.14[29,30].
356 MS/MS spectra were searched with the Andromeda search engine against a
357 custom made decoyed (reversed) version of the Arabidopsis protein
358 database from Araport 11[38] concatenated with a collection of 261 known
359 mass spectrometry contaminants. Trypsin specificity was set to two missed
360 cleavage and a protein and PSM false discovery rate of 1%; respectively.
361 Minimal peptide length was set to seven and match between runs option
362 enabled. Fixed modifications included carbamidomethylation of cysteine
363 residues, while variable modifications included methionine oxidation.

364 DIA files were processed with the Spectronaut direct DIA experimental
365 analysis workflow using default settings without N-acetyl variable
366 modification enabled. Trypsin specificity was set to two missed cleavages
367 and a protein and PSM false discovery rate of 1%; respectively. Data filtering
368 was set to Q-value and normalization set to global normalization.

369 For hybrid library- and library-free DIA analysis, DDA raw files were first
370 searched with the Pulsar search engine implemented in Spectronaut 14 to
371 produce a search archive. Next, the DIA files were searched along with this
372 search archive to generate a spectral library. The spectral library was then
373 used for normal DIA analysis in Spectronaut 14. Default settings (without N-
374 acetyl variable modification) were used in all steps.

375 ## Data analysis

376 Downstream data analysis for DDA samples was performed using Perseus
377 version 1.6.14.0[39]. Reverse hits and contaminants were removed, the data
378 log2-transformed, followed by a data subselection criteria of n=3 of 4
379 replicates in at least one sample. Missing values were replaced using the

normal distribution imputation method with default settings to generate a list of reliably quantified proteins. Subsequently, significantly changing differentially abundant proteins were determined and corrected for multiple comparisons (Bonferroni-corrected p-value < 0.05; q-value).

Direct DIA data analysis was performed on Spectronaut v.14 using default settings. Statistical analysis and plotting were performed using GraphPad Prism 8.

Network analysis was performed on Cytoscape v.3.8.0 using the StringDB plugin.

Data availability

Raw data have been deposited to the ProteomeExchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXDXXXXX. Source data used to produce all graphs is provided.

Acknowledgements

The authors thank Jack Moore (University of Alberta) for assistance with operating the mass-spectrometer. We are grateful to Fabia Simona and Oliver Bernhardt (Biognosys AG) for assistance troubleshooting the Spectronaut software analysis.

Author Information

Affiliations

Department of Biological Sciences, University of Alberta, Edmonton T6G 2E9, Alberta, Canada

Devang Mehta, Sabine Scandola, R. Glen Uhrig

Contributions

D.M., and R.G.U contributed to Conceptualization, Methodology, and Formal Analysis. D.M. and S.S. contributed to Investigation. D.M. contributed to Visualization and Writing (original draft). R.G.U. performed Supervision and Funding Acquisition. D.M., S.S., and R.G.U contributed to Writing (review & editing).

Corresponding author

Dr. R. Glen Uhrig: ruhrig@ualberta.ca

Ethics Declarations

Conflict of Interest

The authors declare no conflict of interest

References

1. Mehta, D. *et al.* Phosphate and phosphite differentially impact the proteome and phosphoproteome of Arabidopsis suspension cell cultures. *Plant J.* Accepted, (2020).

2. Clark, N. M. *et al.* Integrated omics networks reveal the temporal signaling events of brassinosteroid response in Arabidopsis. *BioRxiv* (2020). doi:10.1101/2020.09.04.283788

3. Vanderschuren, H. *et al.* Large-Scale Proteomics of the Cassava Storage Root and Identification of a Target Gene to Reduce Postharvest Deterioration. *Plant Cell* 26, 1913–1924 (2014).

4. Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods* 8, 937–940 (2011).

5. 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).

6. Graf, A. *et al.* Parallel analysis of Arabidopsis circadian clock mutants reveals different scales of transcriptome and proteome regulation. *Open Biol* 7, (2017).

