Improved detection and quantitation of RNA-interactomes using DIA SILAC

Thomas Tan*, Christos Spanos and David Tollervey*

Wellcome Centre for Cell Biology and Institute of Cell Biology, School of Biological Sciences, University of Edinburgh. Edinburgh EH9 3BF, Scotland, UK

* Correspondence to: Thomas.Tan@ed.ac.uk; d.tollervey@ed.ac.uk

Summary

Detection and quantitation of the RNA-interacting proteome is commonly achieved by applying SILAC labeling, following by data-dependent acquisition (DDA) of proteomic data. However, the limited sensitivity of the DDA approach often restricts detection to relatively abundant proteins, necessitating peptide fractionation prior to mass spectrometry. Here we report a pipeline for SILAC analysis using data-independent acquisition (DIA). The spectral library is constructed using gas-phase window separation of light material, followed by *in silico* prediction of heavy spectra. The resulting DIA datasets gave 30-40% greater protein detection relative to the same biological materials analyzed using DDA, while abolishing the requirement for fractionation of peptides following pre-MS reverse-phase separation. For proteins detected in both acquisitions, lower inter-replicate variation was seen with DIA. As a test, we determined the effects of arsenite treatment on the RNA-bound proteome of HEK293T cells, recovered following total RNA-associated proteome purification (TRAPP). The DIA dataset yielded clear GO term enrichment for RNA-binding proteins involved in cellular stress responses, while enrichment in the DDA dataset was relatively weak. Dataset normalization of with Cyclic Loess slightly improved the DDA-DIA correlation over median normalization for DIA datasets, but not for DDA dataset relative to default MaxQuant normalization. Overall, the DIA SILAC approach improved both protein detection and biological significance over conventional DDA SILAC.

Introduction

Proteomics is a rapidly advancing field that focuses on the study of proteins and their functions in biological systems. Most proteomic datasets are currently generated using tandem mass spectrometry (MS) to identify and quantify proteins based on the mass/charge ratio (MS1) followed by fragmentation fingerprint of their constituent peptides MS2. MS data can be acquired using two distinct methods: data dependent acquisition (DDA), the traditional method of acquisition, or data independent acquisition (DIA). In DDA, individual peptide ions from the MS1 cycle are selected for subsequent MS2 analysis, based on pre-set criteria. The technique is so named because data from MS1 is used to determine what data will be acquired in MS2. Generally, the most intense ions are selected, but this is not under control of the operator. This approach has several limitations, including the fact that it relies on the selection of a limited number of ions, which may not be representative of the entire sample. Additionally, DDA is biased towards high-abundance ions, which means that low-abundance ions may be overlooked.

In DIA, the entire MS1 output is sent to MS2. Fractions going into MS2 are separated based solely on time, with as good resolution as can be obtained for the machine (i.e. how fast the MS2 run can be performed), independent of the nature of the peaks present in each fraction. This results in a more comprehensive analysis of the sample, including low-abundance ions that may be missed by DDA. However, the highly complex fragment mass spectra generated from all precursor ions is challenging for data analysis using the conventional genome-wide protein database. Instead, a project-specific spectral library is usually created to optimize matches between the expected fragment ion chromatogram and the ions identified in experimental samples.

Stable isotope labelling by amino acids in cell culture (SILAC) enables the relative quantification of proteins in two or more samples. SILAC involves growing cells in media containing different stable-isotope labelled amino acids, such as heavy ($^{13}C_6$ and/or $^{15}N_2$) lysine or arginine, which are incorporated into newly synthesized proteins. The labelled and unlabeled cells are then mixed, lysed, and subjected to MS analysis. The relative abundance of proteins in each sample is determined based on recovery of heavy and light peptides, which are chemically similar but separate in MS. DDA is currently the predominant method to acquire SILAC data, using quantitation obtained in MS1. DIA is not commonly used for SILAC experiments, since software pipelines were designed to construct spectral libraries based on DDA data, and the generated DDA library often has spectral mismatch with the DIA experimental data. With recent development in both mass spectrometers and analysis software, spectral libraries can be constructed with DIA data. Protocols like openDIA allow library construction entirely *in silico*, using a given set of FASTA protein sequences.

Together these advances allow DIA to be applied to SILAC experiments. The DIA-SILAC workflow was developed using a spectral library constructed using narrow-window gas-phase fractionation of the light peptidome, followed by *in silico* generation of the heavy spectrum based on the light spectrum, and single-injection acquisition for experimental samples (Pino et al., 2021). The quantitation of peptides was based on MS2 chromatogram which is typically more accurate and reproducible than the MS1 quantitation used in DDA.

Isolation and quantitation of the RNA-bound proteome (RBPome) in cells can reveal major regulatory events in RNA function and metabolism that cannot be inferred from total proteomic data. For example, specific ribosomal proteins and translation initiation factors were found to dissociate from mRNA following after glucose withdrawal in yeast, while the cellular abundance of these proteins remained constant (Bresson et al., 2020). Total RNA-associated proteome purification (TRAPP) utilizes UV irradiation to covalently crosslink RNA and associated proteins. Subsequently, a silica column is used to immobilize all RNAs from a denatured cell lysate, allowing RNA-associated proteins to be recovered and identified by MS (Bresson et al., 2020; Shchepachev et al., 2019). There are no reliable internal reference proteins that allow absolute quantitation of proteins in the TRAPP dataset, so relative quantitation of proteins was achieved using DDA-SILAC. For optimal sensitivity, peptides were reverse phase fractionated using an acetonitrile gradient prior to MS injection.

Here we report the application of a DIA-SILAC workflow following TRAPP procedure. Data for the effects of arsenite treatment on human RBPome has been reported, so we adopted this as a test system. TRAPP was applied after acute (10 minutes) and prolonged (1 hour) arsenite treatment in human embryonic kidney HEK293T cells with three replicates. We aimed to improve sensitivity and accuracy of detection, while eliminating the requirement for pre-fractionation of peptides following reverse phase separation on acetonitrile gradients, thereby enhancing dataset qualities while improving workload and machine time economy.

Results

DIA dataset detected more proteins with lower inter-replicate variance compared to DDA for commonly detected RNA-binding proteins

122 µg of RNA extracted from each UV-crosslinked sample was mixed in equal amount with RNA of its SILAC pair, the mixture was purified on silicon columns and the co-purified proteins were digested off the column, and the peptides separated at this stage for either DDA or DIA acquisitions on an Orbitrap Exploris™ 480 Mass Spectrometer (procedures detailed in Material and Methods). The DDA dataset was processed with MaxQuant version

1.6.1.0 (Cox and Mann, 2008) while the DIA dataset with DIA-NN version 1.8.1 (Demichev et al., 2020). A protein was considered as detected if it was represented by 2 or more quantitated peptides, otherwise the protein quantity was replaced with NA for that sample. For DDA dataset, peptide quantity normalization and protein quantity calculation were done by MaxQuant; while for DIA dataset, peptide quantity was normalized by NormalyzerDE (Willforss et al., 2019) using Cyclic Loess, and protein quantity was calculated as the median quantity of all representing peptides.

To assess the sensitivity of the acquisition methods, the total number of detected proteins was compared; initially defined as all proteins with one or more intensity in any condition (Figure 1A left). The untreated DIA dataset had more detected proteins, at the range of 1900 proteins for 10min samples and 2150 for 1h samples; while the untreated DDA datasets had approximately 1420 proteins for 10min and 1650 proteins for 1h. The variation in the number of detected proteins between replicate samples were low indicating consistent sensitivity between samples. Since the non-treated controls (Heavy) at 10min and 1h should be biologically identical, the discrepancy in the number of detected proteins should be due to variations caused by sample preparation. To account for this variation, all analyses downstream were restricted to proteins detected in at least 2 out of 3 replicates in the untreated control at both time points (1399 proteins for DDA and 2179 proteins for DIA, Figure 1A middle). In this way, any protein missing in downstream analyses should be due to effects of arsenite treatment. Variations in quantitation between replicates in the same conditions were measured by the correlation coefficient (R^2) of Log2 raw (not normalized) intensities of proteins detected in any two of the three replicates (Figure 1A right). DDA and DIA datasets had similar correlation coefficients (R^2 ranged from 0.84 to 0.93) and consistency was good between correlations (SD ranged from 0.016 to 0.034). Distribution of raw protein intensities were displayed as violin and box plots (Figure 1B), with missing values imputed with 0.5 x minimum value of the sample. Arsenite-treated 10min samples in both acquisitions had more missing values for all detected proteins (mean of 18.35% in DDA and 21.73% in DIA) than 1h samples (5.11% in DDA and 12.77% in DIA). This was reduced in proteins passing the threshold for analysis (detected in at least 2 out of 3 replicates in the heavy samples), with 8.16% missing values in DDA and 20.38% in DIA at 10min, and 1.35% in DDA and 10.94% in DIA at 1h. DDA data were also more widely distributed, with an average Log2 interquartile range of 4.33 compared to 2.30 in DIA. Maximum inter-replicate standard deviations of Light/Heavy ratios for individual proteins in DDA and DIA were also compared (Figure 1C). In the total dataset, the median and mean maximum SD in DIA were markedly higher than those in DDA. However, the variations in DIA became noticeably lower when the proteins were restricted to those detected in both acquisitions (DDA and DIA), in all

conditions (10min and 1h), and in all conditions with enrichment following UV-crosslinking (protein list presented in Supplementary data 1). In the most restricted protein list, DDA dataset had a median/mean maximum SD of 0.1341/0.1641 while DIA had SDs of 0.0933/0.1339, indicating a better inter-replicate consistency in DIA compared to DDA for these proteins.

