Integrated View of Baseline Protein Expression in Human Tissues using public Data Independent Acquisition datasets

Ananth Prakash[1,*], Andrew Collins[2], Liora Vilmovsky[1], Silvie Fexova[1], Andrew R. Jones[2,*], Juan Antonio Vizcaino[1,*]

Affiliations

[1] European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge. CB10 1SD. United Kingdom

[2] Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, L69 7ZB, United Kingdom

* Corresponding authors

Dr. Ananth Prakash. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. Phone: + 44 (0) 1223 492610. Email: ananth@ebi.ac.uk

Prof. Andrew R. Jones. Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 7ZB, United Kingdom. Email: andrew.jones@liverpool.ac.uk

Dr. Juan Antonio Vizcaíno. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. Phone: + 44 (0) 1223 492686. Email: juan@ebi.ac.uk

Abstract

The PRIDE database is the largest public repository of mass spectrometry-based proteomics data and currently stores more than 40,000 datasets covering a wide range of organisms, experimental techniques and biological conditions. During the past few years, PRIDE has seen an increase in the amount of submitted Data-Independent Acquisition (DIA) proteomics datasets, in parallel with the trends in the field. This provides an excellent opportunity for large scale data reanalysis and reuse.

We have systematically reanalysed 15 public label-free DIA datasets across various healthy human tissues, to provide a state-of-the-art view of the human proteome in baseline conditions (without any perturbations), coming from DIA datasets. We computed baseline protein abundances and compared them across various tissues, samples and datasets. Our second aim was to make a comparison of the protein abundances obtained from previous analyses of human baseline Data-Dependent Acquisition (DDA) datasets. Results were heterogeneous. On one hand, we observed a good correlation across some tissues, especially in liver and colon. On the other hand, weak correlations were however found in others such as lung and pancreas. It is likely this reflects mostly differences the sample preparation and processing for datasets derived from these tissues, as opposed to fundamental differences between DDA and DIA proteomics. The reanalysed results including protein abundances and curated metadata are made available to view and download from the resource Expression Atlas.

Introduction

The most popular quantitative approach for Mass Spectrometry (MS)-based proteomics has historically been data-dependent acquisition (DDA) bottom-up proteomics. However, Data Independent Acquisition (DIA) approaches[1] have matured enormously, thanks to multiple technical developments in e.g. sample preparation, improvements in instrumentation and computational data analysis approaches[2]. In contrast to DDA, where only the most intense peptide ions are measured, in DIA fragmentation products are generated from every peptide ion that is sampled in the MS^1 scans. As a result, a DIA analysis approach can potentially provide a more comprehensive quantitative view on the proteome when compared to DDA approaches.

There are two main approaches of DIA proteomics data analysis: spectrum-centric and peptide-centric. Spectrum-centric methods follow the DDA general approach, by generating "pseudo-MS/MS spectra" where fragment ions are associated with the precursor ion from which they are most likely derived. The pseudo spectra can be analysed like DDA data, for instance through typical sequence database searching tools. Peptide-centric methods are currently much more widely used and rely on deciding in advance which peptidoforms may be present in the sample, by using a spectral library[2]. There are three main approaches for generating the spectral libraries that can be used for DIA analysis. First, it is possible to perform a deeply fractionated DDA analysis using the same instrumentation as will be used by the DIA experiments, to generate an *experimentally matched* spectral library. Second, there are some publicly available libraries, created from a given type of samples, for instance a "pan-species" library, assembled from multiple MS runs[3]. The third approach consists of the use of *in silico* generated spectral libraries, generated using artificial intelligence (AI)-trained models, having learnt retention times and peptide intensities from past DDA datasets[2].

In parallel to other developments, for the last decade or so, the proteomics community has learnt to follow and implement open data practices. The PRIDE database[4], one of the founding members of the ProteomeXchange consortium[5], is the most popular proteomics data repository worldwide. The availability of extensive public proteomics datasets in PRIDE and other ProteomeXchange resources has triggered multiple applications, including meta-analysis studies including some studies applying different AI approaches[6,7]. Additionally, by systematically reanalysing public datasets, original findings can be updated, confirmed and/or

strengthened[8]. Moreover, novel insights beyond the scope of the original studies can be obtained through alternative reanalysis strategies, for instance in the case of proteogenomics approaches[9].

In this context, we are facilitating comparison of transcriptomics and proteomics data in the resource Expression Atlas[10], so that protein expression abundance information is made accessible in the long-term to life scientists, including those not experts in proteomics. We have already performed combined analyses of baseline (i.e. without any perturbation) protein expression studies of public DDA experiments coming from human[11], and from mouse and rat[12] and pig tissues[13], as model organisms. Additionally, we have also performed combined analyses of DDA experiments generated from cell lines and tumour tissue[14], and more recently, a study focused on colorectal cancer related datasets as an approach for biomarker discovery[15]. In all cases, protein abundance results have been made available through Expression Atlas. There, protein abundance data can be visualised together with gene expression information, although rarely from the same samples.

However, in parallel with the trends in the field, an increasing fraction of data deposited in PRIDE and in other ProteomeXchange resources comes from DIA approaches. In this context, a few years back, we carried out a pilot study including ten DIA human datasets including mostly cell lines (all of them, SWATH-MS experiments from SCIEX instruments)[16], using an analysis approach based on the use of the "pan-human" spectral library[17].