7. Uhrig, R. G., Schläpfer, P., Roschitzki, B., Hirsch-Hoffmann, M. & Gruissem, W. Diurnal changes in concerted plant protein phosphorylation and acetylation in Arabidopsis organs and seedlings. *Plant J.* 99, 176–194 (2019).

8. Hartl, M. *et al.* Lysine acetylome profiling uncovers novel histone deacetylase substrate proteins in Arabidopsis. *Mol. Syst. Biol.* 13, 949 (2017).

9. Bekker-Jensen, D. B. *et al.* Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. *Nat. Commun.* 11, 787 (2020).

10. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* 537, 347–355 (2016).

11. Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J. & Mann, M. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat. Methods* 15, 440–448 (2018).

12. 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).

13. Ludwig, C. *et al.* Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* 14, e8126 (2018).

14. Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* 12, 258–64, 7 p following 264 (2015).

15. Li, Y. *et al.* Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. *Nat. Methods* 12, 1105–1106 (2015).

16. Biognosys AG. A new era in proteomics: spectral library free data independent acquisition (directDIA). *The Analytical Scientist* (2017).

17. Bruderer, R., Bernhardt, O. M., Gandhi, T. & Reiter, L. High-precision iRT prediction in the targeted analysis of data-independent acquisition and its impact on identification and quantitation. *Proteomics* 16, 2246–2256 (2016).

18. Reiter, L. MP 125: Direct Searching of DIA Data Catches up with Sample-specific Libraries. in *Proceedings of the 68th ASMS Conference on Mass Spectrometry and Allied Topics, Online Meeting* (American Society for Mass Spectrometry, 2020). at <https://biognosys.com/media.ashx/mp125lukasreiterasms2020.pdf>

19. Yang, Y. *et al.* In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nat. Commun.* 11, 146 (2020).

20. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* 17, 41–44 (2020).

21. Muntel, J. *et al.* Surpassing 10 000 identified and quantified proteins in a single run by optimizing current LC-MS instrumentation and data analysis strategy. *Mol. Omics* 15, 348–360 (2019).

22. Van Leene, J. *et al.* Capturing the phosphorylation and protein interaction landscape of the plant TOR kinase. *Nat. Plants* 5, 316–327 (2019).

23. Van Leene, J. *et al.* Targeted interactomics reveals a complex core cell cycle machinery in Arabidopsis thaliana. *Mol. Syst. Biol.* 6, 397 (2010).

24. Gonzalez, N. *et al.* A repressor protein complex regulates leaf growth in arabidopsis. *Plant Cell* 27, 2273–2287 (2015).

25. Antosz, W. *et al.* The Composition of the Arabidopsis RNA Polymerase II Transcript Elongation Complex Reveals the Interplay between Elongation and mRNA Processing Factors. *Plant Cell* 29, 854–870 (2017).

26. Dejonghe, W. *et al.* Disruption of endocytosis through chemical inhibition of clathrin heavy chain function. *Nat. Chem. Biol.* 15, 641–649 (2019).

27. Marondedze, C., Thomas, L., Serrano, N. L., Lilley, K. S. & Gehring, C. The RNA-binding protein repertoire of Arabidopsis thaliana. *Sci. Rep.* 6, 29766 (2016).

28. Arora, D. *et al.* Establishment of Proximity-dependent Biotinylation Approaches in Different Plant Model Systems. *Plant Cell* (2020). doi:10.1105/tpc.20.00235

29. 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).

30. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* 11, 2301–2319 (2016).

31. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. Upset: visualization of intersecting sets. *IEEE Trans Vis Comput Graph* 20, 1983–1992 (2014).

32. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613 (2019).

33. Jenkins, C. & Orsburn, B. Boxcar assisted MS fragmentation (BAMF). *BioRxiv* (2019). doi:10.1101/860858

34. Erickson, B. K. *et al.* Parallel Notched Gas-Phase Enrichment for Improved Proteome Identification and Quantification with Fast Spectral Acquisition Rates. *J. Proteome Res.* 19, 2750–2757 (2020).