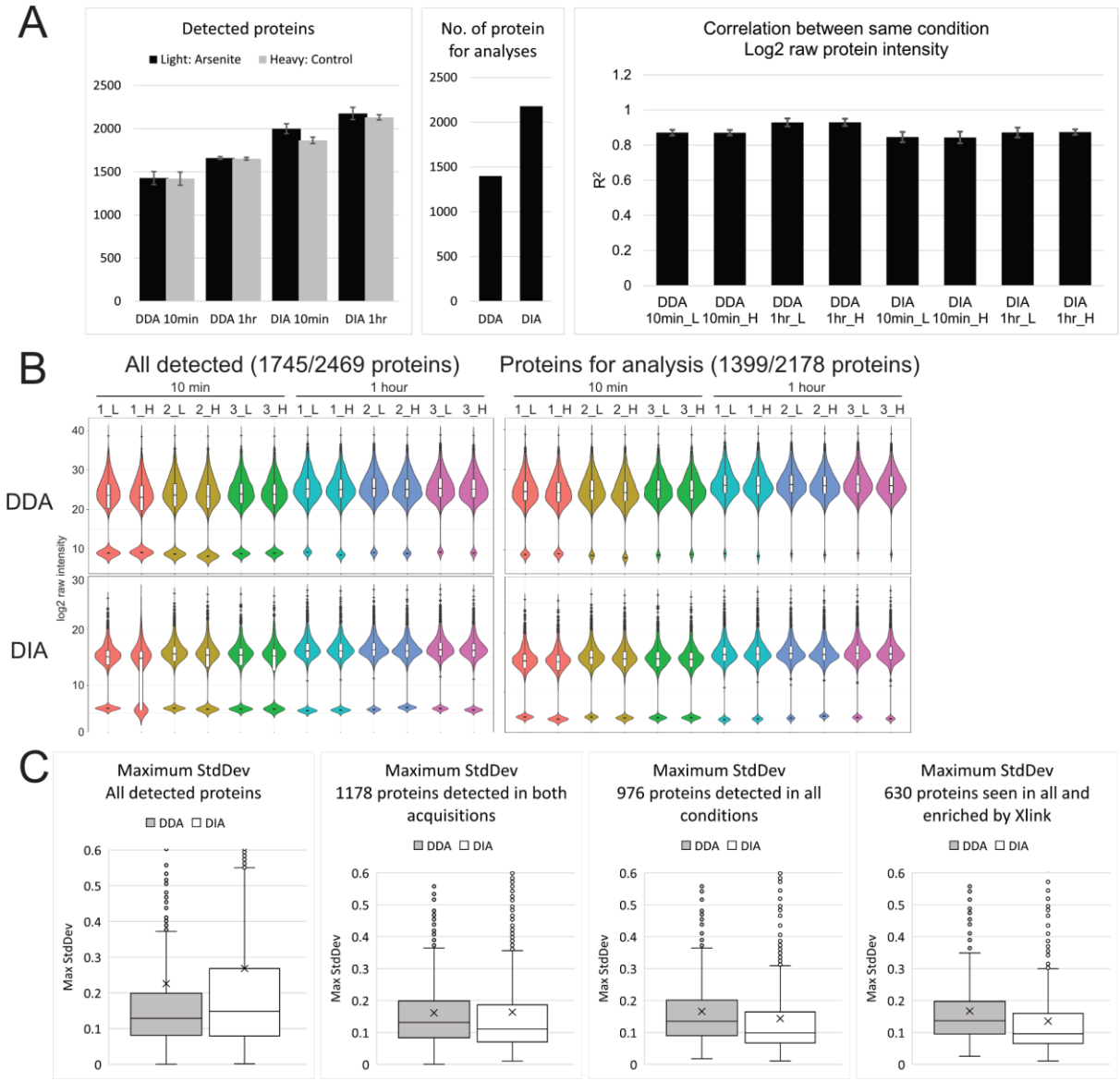


Figure 1. Data-independent acquisition detected more proteins compared to Data-dependent acquisition, with lower median and mean maximum standard deviation of SILAC light/heavy ratios for shared detected proteins and crosslink-enriched proteins. (A) Left: means (bar) and standard deviations (error bar) of number of detected proteins in 3 technical replicates were plotted for all 4 conditions. Middle: Number of proteins for analysis, defined as those represented by more than 2 peptides, and detected in at least 2 out of 3 control samples (heavy) in both time points. Right: Pearson correlation coefficient (R^2) between log2 raw protein intensities in 3 replicates of the same conditions. Bar, mean; error

bar, standard deviation. (B) Violin and box plot showing log2 raw intensity of all detected proteins or proteins for analysis. Colors denote the SILAC pairs. (C) Maximum standard deviation (higher of the two time points) of light/heavy ratios of the 3 replicates for all detected proteins, proteins detected in both acquisitions, proteins detected in all conditions, and proteins detected in all conditions and enriched in UVC-crosslinked over non-crosslinked samples. Center line, median; cross, mean; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers.

Quantitation by DDA and DIA was positively correlated for commonly detected proteins, correlation improved in crosslink-enriched and further-improved in significantly changed proteins

Several dataset normalization approaches were performed in MaxQuant and NormalyzerDE, with MaxQuant, Vsn and Cyclic Loess algorithms for DDA and Median, Vsn and Cyclic Loess for DIA. Table 1A shows the correlation between the two datasets with different combination of normalization methods. MaxQuant normalization was chosen for DDA and Cyclic Loess for DIA. A more detailed comparison between the algorithms is later described.

To eliminate the uncertainty in quantitation for proteins detected in only one sample out of 3 replicates, the Light/Heavy ratios for these proteins in all 3 replicates were replaced by 0. While for those detected in 2 out of 3 replicates, the one missing Light/Heavy ratio was replaced by the average of the other 2 replicates.

Distribution of Light/Heavy ratios across the datasets was plotted in Figure S1. Apart from proteins having missing values at one of the time points (the peak at 0), the ratios distributed normally in both acquisitions, allowing 2-sided T Tests to be applied for the differential RNA-association analysis between 10min and 1h time points. The statistical test was also complemented with Benjamini Hochberg multiple test correction to minimize false discovery. The fold-change in average ratio from 10min to 1h was used as an additional selection criterion for proteins with significantly altered RNA association, and also used to assess the similarity between DDA and DIA datasets. Proteins with a False Discovery Rate (FDR) <0.05 and an absolute fold-change >90th percentile of the dataset, were regarded as significantly changed in RNA interaction between 10min and 1h time points. The 90th percentile for fold change was 1.193 in DDA and 1.274 in DIA, these values were used as the significance threshold instead of an arbitrary value (commonly 1.5 or 2) to maintain a consistent stringency between the datasets.

We identified proteins showing significant changes in RNA association between 10min and 1h of arsenite treatment, that were also detected at both time points in the non-treated controls. DDA identified 324 proteins with significantly changed recovery in the RNA-bound

fraction. Most (210 out of 324) not found at 10min, while 28 were not detected at 1h; with 86 non-missing proteins showing altered recovery between time points. Similarly, DIA identified 305 proteins showing significantly changed RNA-binding, with 213 not recovered at 10min, 35 missing at 1h and 57 detected at both time points. 46 proteins were found to be significant according to both acquisitions. Of these, 4 were missing at 10min in DDA while detected at both time points in DIA, 1 was missing at 10min in DIA while detected at both time points in DDA, and 1 was missing at different time points between the two acquisitions; discrepancies resulting in conflicts in their dynamics (increased or decreased RNA interaction) reported between DDA and DIA. The remaining 40 significant proteins gave consistent dynamics. 16 were detected in all conditions, 2 had missing values at one time point in DDA while detected at both time points in DIA, and 22 were missing at 10min in both acquisitions (Figure 2A). Taken together, significantly changed proteins in both datasets consist mainly of those missing at 10min in arsenite-treated samples, and the agreement between DDA and DIA (40 out of 324/305) was low. Further analysis confirmed low similarity between proteins missing at one time point in one acquisition while detected at both time points in the other acquisition (Figure S2A). However, for proteins detected at both time points in both acquisitions the similarity between the two datasets was higher (R^2 of 0.58 for proteins significant in DDA, and 0.86 for proteins significant in DIA, Figure S2B).