Here, we report the reanalysis and integration of 15 public label-free DIA human baseline tissue datasets including 178 healthy/normal control samples. The objective of this study is two-fold: on one hand, we aim to facilitate access to protein abundance data from baseline human tissues coming from state-of-the-art DIA approaches. On the other, we want to study their similarity with analogous protein abundance data generated using DDA approaches. Unlike our previous study involving DIA datasets[16], we used *in silico* generated spectral libraries for the analysis. The protein abundance results, as in our previous studies, have been incorporated into Expression Atlas. Additionally, we made a comparison of the expression values coming from DIA datasets and the protein abundance results from previous DDA studies including human baseline tissues[18,19], for instance ProteomicsDB data and also our previous combined analysis of DDA public human datasets[11].

Methods

Datasets

Public MS-based human DIA proteomics datasets were queried from the PRIDE database in January 2023. Out of the 288 datasets that were available, several datasets were selected for downstream reanalysis. The selection criteria included: (i) non-enriched samples, only coming from tissues (not including cell lines or fluid samples such as blood/plasma, cerebrospinal fluid, etc.); (ii) datasets from SCIEX (TripleTOF series) and Thermo Fisher Scientific (QExactive series and Orbitrap) instruments; (iii) availability of experimental metadata to link samples to their respective data files; and (iv) in the case of datasets from SCIEX instruments, only datasets that had both .wiff and .scan files deposited were considered (this is a requirement that was added for all SCIEX data submissions to PRIDE in 2021, but there are some previously submitted datasets that do not comply with this requirement). Where enough experimental metadata was not made available in the initial data submissions and/or in the corresponding publications we contacted their respective authors. At the end, we selected 15 datasets for reanalysis (Table 1).

Metadata was annotated using Annotare[20] and saved as Investigation Description Format (IDF) and Sample and Data Relationship Format (SDRF) files. IDF comprises description about dataset experiment protocol, contact information of investigators, and publication details, while SDRF contains information of samples including case/control, donor age, gender, treatment conditions, experimental factors, etc. Both IDF and SDRF files were integrated into Expression Atlas.

Proteomics raw data processing

To process spectral data on a Linux (Rocky Linux 8) platform, the vendor raw files were first converted to mzML format using conversion tools. For Thermo Fisher Scientific instruments (.raw), Thermorawfileparser[21] 1.4.2 was used with default settings. For SCIEX instruments (.wiff and .scan) ProteoWizard's msConvert[22] conversion tool was used with the “peakPicking vendor” option enabled.

As a target database, the UniProt human ‘one protein sequence per gene’ database (UP000005640, May 2023) with 20,838 sequences was used to which cRAP contaminant sequences[23] (245 sequences) were added. The target database was used in generating an *in*

silico spectral library using DIA-NN[24] (version 1.8.1) with default parameters. We also used an entrapment database[25,26] wherein *Arabidopsis thaliana* UniProt 'one protein sequence per gene' database (UP000006548, December 2022) with 41,621 sequences, which was added to the target human protein search database. Similarly, an *in silico* spectral library of the entrapment database was generated with default parameters using DIA-NN.

To improve run time performance, we designed our analysis by separating a conventional DIA-NN[24] run into two phases so that it could be reanalysed efficiently on a Linux high performance computing (HPC) platform, using the Slurm job scheduler[27]. Prior to performing the identification for any given dataset, a calibration run was first performed to identify the search tolerances. By default, DIA-NN will auto-detect the tolerances using the first sample in the batch. This leads to issues when running DIA-NN in parallel as each sample will be run with different tolerance settings. To solve this, we ran DIA-NN using default auto-detect settings on a single sample against the target database. The output of this calibration run provided the MS1, MS2, and scan window values that would be used for the full run.

Using the tolerance values identified above, we then employed a multi-node approach to process all raw files in parallel across the HPC cluster. This parallel process allowed us to reduce the run time of DIA-NN's first phase to the runtime of a single file - a speed-up factor directly proportional to the number of files in each dataset. In this first phase, each MS run was searched against UniProt human with an *Arabidopsis thaliana* entrapment database and the *in silico* spectral library mentioned above. During this phase, cross-run normalisation and MaxLFQ-based protein quantification were disabled (--no-norm, --no-maxlfq), match between runs was not performed, and main report not generated (--no-main-report). Q-value was set to 0.01, peptide cleavage sites assigned to trypsin and the rest of the parameters were set to default. At the completion of the first phase, protein quantification (.quant files) from individual runs were saved in a temporary directory. During the second phase, the existing .quant files generated from the first phase (--use-quant) were collated and a cross-run analysis performed on a single node, wherein cross-run normalisation, match-between runs, MaxLFQ-based protein quantification were performed and main report files, including protein groups, gene groups, unique genes and precursors matrices, were generated.

In a conventional DIA-NN run, each input file is sequentially analysed against the spectral library for protein identification and quantification. By leveraging multi-node processing in the

first phase wherein each file was simultaneously analysed and then collating the .quant files, this significantly reduced the total run time by approximately 50% (Supplementary figure 1).

Postprocessing

The report.tsv output file from DIA-NN was used as the basis for post-processing. Contaminants were removed along with mappings to more than one protein or gene identifier. In each MS run we removed entries with fewer than two unique peptide sequences per protein and the abundances of proteins were aggregated using their median values within each MS run in a dataset. DIA-NN outputs abundances as label-free quantification (LFQ) values. We converted LFQ to intensity based absolute quantification (iBAQ) values by normalising the LFQ abundances with the theoretical number of tryptic peptides of each canonical protein. As explained in previous publications[11-13], we represent protein abundances as abundances of their respective parent gene names, which we will use as equivalent to 'canonical proteins' as described in UniProt (https://www.uniprot.org/help/canonical_and_isoforms). For ease of comparison of protein abundances across tissues and datasets with likely large batch effects between different experiments, iBAQ protein abundances were converted into ranked bins[11]. Briefly, iBAQ abundances were numerically sorted and grouped into five bins of equal size. Proteins in bin 1 are of lowest abundance and those in bin 5 are highly abundant. The heatmap of samples were generated using binned abundance values. Post-processing was performed using R scripts.