35. Meier, F. *et al.* Parallel accumulation – serial fragmentation combined with data-independent acquisition (diaPASEF): Bottom-up proteomics with near optimal ion usage. *BioRxiv* (2019). doi:10.1101/656207

36. Uhrig, R. G. & Moorhead, G. B. Two ancient bacterial-like PPP family phosphatases from Arabidopsis are highly conserved plant proteins that possess unique properties. *Plant Physiol.* 157, 1778–1792 (2011).

37. Leutert, M., Rodríguez-Mias, R. A., Fukuda, N. K. & Villén, J. R2-P2 rapid-robotic phosphoproteomics enables multidimensional cell signaling studies. *Mol. Syst. Biol.* 15, e9021 (2019).

38. Cheng, C.-Y. *et al.* Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* 89, 789–804 (2017).

39. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* 13, 731–740 (2016).

Supplementary Tables

Supplementary Table 1: DDA protein quantification results for Arabidopsis cells

Supplementary Table 2: direct DIA protein quantification results for Arabidopsis cells

Supplementary Table 3: Comparison of protein quantification for Arabidopsis cells between DDA and direct DIA

Supplementary Table 4: DDA protein quantification results for HeLa digests

Supplementary Table 5: direct DIA protein quantification results for HeLa digests

Supplementary Table 6: Proteins identified in Arabidopsis cells using DDA

Supplementary Table 7: direct DIA protein quantification results for Arabidopsis cells filtered for valid values in 3 of 4 replicates.

Supplementary Table 8: DDA protein quantification results for Arabidopsis cells with no imputation

Supplementary Table 9: direct DIA protein quantification results for Arabidopsis cells filtered for valid values in all replicates.

Supplementary Table 10: DDA protein quantification results for Arabidopsis cells filtered for valid values in all replicates.

Supplementary Table 11: Proteins identified in HeLa digests using DDA

Supplementary Table 12: direct DIA protein quantification results for HeLa digests filtered for valid values in 3 of 4 replicates.

Supplementary Table 13: DDA protein quantification results for HeLa digests with no imputation

Supplementary Table 14: direct DIA protein quantification results for HeLa digests filtered for valid values in all replicates.

Supplementary Table 15: DDA protein quantification results for HeLa digests filtered for valid values in all replicates.

Supplementary Table 16: Proteins changing significantly in abundance between light and dark-grown Arabidopsis cells, measured using both direct DIA and DDA.

Supplementary Table 17: Precursor selection mass list table.