DDA/DIA correlation also improved progressively when proteins for analysis became more restricted, with R^2 value of 0.47 for proteins detected in all conditions, 0.53 for proteins detected in all conditions while enriched by crosslinking, and 0.91 for proteins detected in all conditions, enriched by crosslinking, and significant according to DIA (Figure 2B). This result indicates that quantitation in the DIA dataset was similar to the DDA dataset, when potential sources of noises were eliminated. Also, for proteins missing at one time point, their dynamics are not reliably measured in either acquisition. The number of proteins with no missing value at any time points was 1067 in DDA and 1673 in DIA (Figure 2C).

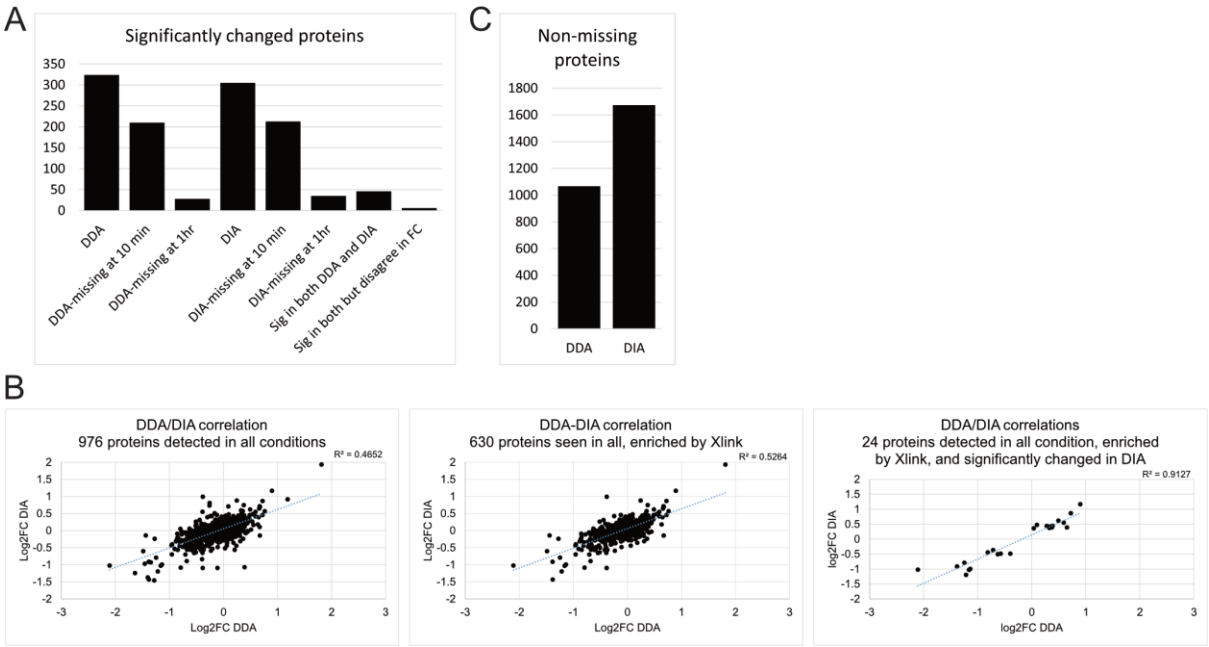


Figure 2. Positive correlation between protein quantitation and fold-change in SILAC light/heavy ratio between time points in the DDA and DIA datasets. (A) Number of significantly changed proteins (2-sided T Test with Benjamini-Hochberg corrected FDR<0.05, absolute FC>90 percentile), proteins missing in 10min samples, in 1h samples, common significant proteins in both datasets, proteins significant in both datasets but have opposite fold-changes. (B) Dot plot of log2 fold change in light/heavy ratio from 10min to 1h time points in DDA and DIA datasets for proteins detected in all conditions, proteins detected in all conditions and enriched by crosslinking, and proteins detected in all conditions, enriched by crosslinking, and significantly changed according to the DIA dataset. (C) Number of proteins with no missing value at any time point in DDA and DIA datasets.

Cellular stress response factors were more clearly enriched in significantly up-regulated proteins in DIA dataset

A well characterized cellular mechanism observed following arsenite exposure is the integrated stress response. Translation of mRNAs into proteins is halted and stress granules are formed to sequester mRNAs of which the translation is inhibited. This occurs between 30min and 1h of treatment with high arsentite dosages (0.5-2.5mM) (Shih et al., 2011; Zurla et al., 2011). Comparing the RNA-interactomes at 10min and 1h of arsenite treatment, should therefore reveal increasing recovery of translation regulators and stress granule-related factors. Since the DDA/DIA correlation is good for proteins detected in all conditions (listed in Figure 2C), a GO-term enrichment analysis was performed using these proteins with significantly altered RNA-association in the two datasets (Figure 3). For both DDA and DIA, proteins with significantly increased RNA-interactions from 10min to 1h were frequently linked to GO terms related to mRNA binding, single or double-stranded RNA binding and ribonucleoprotein complexes. However, only for the DIA dataset was GO term enrichment

statistically significantly (FDR < 0.05, single- and double-stranded RNA binding). Both datasets included translation regulators, including translation initiation factor eIF4H, DEAD/DExH-box helicases DDX3X and DHX9, and other stress granule related factors, incuding FMR1 in DIA and FXR2 in DDA (Lai et al., 2020). For proteins with significantly decreased RNA-interaction, both datasets included specific ribosomal proteins with known extraribosomal functions (Warner and McIntosh, 2009) including RPS3, RPS14 and RPS28. These decreases in RNA interactions likely indicate changes in specific mRNA regulation, rather than a global reduction in mRNA translation. Together, proteins significantly changed in the DIA dataset showed clearer enrichment for mRNA translation regulators, and the DIA dataset gave better discrimination in selecting proteins with significantly reduced RNA association, especially the ribosomal proteins. A more relaxed comparison was made using also proteins deemed significantly changed using T Test P value without the Benjamini-Hochberg correction (Figure S3). In this case both datasets had clear enrichment in mRNA binding and double-stranded RNA binding (FDR < 0.05) and more ribosomal proteins in both increased and decreased RNA interaction sets. We conclude that the DIA dataset, with or without the BH correction, was more stringent in selecting proteins with significantly altered RNA-binding. Fewer significant proteins were called, with better enrichment for known mRNA, single or double stranded RNA binding factors, which are very likely involved in stress response.

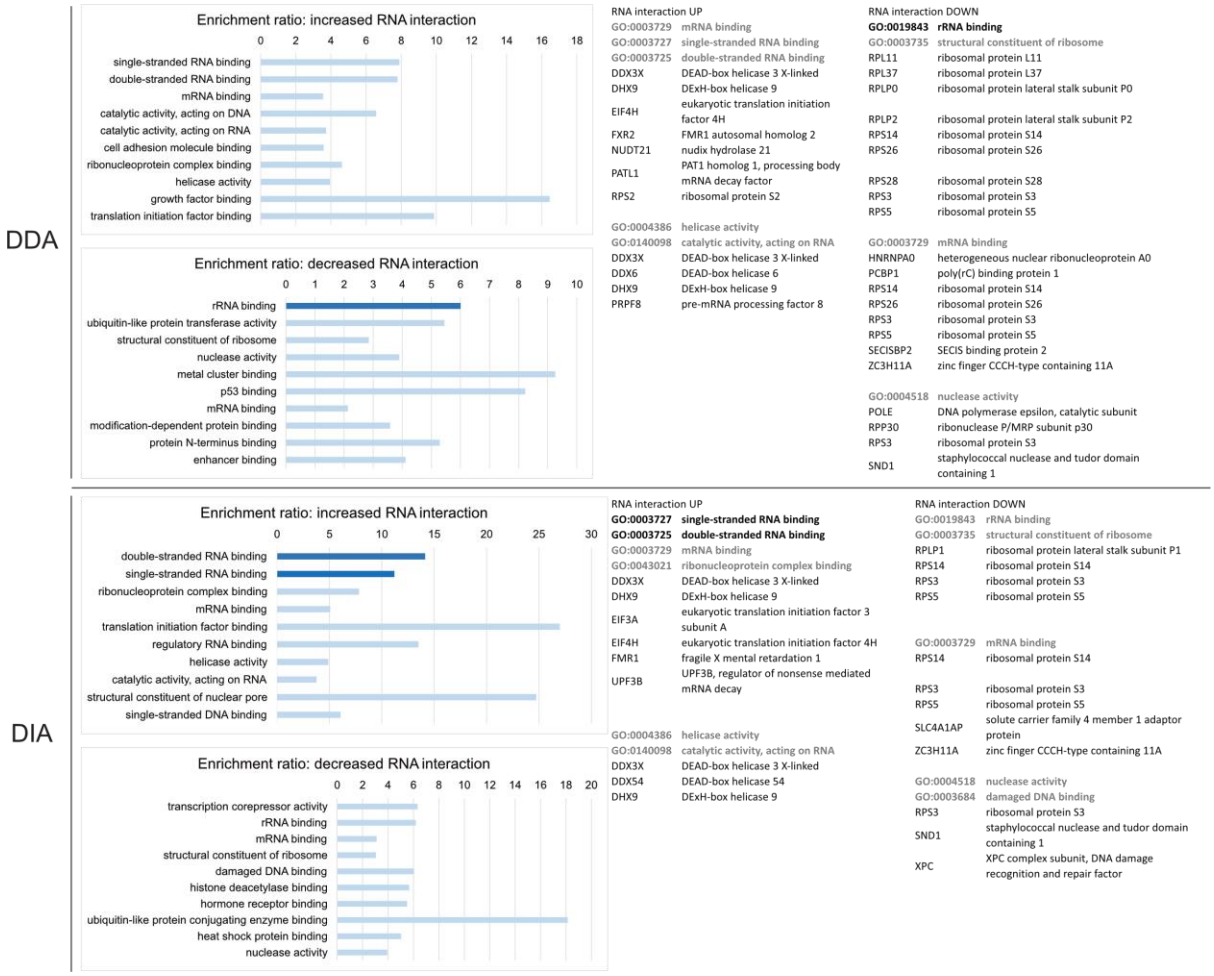


Figure 3. Top-ten enriched GO terms in molecular function for proteins with increased or decreased RNA interaction from 10min to 1h and detected at both time points according to the DDA or DIA dataset, and selected gene sets possibly related to the stress-induced RNA translation shutdown. Dark bar, FDR<0.05; Pale bar, FDR>0.05. GO terms were sorted by FDR in ascending order, or by pre-adjusted P value if all FDRs were 1.