Integration into Expression Atlas

Post-processed results comprising protein abundances, expressed as their canonical gene identifiers (proteins were mapped to gene identifiers by DIA-NN), were integrated into Expression Atlas[28] along with metadata files (IDF and SDRF) and a quality assessment summary. Expression Atlas dataset identifiers (E-PROT) for each dataset are shown in Table 1.

Comparison of baseline protein abundances generated using DDA data

Ensembl gene identifiers of the respective canonical proteins were used for comparing proteins identified between this study and previous DDA studies performed in human baseline tissues. Ensembl gene identifiers and normalised iBAQ protein abundance values in supplementary table 2 available from[11] was gathered for tissues between two studies for comparison. The

iBAQ protein abundances of DIA datasets were compared with the fraction of total normalised iBAQ protein abundances from DDA datasets.

Additionally, normalised protein intensities from ProteomicsDB[18] were queried for tissues that were in common in our study (8 tissues). Values were obtained using the ProteomicsDB Application Programming Interface (downloaded in April 2022). For different tissue samples we aggregated the normalised intensities using the median of their respective tissues. The intensities were log2 normalised and compared.

Furthermore, protein abundances calculated across various baseline human tissues using the TMT-labelling method were obtained from[19] (Supplementary file 'NIHMS1624446-supplement-2', sheet: 'C protein normalized abundance'). Protein abundances of the respective tissues measured across different TMT channels and runs were aggregated using the median and log2 transformed. Different tissue samples from esophagus, heart, brain and colon were aggregated into their respective tissues. Binned protein abundances of the samples were used for calculating Pearson's correlation within each tissue. Brain is the only tissue in this study which is available in multiple datasets. The correlation of protein abundances between brain samples from multiple datasets were calculated and median correlation value presented. Correlations were calculated using the R programming language.

Comparison of missing values between DDA and DIA datasets

To compare the proportion of missing protein identifications between DDA and DIA techniques we used the protein abundances from human DDA samples calculated in our previous study[11]. Protein abundances from DDA samples were represented as their canonical gene identifiers as described for DIA samples above. To compare the completeness of protein detection among samples of their respective tissues within a dataset, first for a given tissue we calculated the total number of missing abundances (NA) among its samples across all genes and then normalised it by the total number of observations as shown in the following equation below.

$$Fm = \frac{\sum_{1}^{n}(NA)}{(g * n)}$$

where 'Fm' is the fraction of missingness, 'n' is total number of samples of a particular tissue within a dataset, 'g' is the total number of genes identified in those respective tissue samples

within the dataset, 'NA' is the missing abundance of a canonical gene. Where a dataset had more than one tissue, the 'Fm' was calculated individually for each group of tissue samples and then a median of 'Fm' over all tissues was calculated for the dataset.

False Discovery Rate analysis using entrapment database

All datasets were analysed individually. To estimate the protein False Discovery Rate (FDR) across datasets we used the *Arabidopsis thaliana* entrapment search database as explained above. From the protein group matrix report output file from DIA-NN, we treated the protein groups that had gene identifiers belonging exclusively to *Arabidopsis thaliana* as decoys and the rest of the protein groups were regarded as targets. The distribution of number of datasets in which all decoys and targets were found in common were computed and the FDR for a target protein identified across a number of datasets was calculated using the equation-

$$FDR = \frac{Number\ of\ Arabidopsis\ decoys\ * Dn}{Number\ of\ targets}$$

where 'Dn' = 0.5, is the database normalising factor i.e., size of UniProt human one protein per gene set database (20,838 sequences) normalised by UniProt *Arabidopsis thaliana* one protein per gene set database (41,621 sequences).

Results

DIA human proteomics datasets reanalysis

We selected 15 DIA proteomics datasets, which represented the available range in the public domain of human tissues in healthy/baseline conditions. The spectra from these datasets were acquired either using SCIEX or Thermo Fisher Scientific instruments, as explained in 'Methods'. In total there were 1,361 MS runs from all samples in aggregated datasets, which included 356 MS runs from 178 healthy or normal control samples. Each dataset was analysed individually and post-processed. The complete list of datasets along with the number of peptides and proteins identified and quantified are shown in Table 1. Protein abundances computed only from healthy or normal tissue samples are discussed in this manuscript. However, abundances from all samples along with a quality assessment summary and sample metadata are made available in Expression Atlas (https://www.ebi.ac.uk/gxa/) to view and download. The overall data reanalysis protocol is summarised in Figure 1.