Supplementary Figures

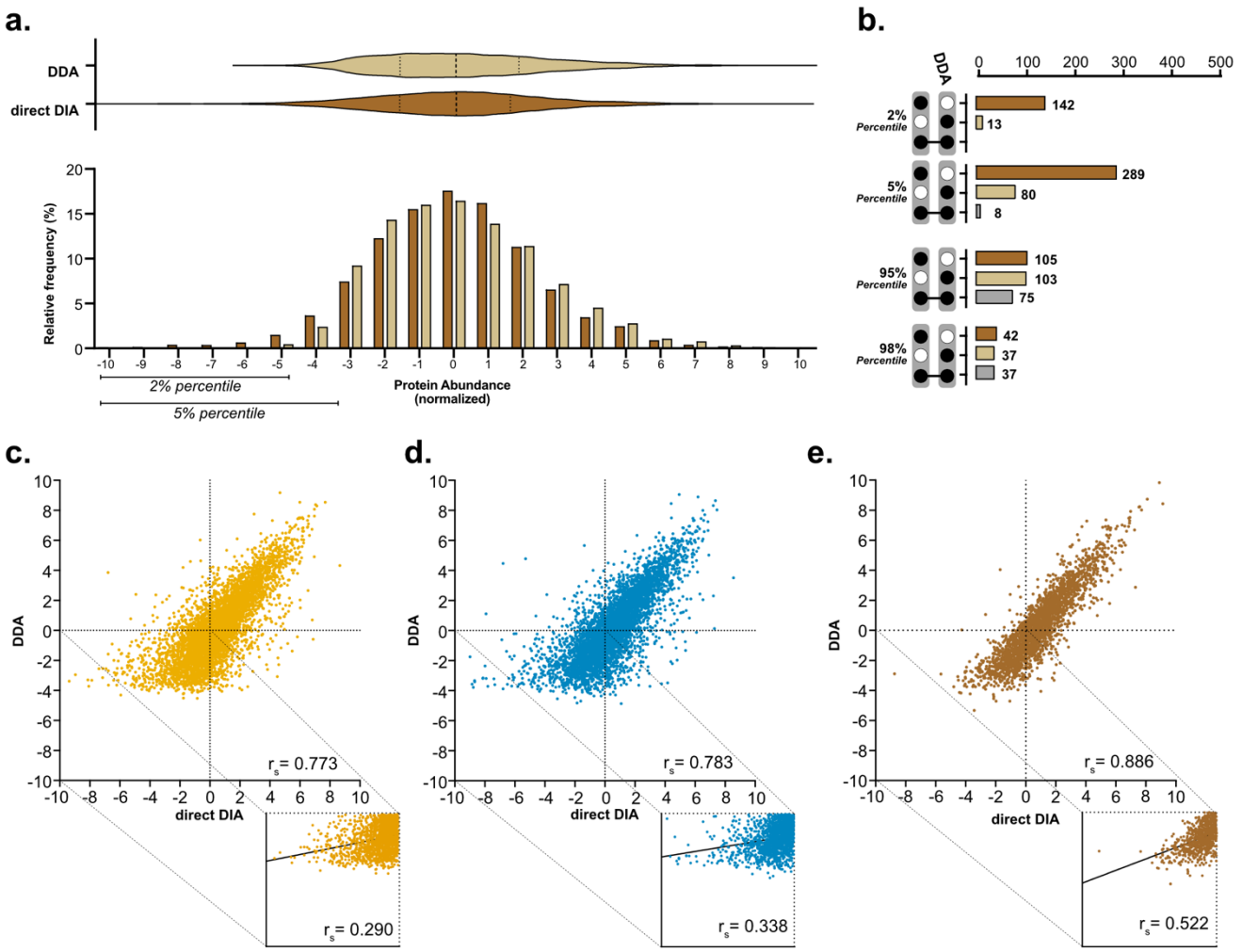

Figure S1: Comparison of protein quantification results using DDA and direct DIA analysis.
(a.) Frequency distribution of normalized protein abundances for DDA and direct DIA analysis and corresponding violin plots with median and quartile lines marked for HeLa digests. **(b.)** Upset plots depicting intersections in protein groups quantified by DDA and direct DIA at either extreme of the abundance distribution for HeLa digests. **(c.-e.)** Scatter plots of protein groups quantified by DDA and direct DIA for light-grown Arabidopsis cells, dark-grown Arabidopsis cells, and HeLa digests. Insets show correlations for protein groups with abundances less than the median.