Vsn and Cyclic Loess data normalization did not improve DDA-DIA correlation over MaxQuant/Median normalization but reduced inter-replicate variation for DIA datasets

Previous analyses on DDA/DIA correlation and differential RNA interaction were based on MaxQuant-normalized DDA dataset and Cyclic Loess-normalized DIA dataset. To test whether other methods of normalization enhance DDA/DIA correlations and/or inter-replicate variation, the two datasets were separately normalized by MaxQuant (DDA) and Median (DIA), Vsn, or Cyclic Loess. The Vsn algorithm was chosen due to its reported superior performance in reducing inter-replicate variation in total proteomic datasets (Välikangas et al., 2018). Cyclic Loess was originally designed to handle two-color microarrays with different efficiencies of dye incorporation (Dudoit et al., 2002), similar to the condition in Light

and Heavy incorporation in SILAC experiments. Correlation between DDA and DIA datasets was measured by the R^2 value between Log2FC of proteins detected in all conditions (Table 1A). Variation between replicates was measured by the maximum standard deviation (greater of the 10min and 1h time points) between normalized Light/Heavy ratios (Table 1B). Amongst these methods, the MaxQuant/Median pair produced the highest R^2 value of 0.4732, while Vsn and Cyclic Loess normalization produced similar but lower R^2 values. Cyclic Loess reduced the median and mean maximum SD for DIA, but not for the DDA dataset. The standard analyses in this article were therefore based on the MaxQuant for DDA and Cyclic Loess for DIA, giving relatively good correlations with optimized inter-replicate variation for the DIA dataset.

A R^2

DDA \ DIA	Median	Vsn	CycLoess
MaxQuant	0.4732	0.4718	0.4685
Vsn	0.4046	0.4030	0.3921
CycLoess	0.3915	0.3897	0.3872

B Proteins detected in all conditions, enriched by crosslink

Max SD	Median	Mean
DDA-MaxQuant	0.1341	0.1641
DDA-Vsn	0.1386	0.1751
DDA-CycLoess	0.1355	0.1703
DIA-Median	0.1055	0.1413
DIA-Vsn	0.0958	0.1361
DIA-CycLoess	0.0933	0.1339

Table 1. Correlation between, and level of variation in DDA and DIA datasets normalized with MaxQuant/Median, Vsn, or Cyclic Loess. (A) R^2 values for Log2FC of individual proteins between datasets normalized by different methods, calculated using proteins detected in all conditions. (B) Median and mean of maximum standard deviation between replicates for proteins detected in all conditions and enriched by crosslink, in datasets normalized with different algorithms.

Imputation by Random Forest dramatically reduced significantly changed proteins compared to Group Minimum

All analyses above were based on the processed datasets in which proteins with 1 missing value in 3 replicates had the missing value replaced by the mean value of the other two replicates, while proteins with 2 or more missing values had all the 3 replicates replaced by 0. We also tested two imputation methods for handling missing values affect DDA/DIA correlations. Based on the assumption that values are missing due to low abundance, we tested the use of 0.5 x the minimum Light/Heavy ratio in the sample (MINx0.5). Alternatively, machine learning can be applied, training the algorithm to predict the missing ones based on the observed values. Random Forest (RF) (Stekhoven and Bühlmann, 2011) uses such a strategy and was reported to out-perform other imputation algorithms in LC-MS metabolomic (Kokla et al., 2019) and label-free proteomic data (Zhang et al., 2021). The two methods were used here to impute both DDA and DIA datasets at the protein level (Light/Heavy ratios of the 3 replicates at 2 time points, normalized using the MaxQuant/Cyclic Loess pair). Proteins with more than 4 missing values in the total 6 samples were removed prior to the analysis, leaving at least 2 non-missing values to aid prediction. DDA/DIA correlations are shown in Figure 4A, for imputed datasets for proteins consistently detected in the non-treated (heavy) control, and enriched by UV-crosslinking while significantly changed in DIA. Similar to previous analysis, quantitation for proteins missing at one of the time points was not reproducible between datasets, causing populations with extreme Log2FC values mainly in DDA with inconsistent values in DIA, and an overall poor correlation between datasets (R^2 = 0.08) when imputed by MINx0.5. These populations could not be seen when RF was used and the correlation improved (R^2 = 0.27). However, this was still lower than the previously processed dataset with proteins detected in all conditions (R^2 = 0.47). Using the same previous criteria for selecting significantly changed (FDR <0.05, absFC >90th percentile) and UV-crosslink-enriched proteins, both MINx0.5 and RF called fewer significantly altered proteins (6 and 3, respectively) compared to the previously processed datasets (24 proteins). DDA/DIA correlation for significant proteins was poorest in RF (R^2 = 0.46), followed by the previously processed datasets (R^2 = 0.91) and MINx0.5 (R^2 = 0.97). It appears that the Random Forest algorithm provided the most stringent selection for proteins with significant alteration in RNA association. However, the number of significantly altered proteins may be too few to properly assess DDA-DIA correlations.

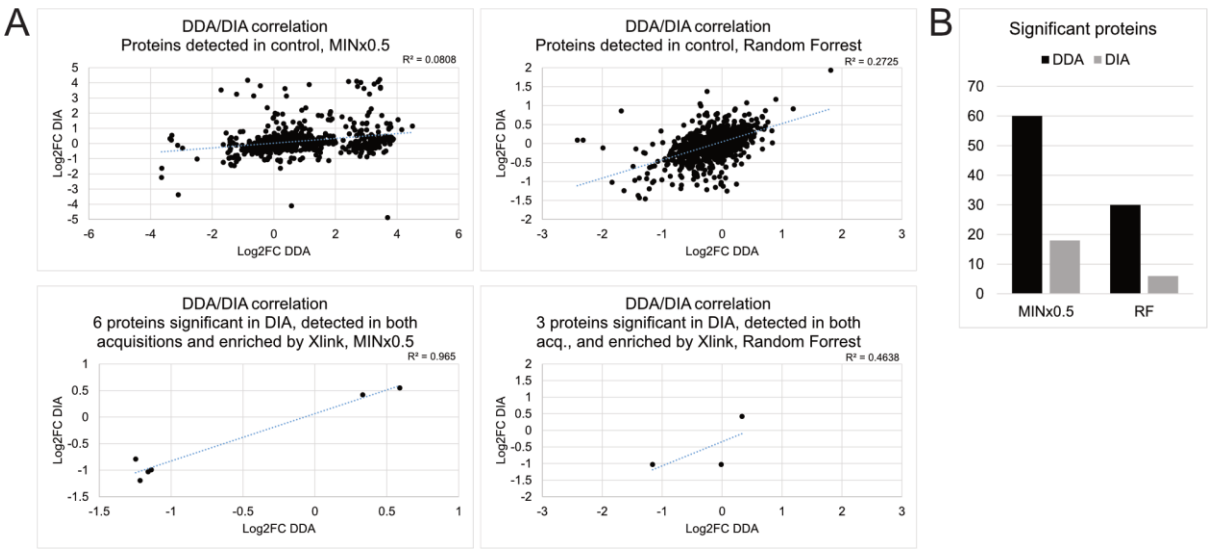


Figure 4. Imputation by Random Forrest eliminated the contrast between detected and non-detected, and thereby reduced the number of significantly changed proteins. (A) Like-for-like comparison of DDA/DIA correlation between original datasets with proteins having more than 2 non-missing values, with missing values replaced using Random Forrest or the minimum value of the sample x 0.5. Comparison was restricted to all proteins detected in non-treatment controls and proteins enriched by UV-crosslinking and significantly changed between 10min and 1h of treatment according to DIA. (B) Number of proteins regarded as significantly changed between 10min and 1h of treatment (2-sided T test with Benjamini Hochberg multi-testing correction, FDR<0.05, absolute FC>90 percentile) in datasets imputed by Random Forrest or the minimum value of the sample x 0.5.