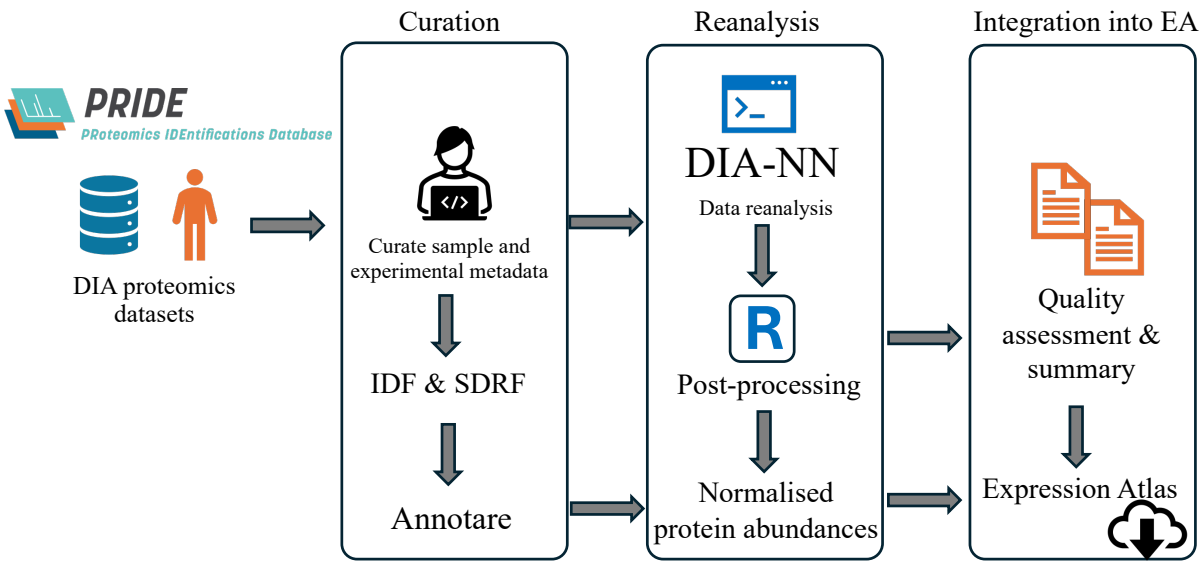


Figure 1. Overview of DIA datasets reanalysis pipeline. EA = Expression Atlas

Expression Atlas accession number	PRIDE proteomics dataset identifier	Tissue	Mass spectrometer	Fractionation	No. of MS runs*	Number of samples*	Number of PSMs*	Number of unique peptides †	Number of unique genes (canonical proteins) ‡
E-PROT-137	PXD031419 [29]	Brain	TripleTOF 5600	No	75	19	1,424,053	49,709	5,069
E-PROT-138	PXD012254 [30]	Colon	Q Exactive	No	19	19	871,632	53,273	5,567
E-PROT-139	PXD001506 [31]	Duodenum	TripleTOF 5600	No	6	6	104,435	14,424	1,955
E-PROT-140	PXD001764 [32]	Esophagus	TripleTOF 5600	No	3	3	58,131	15,545	2,546
E-PROT-141	PXD019594 [33]	Heart	TripleTOF 5600	No	10	10	93,177	10,456	1,131
E-PROT-142	PXD025705 [34]	Liver	Q Exactive Plus	No	33	11	3,099,160	85,833	6,403
E-PROT-143	PXD004684 [35]	Lung	Q Exactive	No	8	4	155,588	25,179	2,905
E-PROT-144	PXD032076 [36]	Pancreas	Orbitrap Fusion	No	25	25	417,648	27,139	3,507
E-PROT-145	PXD025431 [37]	Skin	Q Exactive HF	No	10	10	437,823	48,343	4,134
E-PROT-146	PXD002732 [38]	Thyroid	TripleTOF 5600	No	8	8	160,326	21,965	2,999
E-PROT-147	PXD022872 [29]	Brain	TripleTOF 5600	No	71	11	1,432,467	50,605	5,119
E-PROT-148	PXD033060 [39]	Brain	TripleTOF 5600	Yes	17	17	621,054	44,306	4,987
E-PROT-149	PXD018830 [40]	Breast epithelium	Q Exactive	No	4	4	113,729	27,855	4,159
E-PROT-150	PXD039665 [41]	Skin	Orbitrap Exploris 480	No	13	13	124,942	13,463	1,759
E-PROT-151	PXD034908 [42]	Skeletal muscle	TripleTOF 6600	No	54	18	1,174,230	18,406	1,413
Total	**15 datasets**	**12 tissues**			**356 MS runs**	**178 Samples**			

Table 1. Summary of protein identifications among datasets. * Healthy/normal control samples only. † Post-processed results from control samples, wherein each peptide has at least 2 PSMs mapping to it. ‡ Post-processed results from control samples, where a gene has at least two unique peptide mappings. Protein expression data in Expression Atlas can be accessed using the link https://wwwdev.ebi.ac.uk/gxa/experiments/E-PROT-XXX/Results , where XXX should be replaced with an E-PROT identifier.

Protein coverage across samples

From each sample analysed in a dataset, we kept those proteins which were identified by at least two peptides. Those mapping to more than one gene identifier were also filtered out for downstream analysis. We identified a total of 9,299 canonical proteins, of which 521 proteins were observed in all 12 tissues and 2,679 were only found in one tissue (Supplementary table 1). We identified the largest number of proteins in liver (6,401, 68.8%) followed by brain (6,018, 64.7%) and colon (5,564, 59.8%). The fewest proteins were identified in heart (1,131, 12.2%), skeletal muscle (1,412, 15.2%) and duodenum (1,955, 21.0%) (Figure 2A). We observed significant variations in the protein abundances across each tissue. Figure 2B shows the aggregated (median) protein abundances of each tissue over several samples. Among datasets, PXD025705 (6,401, 68.8%) from liver had the largest number of proteins identified, followed by PXD012254 (5,564, 59.8%). As seen in Table 1, each tissue is represented by one dataset, except brain which is represented by 3 datasets: PXD022872 (5,119, 55.0%), PXD031419 (5,069, 54.5%), and PXD033060 (4,987, 53.6%) (Figure 2C). The protein abundances in each dataset vary reflecting those observed for every tissue except brain (Figure 2D).

Protein abundance comparison across tissues

To compare protein abundances across different tissues we first transformed the LFQ protein abundance values obtained from DIA-NN to iBAQ as explained in the 'Methods' section. To make comparison easier across datasets and tissues we grouped iBAQ values equally into five categorical bins. Proteins within bin 1 are of lowest abundances and those in bin 5 are of highest abundances. These binned abundances are available in Supplementary table 2.