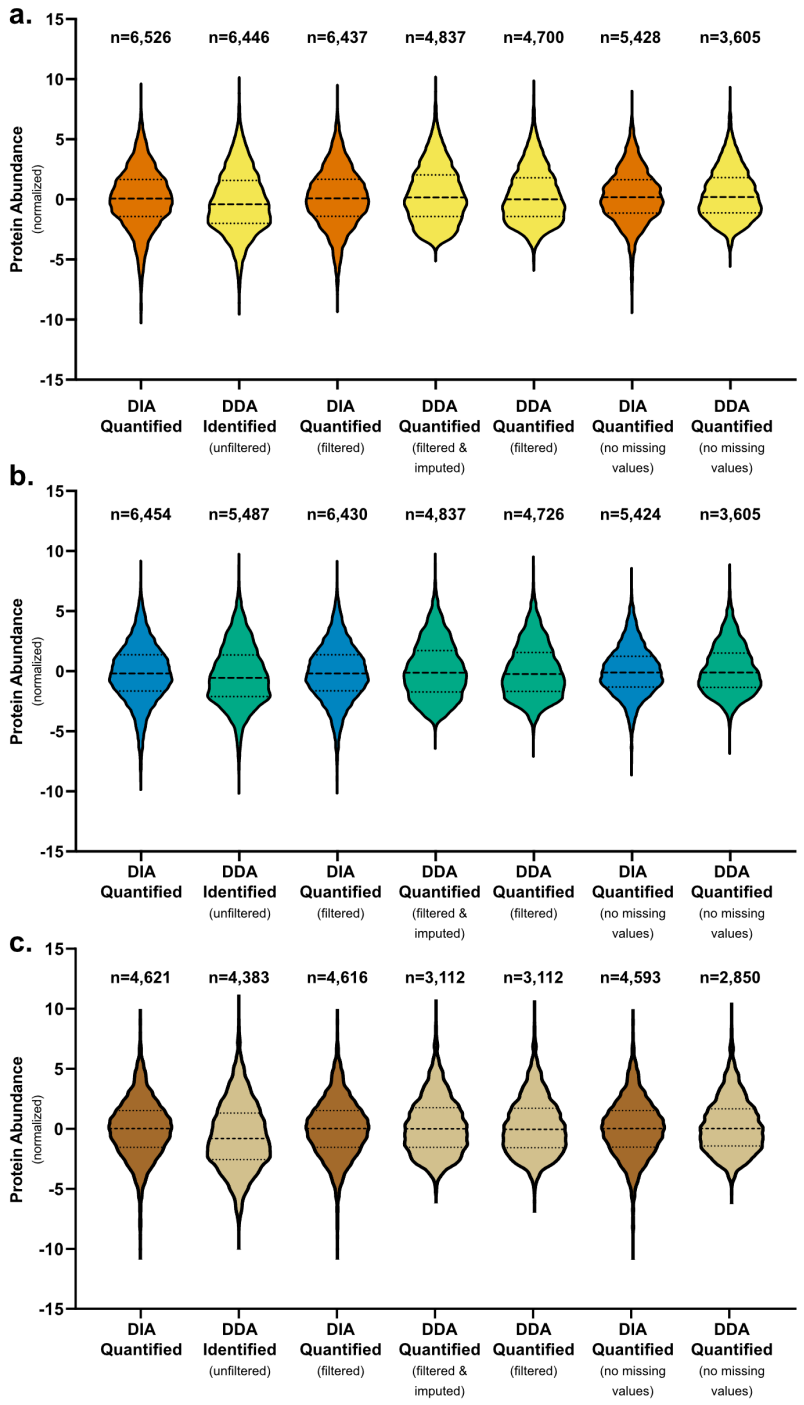

Figure S2: Protein abundance distributions by analysis type and data filtering settings.

Violin plots showing normalized protein abundance for proteins quantified by direct DIA (default setting), identified by DDA, quantified by DIA (filtered for protein groups present in at least 3 samples in any one condition), quantified by DDA (filtered for protein groups present in at least 3 samples in any one condition with missing values imputed), quantified by DDA (filtered for protein groups present in at least 3 samples in any one condition with missing values left blank), quantified by DIA (counting only protein groups found in all samples), and quantified by DIA (counting only protein groups found in all samples), respectively for **(a.)** light grown Arabidopsis cells **(b.)** dark grown Arabidopsis cells and **(c.)** HeLa cell digestion standards. (n= number of protein groups).