Discussion

Quantitation of dynamic RNA-protein associations requires consistent relative measures of ratios between recovery of experimental samples and controls, which can be achieved using the SILAC labelling. However, conventional DDA-SILAC tandem-MS involves selective, and usually not customizable, peptide identification and quantitation. In contrast, DIA-SILAC does not apply selection between MS1 peptides, potentially quantifying a greater fraction of the total population.

In the DDA datasets reported here, 1067 proteins were reliably quantified between replicates. However, proteins classed as significantly altered in RNA association were not significantly enriched for mRNA translation regulators and stress granule formation; factors expected to be dynamically controlled during cellular stress following arsenite exposure. In comparison, the DIA dataset reliably quantified ~57% more proteins than to DDA, together with better discrimination for significantly changed proteins. In consequence, there was clear enrichment of expected, stress-associated regulators among proteins showing significantly increased RNA association.

Both datasets included many proteins that failed to be detected at one time point. These proteins would be classed as low abundance at these time points, and called as significantly increased/decreased. However, quantitation in other acquisitions frequently revealed quite different dynamics. Reliable quantitation therefore requires proteins to be detected at both time points, and the DIA dataset was superior in detecting such proteins. This was the case even without peptide fractionation following the pre-MS reverse phase column, and the consequential multiple MS injections per biological sample. DIA produced dataset that appeared more biologically informative, while reducing labor and machine time expenditure.

Due to the analysis of all detected peptide fragment ions, raw DIA datasets contain more noise than DDA datasets, resulting in higher median and mean standard deviations between replicates. However, when suitable filters were present, in this case a list of proteins enriched by UV-crosslinking and significant according to statistical tests, the inter-replicate variation in the DIA dataset became lower than the DDA dataset, and correlation between datasets was high ($R^2 > 0.90$).

All methods of dataset normalization tested gave similar DDA/DIA correlations and inter-replicate variation. However, normalization by MaxQuant for DDA, and by Cyclic Loess for DIA gave the best balance between correlation and variation. Both MINx0.5 and Random Forest imputation strategies increased the stringency of significant protein selection. However, the MINx0.5 method did not allow selection of proteins detected at both time points, adversely impacting the DDA/DIA correlation. Random Forest imputation was the most stringent, resulting in only three proteins being classed as significantly changed (identified in DIA, enriched by UV-crosslinking, detected in both acquisitions), from 24 without imputation. Given the low numbers of proteins with significantly altered RNA-interactions, datasets in the main part of this article were analyzed without imputation to maximize significant protein detection.

We conclude that analyses of protein-RNA interactome using DIA can give superior results to the more conventional DDA approach, as well as reducing the machine time required for the analysis.

Material and Methods

Cells

HEK293T cells were grown in DMEM with GlutaMax (Gibco; cat. 10566016) supplemented with 10% fetal calf serum (Sigma; cat. F2442). Cells were seeded at a confluency of 10% and grown to 70 - 90% before being split or used for experiments. To split cultures, cells

were washed once with sterile PBS, then incubated with 0.25% Trypsin-EDTA (Gibco; cat. 25200056; 0.2 times volume of culture media removed) until detached. Trypsin was inactivated with 5 volumes of media, and cells pelleted at 300 RCF for 5 minutes before resuspension in culture media.

For SILAC experiments, cells were grown for 10 divisions in SILAC DMEM (Thermo Fischer; cat. A33822) supplemented with 10% dialyzed FBS (Sigma-Aldrich; cat. F0392) and 50 µg/L each of lysine and arginine. For "light" cultures, these amino acids were obtained from Sigma. For "heavy" cultures, $^{13}C_6$-lysine and $^{13}C_6$-arginine were obtained from Cambridge Isotope Laboratories (cat. CLM-226 and CLM-2247).

Immediately prior to harvest, sodium arsenite (Merck; cat. 1.06277) was added to the light culture to a final concentration of 1mM and incubated for 10 minutes or 1 hour. An equal volume of water was added to the heavy culture and incubated for the same duration. At the end of the incubation cells were washed with PBS and snap-frozen on the dish in dry ice/methanol and crosslinked with 400 mJ/cm^2 UVC using a Stratalinker 2400 (Stratagene). Cells were stored at -80°C.

All cell cultures were maintained at 37°C with 5% CO2.

Isolation of RNA interactome using TRAPP

Matching SILAC pairs were each resuspended and homogenized in 3.6 mL of a 1:1 mix between phenol pH 8.0 (Sigma-Aldrich; cat. P4557) and GTC lysis buffer (4 M guanidine thiocyanate, 50 mM Tris-HCl pH 8.0, 10 mM EDTA, and 1% b-mercaptoethanol), and combined in equal proportion (244 µg RNA total in 4 mL phenol-GTC). 0.1 x volume of 3M sodium acetate pH 4.0 was added to the mixed extract and incubated at 80 ^{o}C for 5 minutes, with 1000 rpm agitation. The mixture was spun at 13,000 rcf for 10 minutes at RT and supernatant was transferred to a new tube, in which 0.4286 x volume of 100% ethanol was slowly added and mixed by vortexing. The extract was loaded onto three EconoSpin column for RNA (Epoch Life Science; cat. 1940-250) and each column washed once with 600 µL WB1A (3 M guanidine thiocynate, 1 M sodium acetate pH 4.0, 30% ethanol). Collection tubes were replaced and the columns were washed once with 600 µL WB1B (2 M guanidine hydrochloride, 20 mM Tris-HCl pH 6.4, 50% acetone, 20% ethanol), twice with 600 µL WB2B (20 mM Tris-HCl pH 6.8, 0.5 mM EDTA, 80% ethanol) and then once with 500 µL WB3B-80 (80% acetonitrile, 50 mM Tris-HCl pH 7.8, 2 mM $CaCl_2$). After the last wash columns were spun for 2 minutes at 16,000 rcf. Collection tubes were replaced with 1.5 mL low protein binding tubes (Thermo Scientific; cat. 90410). To each column was added 0.25 µg Trypsin+Lys-C (Promega; cat. V5073) in 75 µL WB3B-60 (60% acetonitrile, 50 mM Tris-HCl pH 7.8, 2 mM $CaCl_2$) with incubation for 16 hours at 37 ^{o}C. The three columns were eluted

by spinning at 16,000 rcf for 30 seconds, and each column was washed with 75 µL WB3B-60. Eluates from the three columns were pooled and spun at 16,000 rcf for 5 minutes. Supernatant was transferred to a new tube. 1M DTT (Sigma-Aldrich; cat. 43816) was added to the eluate to a final concentration of 5mM and the mixture was incubated at 37 ^{o}C for 20 minutes. Peptide alkylation was done with 10 mM final Iodoacetamide (Merck; cat. A3221) at RT in the dark for 20 minutes, followed by inactivation with 10 mM final DTT, incubated at RT for 5 minutes. 0.15 µg Trypsin+Lys-C in 7.5 µL WB3B-60 was added and incubated at 37 ^{o}C for 1 hour for complete digestion of the peptides. The volume of the peptides was evaporated to ~10 µL and resuspended in 200 µL Buffer A (0.1% NH_4OH). The material was split at this stage, half was reverse phase fractionated using an SDB-XC stage tip with 10%, 20% and 80% acetonitrile (flow-through was acidified, cleaned with a C18 stage tip and pooled with the 80% fraction) and used for DDA; another half was acidified with 10% TFA to pH ~1.0 and cleaned with a C18 stage tip and used for DIA.

Mass spectrometry

For library generation, 30 µg of total protein from HEK293T cells grown in Light medium were digested using the Filter Aided Sample Preparation (FASP) protocol published previously (Wiśniewski et al., 2009) with minor modifications. In brief, protein sample was added on top of a 30 kDa MWCO filter unit (Vivacon, UK), along with 150 µl of denaturation buffer (8M Urea in 50mM ammonium bicarbonate (ABC) (Sigma Aldrich)) and spun at 14,000 x g for 20 min. An additional wash with 200 µl of denaturation buffer was performed under the same conditions. The protein sample was then reduced with 100 µl of 10 mM dithiothreitol (Sigma Aldrich, UK) in denaturation buffer for 30 min at ambient temperature, and alkylated by adding 100 µl of 55 mM iodoacetamide (Sigma Aldrich, UK) in denaturation buffer and incubated for 20 min at ambient temperature in the dark. Two washes with 100 µl of denaturation buffer and two with digestion buffer (50mM ABC) were performed under the conditions described above, before the addition of trypsin/LysC mix (Pierce, UK). The protease:protein weight ratio was 1:50 and proteins were digested overnight at 37°C. Following digestion, the sample was spun at 14,000 x g for 20 min and the flow-through containing digested peptides was collected. Filters were then washed one more time with 100 µl of digestion buffer and the flow-through was collected again. The eluates from the filter units were acidified using 20 µl of 10% Trifluoroacetic Acid (TFA) (Sigma Aldrich), and spun onto C_{18} StageTips as previously described (Rappsilber et al., 2007). Peptides were eluted in 40 µL of 80% acetonitrile in 0.1% TFA and concentrated down to 1 µL by vacuum centrifugation (Concentrator 5301, Eppendorf, UK). The peptide sample was then prepared for LC-MS/MS analysis by dilution to 5 µL using 0.1% TFA.