We carried out pairwise comparison across all samples (n=178) using binned protein abundances (Figure 3). Within brain samples we observed moderate correlation of protein abundances (median R^2=0.40), across the three datasets. In comparison, the correlation of protein abundances in DDA brain samples from our previous study[11] was slightly higher (median R^2=0.61). It is important to note that the number of DDA brain samples was much larger (n=339) than those in analysed here by DIA approaches (n=47).

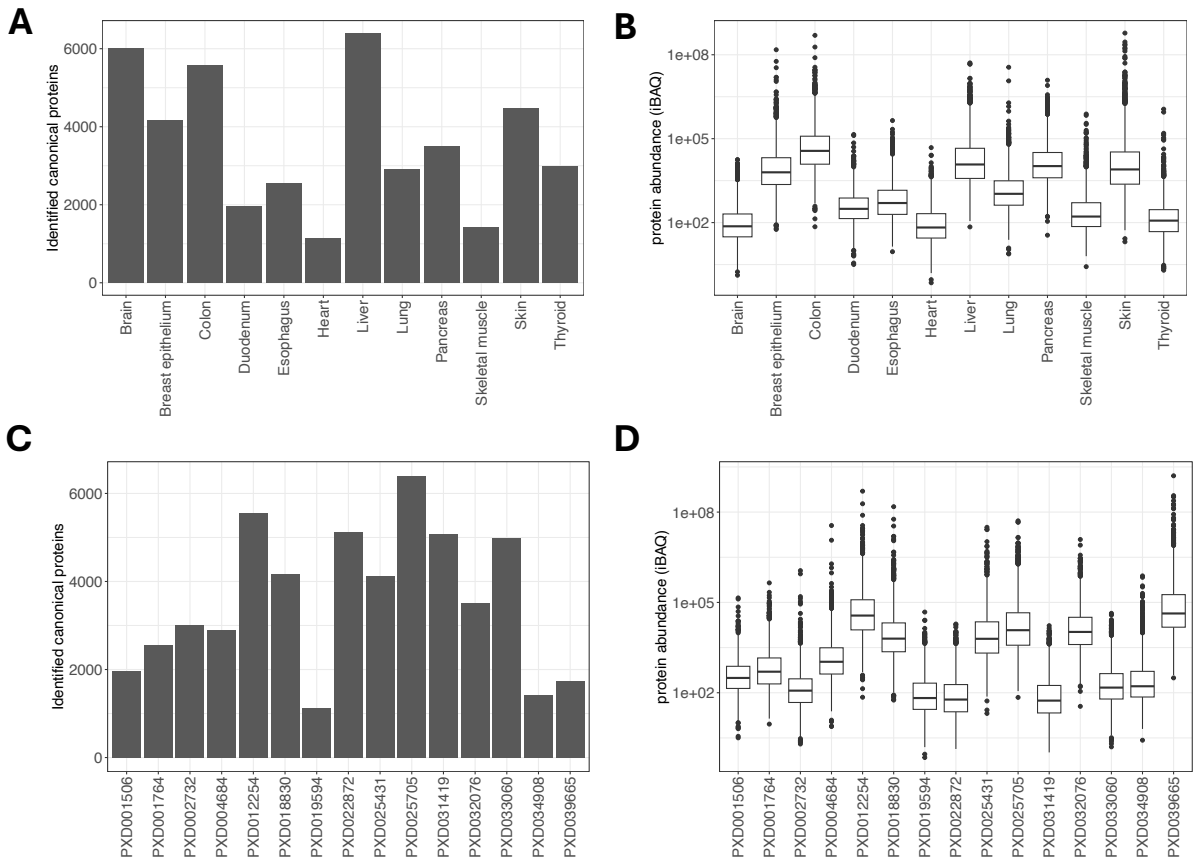


Figure 2. (A) Number of canonical proteins identified across different tissues. (B) iBAQ protein abundances of canonical proteins across different tissues. (C) Number of canonical proteins identified across different datasets. (D) iBAQ protein abundances of canonical proteins identified across different datasets.