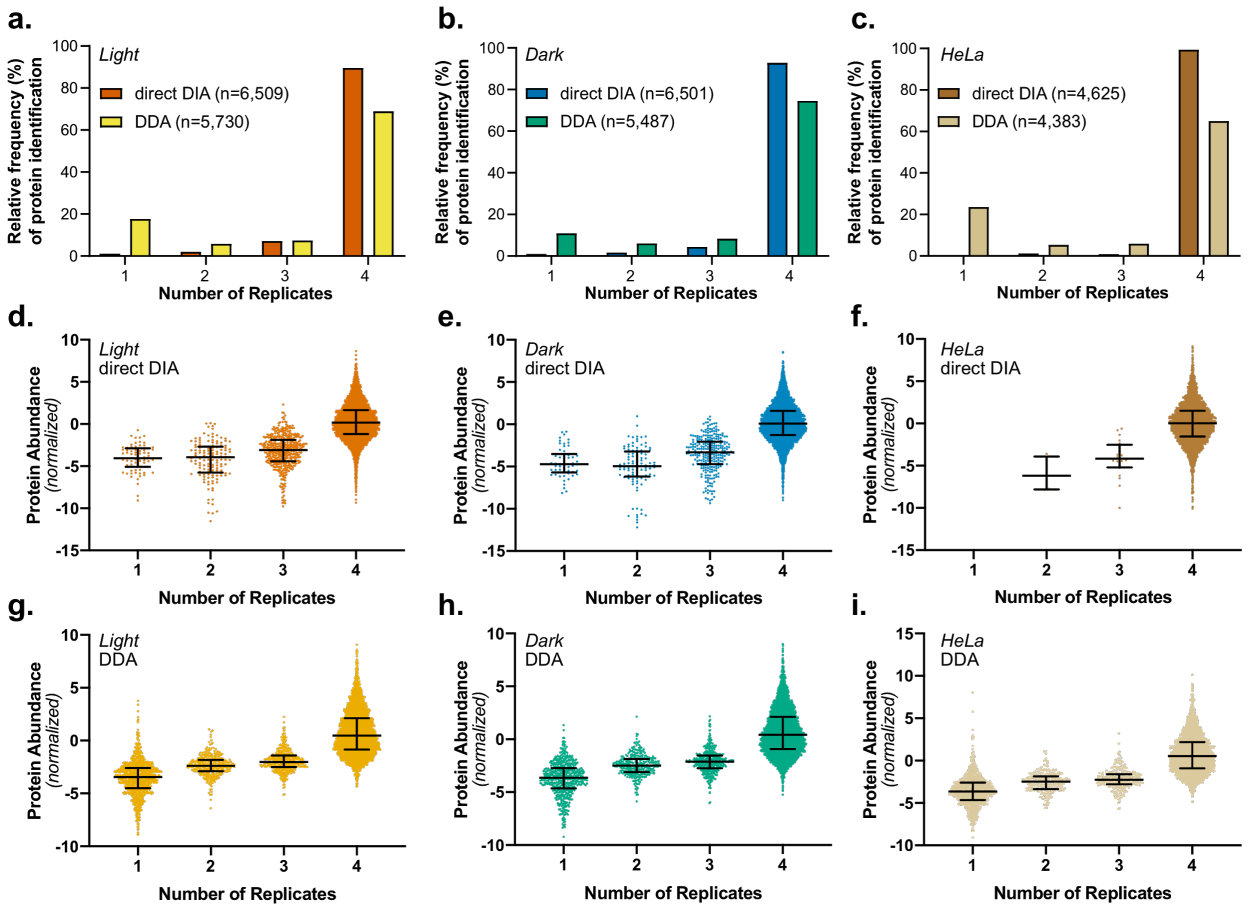

Figure S3: The DDA missing value problem explains the gap in quantification of low abundant proteins compared to direct DIA.

(a.-c.) Histograms of direct DIA or DDA protein group identifications across replicate samples for light-grown, dark-grown Arabidopsis cells, and HeLa cell digestion standards, respectively. **(d.-i.)** Normalized abundances of proteins binned by the number of replicates containing each protein for direct DIA and DDA. Bars represent median and interquartile range.