LC-MS analyses were performed on an Orbitrap Exploris™ 480 Mass Spectrometer (Thermo Fisher Scientific, UK), coupled on-line to an Ultimate 3000 HPLC (Dionex, Thermo Fisher Scientific, UK). Four 2µg-injections of the HEK cell digest were performed. Peptides were separated on a 50 cm (2 µm particle size) EASY-Spray column (Thermo Scientific, UK), which was assembled on an EASY-Spray source (Thermo Scientific, UK) and operated constantly at 50°C. Mobile phase A consisted of 0.1% formic acid in LC-MS grade water and mobile phase B consisted of 80% acetonitrile and 0.1% formic acid. Peptides were loaded onto the column at a flow rate of 0.3 µL min^{-1} and eluted at a flow rate of 0.25 µL min^{-1} according to the following gradient: 2 to 40% mobile phase B in 120 min and then to 95% in 11 min. Mobile phase B was retained at 95% for 5 min, then returned to 2% in a minute, and stayed at 2% until the end of the run (160 min).

For all four injections survey scans were recorded at 120,000 resolution (scan range 350-1650 m/z) with an ion target of 5.0e6, and injection time of 20ms. MS2 DIA was performed in the orbitrap at 30,000 resolution, maximum injection time of 55ms and AGC target of 3.0E6 ions. We used HCD fragmentation (Olsen et al., 2007) with stepped collision energy of 25.5, 27 and 30. For the first injection the scan range was between 400 and 600 m/z with isolation windows of 5 m/z. For the other three injections the scan range was 600-800, 800-1000 and 100-1400, with 5 m/z isolation windows and default charge state of 3.

For SILAC TRAPP samples acquired with DDA, the LC-MS preparation and the injection conditions were as described above. MS1 was performed at 120,000 resolution with scan range 350-1500 m/z, AGC target of 3.0E6 ions and injection time of 50 ms. MS2 scans were performed at resolution of 15,000 with isolation window of 1.4 Thompson AGC target of 8.0E5 and injection time of 60 ms.

For SILAC TRAPP samples acquired with DIA, LC-MS conditions were the same as above. Variable isolation windows was used as the table below:

m/z	z	t start (min)	t stop (min)	Isolation Window (m/z)
383.375	3	0	185	66.8
423	3	0	185	13.5
435	3	0	185	11.5
446.5	3	0	185	12.5
458	3	0	185	11.5
469	3	0	185	11.5
480	3	0	185	10.5
490.5	3	0	185	10.5
501	3	0	185	11.5
512	3	0	185	11.5

523	3	0	185	11.5
533.5	3	0	185	10.5
544	3	0	185	11.5
554.5	3	0	185	10.5
565	3	0	185	11.5
575.5	3	0	185	10.5
586	3	0	185	11.5
597.5	3	0	185	12.5
609.5	3	0	185	12.5
621.5	3	0	185	12.5
633	3	0	185	12.5
645	3	0	185	13.5
657.5	3	0	185	12.5
670.5	3	0	185	14.5
684	3	0	185	13.5
697	3	0	185	13.5
710.5	3	0	185	14.5
725.5	3	0	185	16.5
741	3	0	185	15.5
756.5	3	0	185	16.5
773.5	3	0	185	18.5
791	3	0	185	17.5
808.5	3	0	185	18.5
827	3	0	185	19.5
846.5	3	0	185	20.5
866.5	3	0	185	20.5
887.5	3	0	185	22.5
910.5	3	0	185	24.5
935.5	3	0	185	26.5
962.5	3	0	185	28.5
992	3	0	185	31.5
1025	3	0	185	35.5
1063	3	0	185	41.5
1108.5	3	0	185	50.5
1391.62	3	0	185	516.8

For DDA-SILAC samples, MaxQuant software platform (Cox and Mann, 2008) version 1.6.1.0 was used to process the raw files and the search was conducted against the *Homo sapiens* (released 14/05/19) protein database, using the Andromeda search engine (Cox et al., 2011). For the first search, peptide tolerance was set to 20 ppm while for the main search peptide tolerance was set to 4.5 ppm. Isotope mass tolerance was 2 ppm and maximum charge set to 5. Digestion mode was set to specific with trypsin and LysC allowing maximum of two missed cleavages. Carbamidomethylation of cysteine was set as fixed modification. Oxidation of

methionine was set as variable modification. Multiplicity was set to 2, specifying Arg6 and Lys6 for heavy labels. Peptide and protein identifications were filtered to 1% FDR.

DIA-NN software platform (Demichev et al., 2020) version 1.8.1 was used to process the DIA raw files and search was conducted against the Uniprot one protein sequence per gene database (Proteome ID UP000005640, released in July, 2017). Precursor ion generation was based on the chosen protein database (automatically generated spectral library for the raw files used to generate a library) with deep-learning based spectra, retention time and IMs prediction. Digestion mode was set to specific with trypsin, allowing maximum of one missed cleavage. Carbamidomethylation of cysteine was set as fixed modification. Oxidation of methionine, and acetylation of the N-terminus were set as variable modifications. The parameters for peptide length range, precursor charge range, precursor m/z range and fragment ion m/z range as well as other software parameters were used with their default values. The precursor FDR was set to 1%. For SILAC TRAPP samples the parameters used were similar with some modifications according to a published protocol (Derks et al., 2023).

DDA data processing

For MaxQuant normalized dataset, normalized Heavy/Light ratios for each protein was obtained from the proteinGroups.txt file, Heavy/Light ratios were inverted to Light/Heavy. For Vsn and Cyclic Loess normalized dataset, Heavy and Light intensities of each peptide were obtained from the peptides.txt file, only peptides labelled as unique were used. Peptide intensities were normalized on NormalyzerDE (https://normalyzerde.immunoprot.lth.se/) with 10min light as Group1, 10min heavy as Group2, 1h light as Group3 and 1h heavy as Group4. Light/Heavy ratio for each peptide in the normalized dataset was calculated; missing values were replaced with “NA”. For every gene in each sample, the number of non-NAs was calculated; if the number is 1, the non-NA value was replaced by “NA”; otherwise the value was retained. Median value of all peptides belonging to a gene was calculated and used to represent the level of the protein in the sample. Proteins annotated as contaminants, reverse, or keratins were removed. Proteins for analysis were restricted to those represented by equal or more than 2 peptides and detected in at least 2 out of 3 replicates in both time points in non-treated controls (SILAC heavy). Proteins with 1 missing value in 3 replicates had the missing value replaced by the mean value of the other two replicates, while proteins with 2 or more missing values had all the 3 replicates replaced by 0. Proteins containing all blanks were removed. All NAs were then replaced by 0. At this point the datasets were named the “processed datasets”.

Differential expression analysis was performed with Light/Heavy ratios in 10min and 1h samples, using 2-sided T Test with Benjamini Hochberg multiple test correction. Significantly

changed proteins were defined as those with FDR<0.05 and absolute FC>1.4. GO term enrichment analysis was done on WebGestalt (Liao et al., 2019) using Molecular Function noRedundant database and a list of all detected proteins in the DDA dataset as reference.

DIA data processing

Intensity for each peptide was obtained from the “Precursor.Translated” column of the Report.tsv generated by DIA-NN. Peptides with Precursor.Id containing (SILAC-K-H) and/or (SILAC-R-H) were labelled as heavy and those without were labelled as light. Sample to which a peptide belongs was determined by the “Run” column of Report.tsv and concatenated with either H or L depending on whether the peptide is heavy or light. “Protein.Ids”, “Genes”, and “Precursor.Id” with (SILAC-K-L/H) and (SILAC-R-L/H) removed were concatenated to form the identity of each peptide. Together a long-formatted file containing 3 columns: Sample, Peptide, Intensity was formed. A matrix file was generated using Pivot Table in Excel with Sample as columns, Peptide as rows, and Intensity as values. The concatenated Peptide column was separated again. This matrix file was used as the input file on NormalyzerDE with 10min light as Group1, 10min heavy as Group2, 1h light as Group3 and 1h heavy as Group4. Datasets normalized by Median, Vsn, and Cyclic Loess were analyzed.

Light/Heavy ratio for each peptide in each sample was calculated, missing values were replaced with “NA”. For every gene in each sample, the number of non-NAs was calculated; if the number is 1, the non-NA value was replaced by “NA”; otherwise the value was retained. Median value of all peptides belonging to a gene was calculated and used to represent the level of the protein in the sample. Proteins for analysis were restricted to those represented by equal or more than 2 peptides and detected in at least 2 out of 3 replicates in both time points in non-treated controls (SILAC heavy). Proteins with 1 missing value in 3 replicates had the missing value replaced by the mean value of the other two replicates, while proteins with 2 or more missing values had all the 3 replicates replaced by 0. Proteins containing all blanks were removed. All NAs were then replaced by 0. At this point the datasets were named the “processed datasets”.