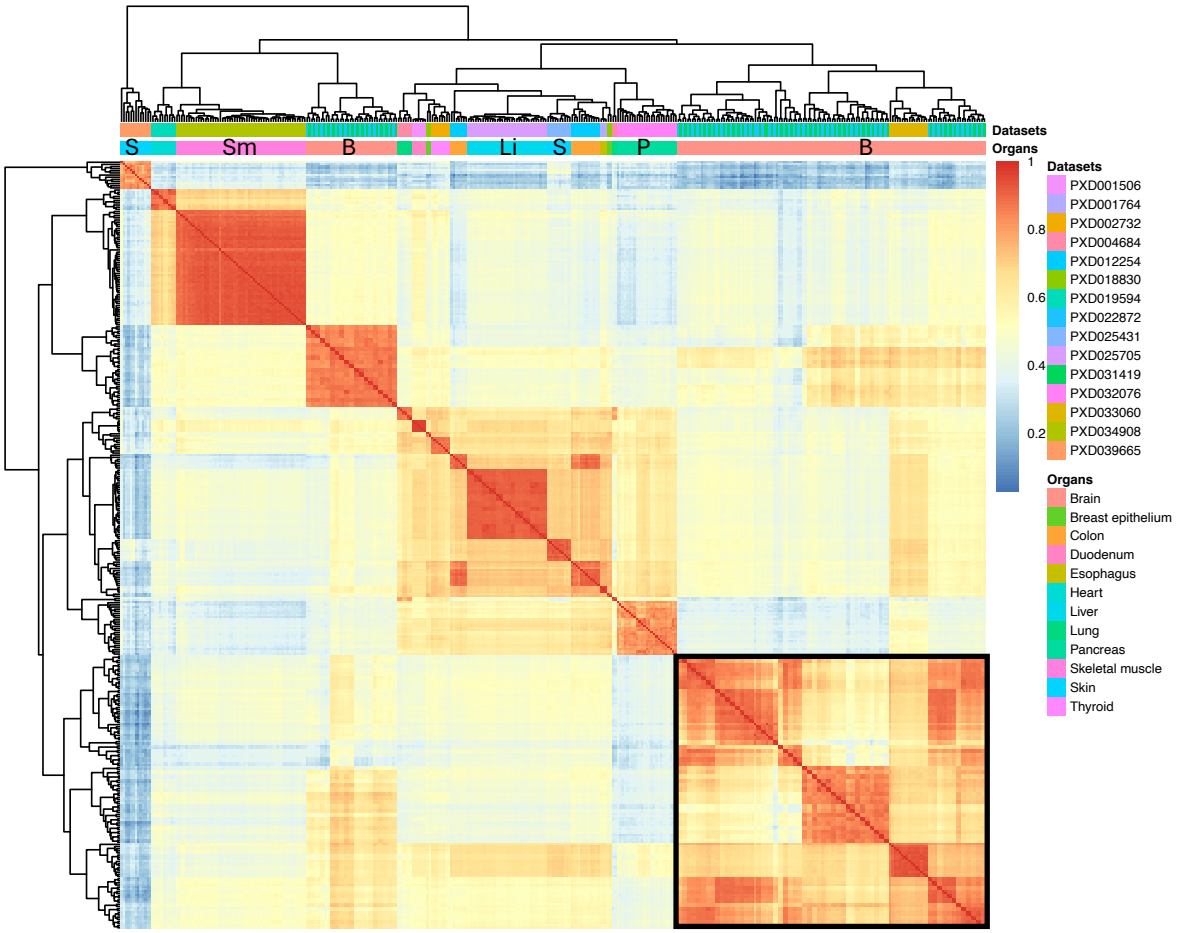
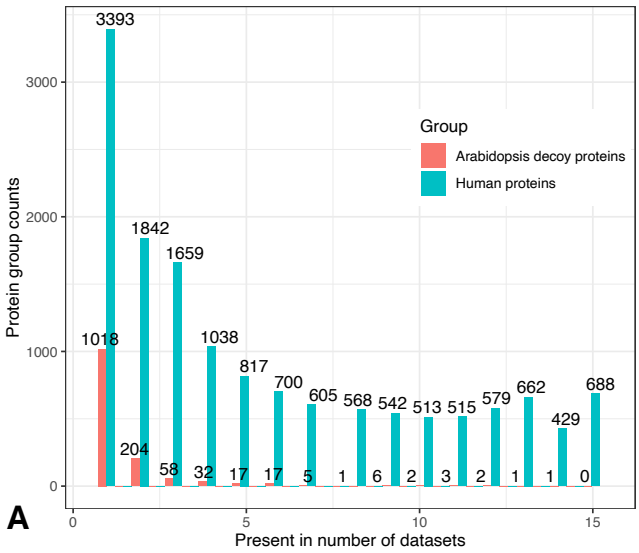


Figure 3. Heatmap of binned protein abundances across all samples between various tissues and datasets. Brain samples clustered together are highlighted in black border. S: Skin, Sm: Skeletal muscle, B: Brain, Li: Liver, P: Pancreas.

Arabidopsis thaliana entrapment analysis

We designed our study to reanalyse each dataset individually, as in previous studies. The protein False Discovery Rate (FDR) was set to 1% with the Benjamini-Hochberg correction for multiple hypothesis testing within each dataset, and although this is a stringent filter, but as protein abundances from multiple datasets are collated at the end it could result in a protein FDR much greater than 1%. To check this, we used an entrapment target sequence approach, wherein *Arabidopsis thaliana* protein sequences were added to the target human sequence database. The results from the reanalysis of datasets were searched against the entrapment database for protein groups comprising exclusively of canonical *Arabidopsis thaliana* proteins (false targets). We found a total of 1,367 decoy protein groups across all 15 datasets including 1,018, which were observed in only one dataset (Figure 4). The largest fraction of decoy hits was found in dataset PXD039665 (4.3%) and the lowest in dataset PXD012254 (1.1%). From

the number of decoys detected in common across all datasets, we estimated a protein FDR of less than 1% when proteins were observed in at least 7 datasets and an FDR of less than 5% when proteins were observed in at least 3 datasets. In all cases, *Arabidopsis thaliana* proteins with evidence of at least 2 peptides were considered. We concluded that the protein FDR values were comparable to our previous human study and appropriate for the objectives of this work. We recommend that DIA analyses use an entrapment approach such as this, to ensure that there is robust control of FDR, following any post-processing or merging steps that may be done.



B

Present in number of datasets	# of Arabidopsis decoys	# of Human proteins	FDR
1	1018	3393	0.150 (15.0%)
2	204	1842	0.055 (5.5%)
3	58	1659	0.017 (1.7%)
4	32	1038	0.015 (1.5%)
5	17	817	0.010 (1.0%)
6	17	700	0.012 (1.2%)
7	5	605	0.004 (0.4%)
8	1	568	0.000 (0.0%)
9	6	542	0.005 (0.5%)
10	2	513	0.001 (0.1%)
11	3	515	0.002 (0.2%)
12	2	579	0.001 (0.1%)
13	1	662	0.000 (0.0%)
14	1	429	0.001 (0.1%)
15	0	688	0 (0%)

Figure 4. (A) Distribution of decoy and target protein groups present across all datasets identified from using the *Arabidopsis thaliana* sequence entrapment search database. (B) FDRs of target proteins present in common in number of datasets, the calculation of FDR is described in detail in the 'Methods' section.