Differential expression analysis was performed with Light/Heavy ratios for proteins in 10min and 1h samples, using 2-sided T Test with Benjamini Hochberg multiple test correction. Significantly changed proteins were defined as those with FDR<0.05 and absolute FC>1.4. GO term enrichment analysis was done on WebGestalt using Molecular Function noRedundant database and a list of all detected proteins in the DIA dataset as reference.

Data imputation

The effect of imputation by Random Forest (using R 4.2.2 with missForest library version 1.5) and MIN*0.5 (replacement of missing values by 0.5 x the minimum value of the sample) on the normalized datasets (MaxQuant normalized for DDA and Cyclic Loess normalized for DIA) was analyzed on all proteins detected in non-treatment controls, and proteins enriched by UV-crosslinking and significantly changed (according to DIA) between 10min and 1h of treatment.

Acknowledgements

The authors thank Dr. Van Kelly for help in confirming the DDA/DIA correlations at both peptide and protein levels; and Dr. Nic Robertson and Dr. Aleksandra Helwak for assistance with mammalian cell cultures. This work was funded by a Wellcome Principal Research Fellowship to DT (222516). Work in the Wellcome Centre for Cell Biology is supported by core funding (203149).

References

Bresson, S., Shchepachev, V., Spanos, C., Turowski, T.W., Rappsilber, J., and Tollervey, D. (2020). Stress-Induced Translation Inhibition through Rapid Displacement of Scanning Initiation Factors. Molecular Cell *80*, 470-484.e478.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nature Biotechnology *26*, 1367-1372.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., and Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res *10*, 1794-1805.

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

Derks, J., Leduc, A., Wallmann, G., Huffman, R.G., Willetts, M., Khan, S., Specht, H., Ralser, M., Demichev, V., and Slavov, N. (2023). Increasing the throughput of sensitive proteomics by plexDIA. Nature Biotechnology *41*, 50-59.

Dudoit, S., Yang, Y.H., Callow, M.J., and Speed, T.P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Stat Sinica *12*, 111-139.

Kokla, M., Virtanen, J., Kolehmainen, M., Paananen, J., and Hanhineva, K. (2019). Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. BMC Bioinformatics *20*, 492.

Lai, A., Valdez-Sinon, A.N., and Bassell, G.J. (2020). Regulation of RNA granules by FMRP and implications for neurological diseases. Traffic *21*, 454-462.

Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic acids research *47*, W199-W205.

Olsen, J.V., Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007). Higher-energy C-trap dissociation for peptide modification analysis. Nature Methods *4*, 709-712.

Pino, L.K., Baeza, J., Lauman, R., Schilling, B., and Garcia, B.A. (2021). Improved SILAC Quantification with Data-Independent Acquisition to Investigate Bortezomib-Induced Protein Degradation. J Proteome Res *20*, 1918-1927.

Rappsilber, J., Mann, M., and Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nature Protocols *2*, 1896-1906.

Shchepachev, V., Bresson, S., Spanos, C., Petfalski, E., Fischer, L., Rappsilber, J., and Tollervey, D. (2019). Defining the RNA interactome by total RNA-associated protein purification. Molecular systems biology *15*.

Shih, J.-W., Wang, W.-T., Tsai, T.-Y., Kuo, C.-Y., Li, H.-K., and Wu Lee, Y.-H. (2011). Critical roles of RNA helicase DDX3 and its interactions with eIF4E/PABP1 in stress granule assembly and stress response. Biochemical Journal *441*, 119-129.

Stekhoven, D.J., and Bühlmann, P. (2011). MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics *28*, 112-118.

Välikangas, T., Suomi, T., and Elo, L.L. (2018). A systematic evaluation of normalization methods in quantitative label-free proteomics. Briefings in bioinformatics *19*, 1-11.

Warner, J.R., and McIntosh, K.B. (2009). How Common Are Extraribosomal Functions of Ribosomal Proteins? Molecular Cell *34*, 3-11.

Willforss, J., Chawade, A., and Levander, F. (2019). NormalyzerDE: Online Tool for Improved Normalization of Omics Expression Data and High-Sensitivity Differential Expression Analysis. Journal of Proteome Research *18*, 732-740.

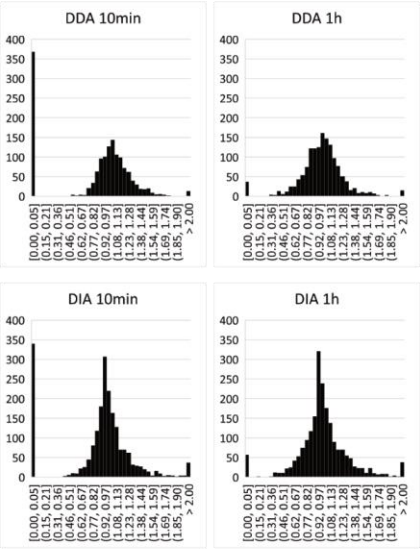
Wiśniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. Nature Methods *6*, 359-362.

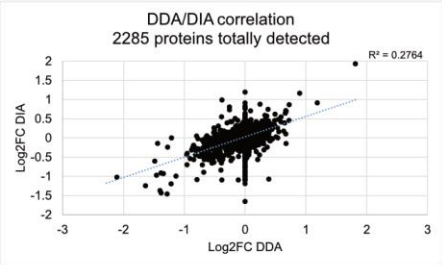
Zhang, M.J., Pisco, A.O., Darmanis, S., and Zou, J. (2021). Mouse aging cell atlas analysis reveals global and cell type-specific aging signatures. eLife *10*, e62293.

Zurla, C., Lifland, A.W., and Santangelo, P.J. (2011). Characterizing mRNA Interactions with RNA Granules during Translation Initiation Inhibition. PLOS ONE *6*, e19727.

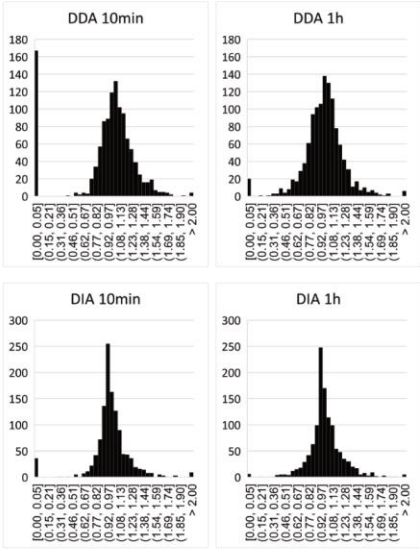
Supplementary data

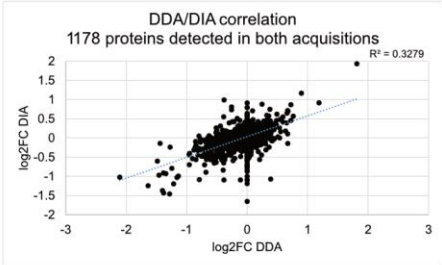
All detected proteins



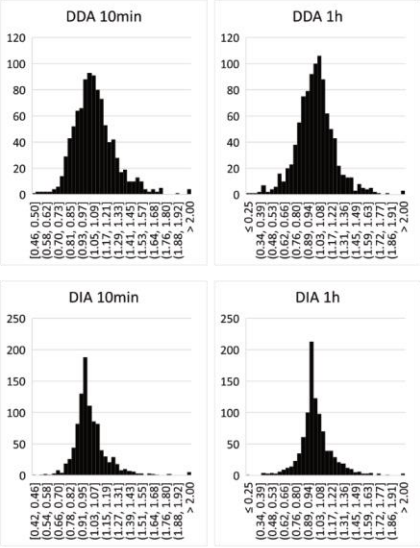


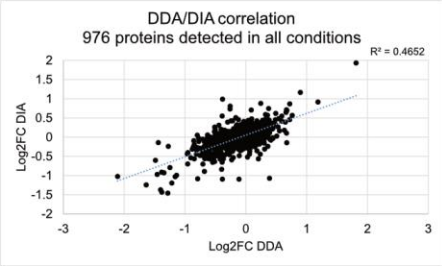
1178 proteins detected in both acquisitions



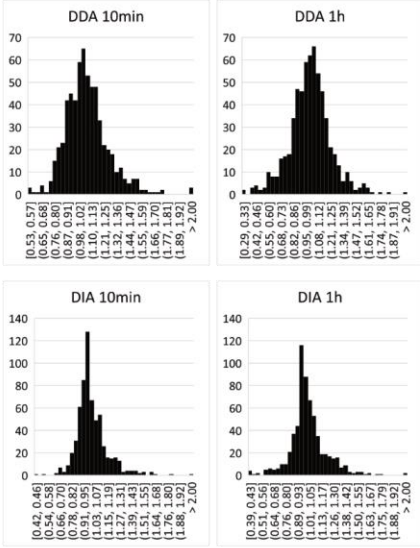


976 proteins detected in all conditions





630 proteins detected in all conditions, and enriched by crosslinking



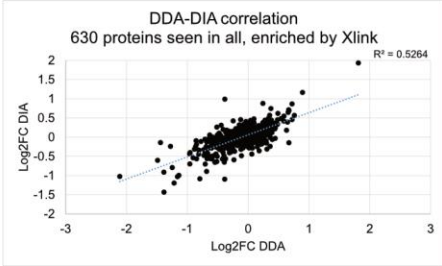


Figure S1. Distribution of Light/Heavy ratios and correlation of Log2FC between 10min and 1h time points, calculated from normalized processed datasets. DDA dataset was normalized by MaxQuant, DIA by Cyclic Loess. All detected proteins, proteins detected in both acquisitions, proteins detected in all conditions, and proteins detected in all conditions and enriched by UV-crosslinking were separately plotted.

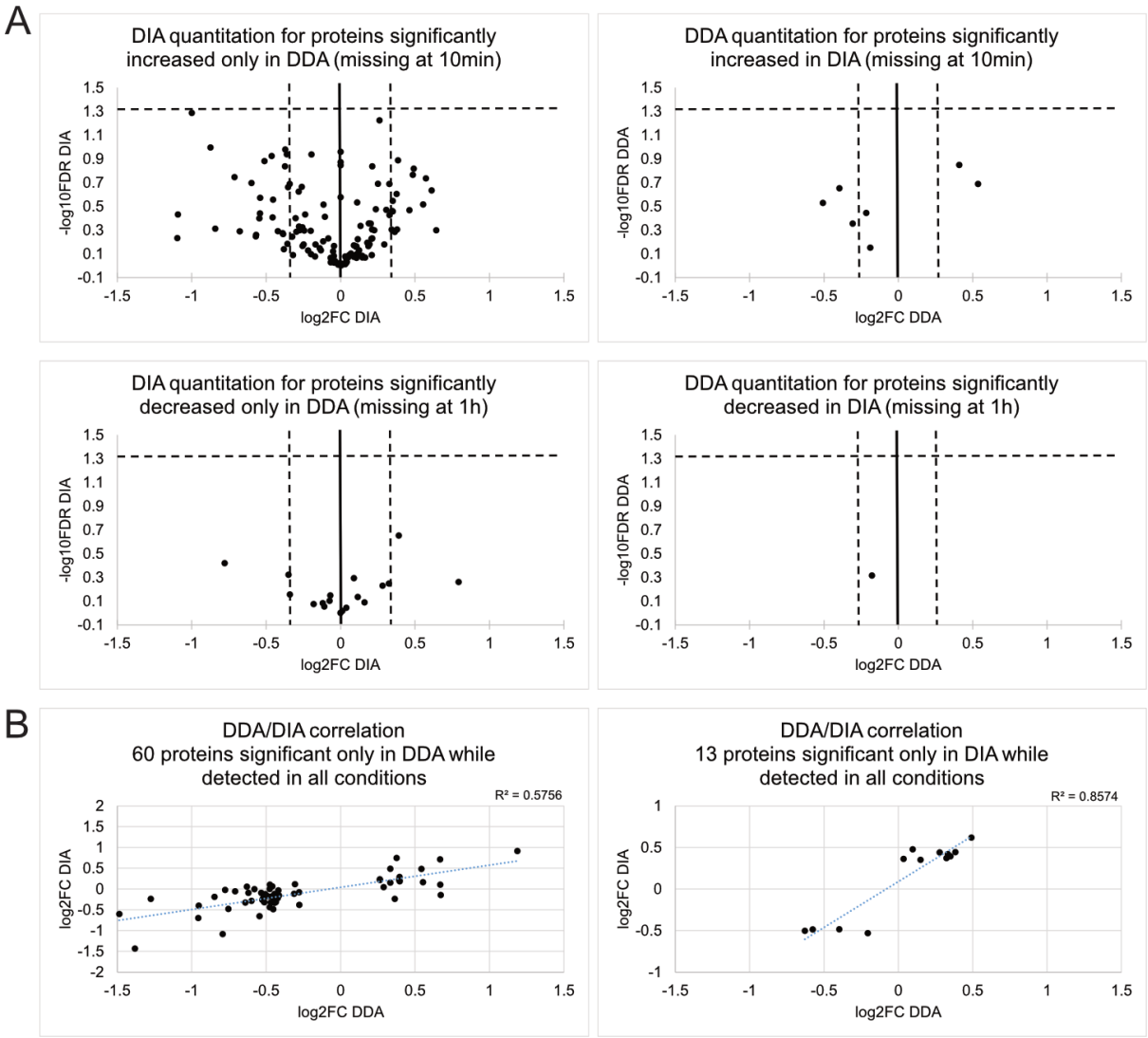


Figure S2. DDA and DIA datasets were similar only for proteins detected in all conditions. (A) DDA or DIA quantitation for proteins detected at both time points while significantly changed according to the other acquisition due to missing at one time point. (B) DDA/DIA correlation for proteins significantly changed in either dataset and detected in all conditions (i.e. both time points in both acquisitions).

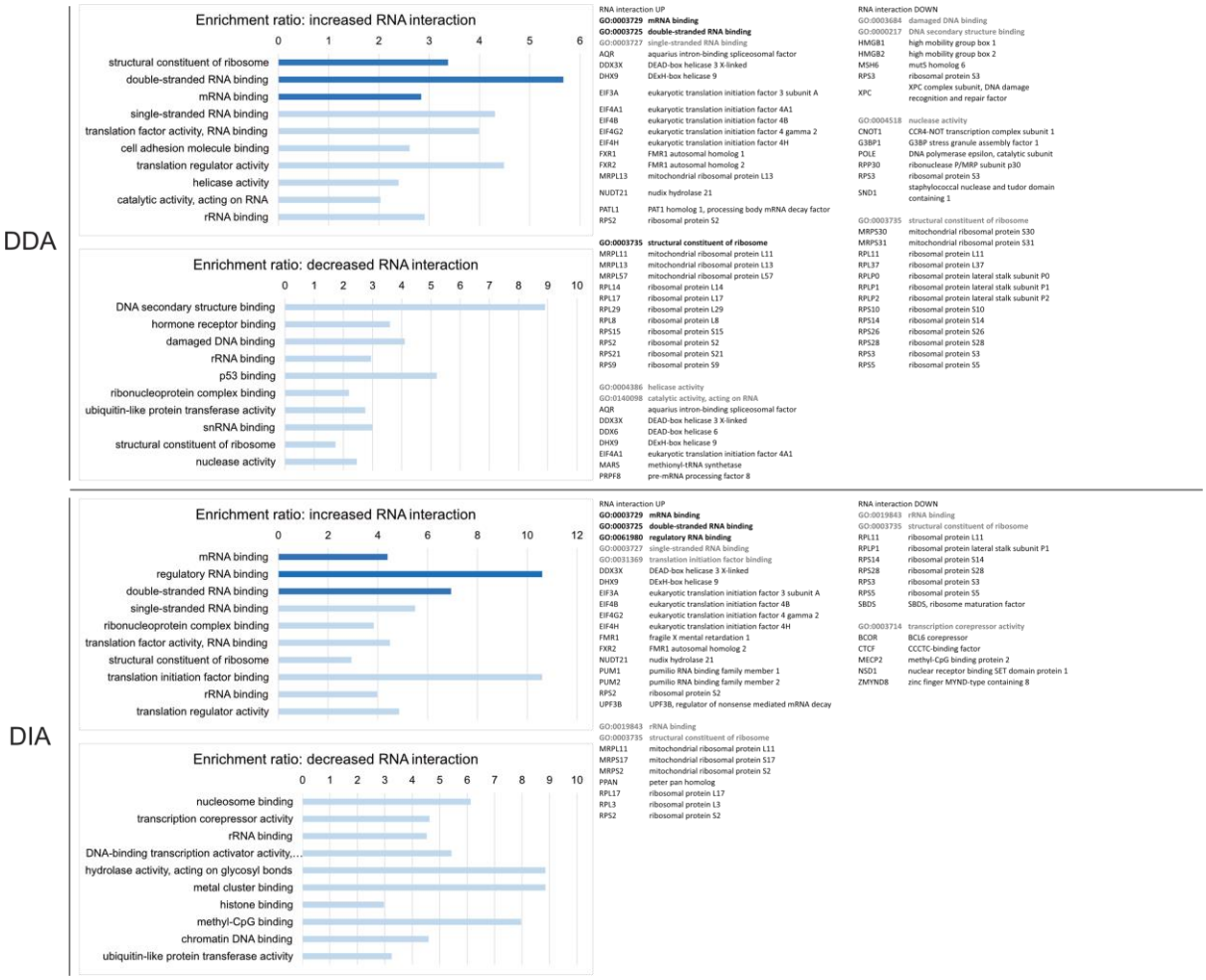


Figure S3. Top-ten enriched GO terms in molecular function for proteins with increased or decreased RNA interaction from 10min to 1h according to the DDA or DIA dataset, selected using 2-sided T Test P value <0.05 and absolute fold-change > 90 percentile. Selected gene sets in the right panel are possibly related to the stress-induced RNA translation shutdown. Dark bar, FDR<0.05; Pale bar, FDR>0.05. GO terms were sorted by FDR in ascending order, or by enrichment ratio if all FDRs were 1.