Protein abundance comparison between the DIA datasets and previous DDA studies

We had previously reanalysed 24 public proteomics dataset representing 31 healthy human tissues[11] coming from DDA studies. We first compared the number of canonical proteins identified between datasets analysed using DDA and DIA. We had previously identified a total of 13,071 canonical proteins from DDA studies and 9,299 canonical proteins from the reanalysis of the DIA datasets in this study, all in baseline conditions. When comparing DDA and DIA results, we found 8,449 (64.6%) common canonical proteins obtained from the two groups of datasets, with 4,621 (35.3%) canonical proteins identified only in DDA datasets and 853 (9.1%) canonical proteins identified only in DIA datasets (Supplementary table 3, Supplementary figure 2). It is important to note here that in comparison to DIA there were

many more MS runs analysed in DDA datasets, and additionally many DDA datasets were fractionated increasing the depth of the study. By analysing the missingness of protein abundance values among samples and comparing it with samples in DDA datasets from our earlier study[11] we observed similar trends in the fraction of missing values among samples analysed by DIA and DDA techniques, implying that DIA does not confer substantial advantages in terms of avoiding missing values, at least in the datasets we analysed (Supplementary figure 3).

We then compared protein abundances between DIA datasets and DDA datasets from[11]. Protein abundances from DDA datasets were computed as iBAQ values and as mentioned in 'Methods' and the abundances from DIA datasets were transformed from LFQ to iBAQ values. We observed strong correlation of protein abundances in various tissues including colon (R^2=0.64), liver (R^2=0.63) and brain (R^2=0.55) and weak correlation in lung (R^2=0.24), pancreas (R^2=0.33), duodenum (R^2=0.34) and thyroid (R^2=0.35) (Figure 5).

In addition, we compared our results with protein expression data from DDA studies available in ProteomicsDB[18]. We observed a good correlation in protein abundance across various tissues (Supplementary Figure 4). The strongest correlations were observed in liver (R^2=0.71) and colon (R^2=0.68) and the lowest in thyroid (R^2=0.39). We also compared abundances with protein expression data across various human tissues at baseline conditions from a large-scale study using TMT[19]. We did not observe a strong correlation across various tissues (Supplementary Figure 5).

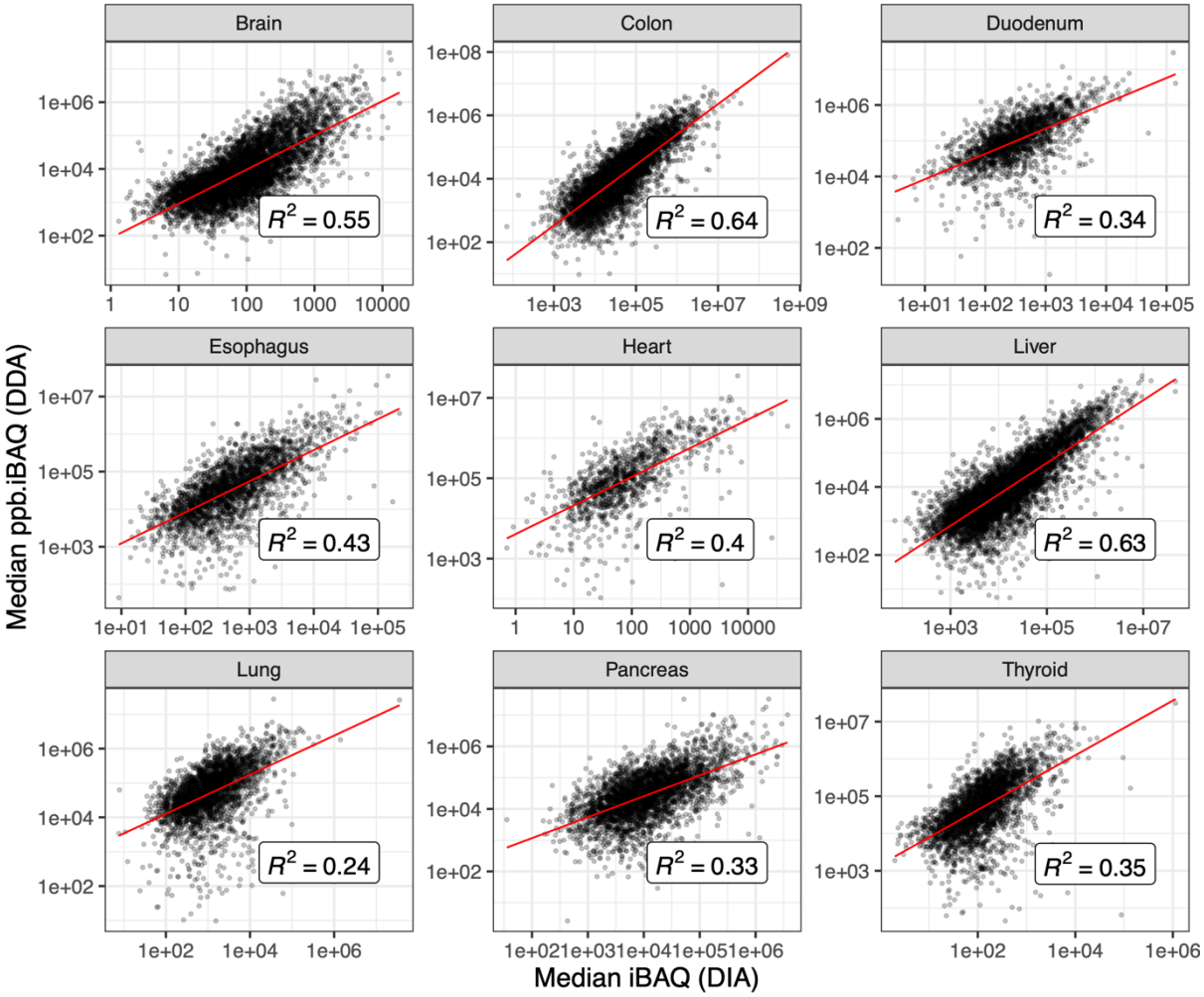


Figure 5. Correlation of protein abundances between human baseline DIA and the DDA datasets from a previous study[11].

Discussion

Here we present a detailed reanalysis of 15 DIA proteomics datasets from 12 human tissues in baseline conditions. We selected datasets from PRIDE, manually curated them connecting samples with the files, and generated protein abundances for all datasets. The objective was two-fold. First, we wanted to facilitate access to protein abundance data from state-of-the-art DIA approaches. Second, we wanted to explore their comparability with analogous protein abundance data generated using DDA approaches.

Different benchmarking studies have been published, pointing out advantages and disadvantages of using different approaches (and software tools) for DIA analysis[43-45]. High-quality spectral libraries can be generated *in silico* in tools such as DIA-NN including every

peptide sequence. However, using a public or a 'matched experimental' library, containing peptidoforms likely to be present in the sample (e.g. for particular biological conditions, tissues or fluids) would likely provide a better statistical power, at the expense of losing a few low abundant peptidoforms absent from a DDA library. An *in silico* spectral library can be potentially much bigger, and thus may give lower sensitivity overall. We here assessed the feasibility of using them to obtain reliable results.

It is important to highlight that the use of *in silico* spectral libraries makes public data reanalysis efforts feasible (like in the case of DDA datasets), without having to rely on the availability of a 'pan-species' public library. The other alternative option is that submitters make the in-house spectral libraries publicly available as well, but at present, this does not happen very often. Indeed, although spectral libraries can be submitted to ProteomeXchange resources (and PRIDE in particular) as part of DIA datasets, it is not mandatory to do it and then, it is not a common practice yet. However, there are cases where they are made available, and this is a key factor for reproducing the results of the original studies.

The approach followed in this study is different from a previous reanalysis effort of DIA datasets, where the 'pan-human' library was used[17], but also a different analysis software (the pipeline was based on OpenSWATH)[46]. No common datasets were included in both studies, since none of the datasets included were common. By using an entrapment database, we checked that the resulting protein FDR per dataset was at an adequate level.

We also compared the protein abundance values in DIA datasets with the results generated from previous DDA studies. One caveat is that at present, the number of public DIA human datasets from baseline human tissues is much smaller when compared with DDA. The results were heterogeneous, with good level of overall correlation for some of the tissues (especially for liver and colon), but much lower for others (e.g. lung and pancreas). Also, we studied the groups of proteins that were detected by either DDA, DIA or both approaches. Finally, we observed a similar percentage of protein abundance missing values among samples analysed by DIA and DDA techniques. It needs to be taken into account that there are some limitations in this comparison. On one hand, human samples are not matched between the different studies. On the other hand, for our previous DDA metanalysis study, we used a different protein sequence database, including several proteins per gene. Last, there is only one DIA dataset where fractionation was used.

Whereas reuse of public DIA proteomics datasets is common for benchmarking efforts in the context of development of software tools and analysis approaches (e.g. [47-49]) data reanalysis of DIA datasets is still very limited, unlike in the case of DDA datasets. However, the trend is changing, as more data make it into the public domain. Some recent efforts such as the development of the open quantMS pipeline[50] can facilitate the reanalysis of very large datasets.

In conclusion, we here present a metaanalysis study of public DIA human datasets generated from tissues in baseline conditions from the PRIDE database. Analogous studies in model organisms (such as those possible for DDA data) are not feasible yet because the amount of DIA baseline data from them is still small. The resulting protein abundance data has been made available via Expression Atlas.

Code availability

https://github.com/Ananth-Prakash/PRIDE-DIA-DataReuse

Data availability

Supplementary Tables:
https://docs.google.com/spreadsheets/d/1__PtA0Ly9FG27RadlVngi_6cw_clKCYV/edit?usp=drive_link&ouid=107719713788096579220&rtpof=true&sd=true

Supplementary Figures:
https://drive.google.com/file/d/1MVYesCbW7gHYv7IgFEVGBpIZUx28mREG/view?usp=drive_link

Supplementary data

Supplementary Table 1: Canonical protein abundances (iBAQ) across various tissues.

Supplementary Table 2: Binned canonical protein abundances across various tissues.

Supplementary Table 3: Canonical proteins identified by DIA only, DDA only and by both techniques.

Supplementary Figure 1: Comparison of DIA-NN runtimes of an example dataset when run in sequential and parallel modes.

Supplementary Figure 2: Venn diagrams showing the number of canonical proteins identified by DIA in this study and by DDA in our previous study[11] in various tissue samples.

Supplementary Figure 3: Comparison of missing values in samples between DDA[11] and DIA datasets.

Supplementary Figure 4: Comparison of protein expression across various human tissues in baseline conditions analysed from DIA datasets (this study) and from ProteomicsDB.

Supplementary Figure 5: Comparison of protein expression across various human tissues at baseline condition analysed from DIA datasets (this study) and from Jiang *et al.*[19] using the TMT-labelling method.

Acknowledgements

First, we would like to thank all data submitters who made their datasets available via PRIDE and ProteomeXchange. This work has been funded by the BBSRC/NSF grant 'DIA-Exchange' [BB/X001911/1], BBSRC 'GRAPPA' [BB/T019670/1 and BB/T019557/1], Wellcome [grant number 221401/Z/20/Z] and EMBL core funding.

Abbreviations

EA: Expression Atlas
DDA: Data Dependent Acquisition
DIA: Data Independent Acquisition
FDR: False Discovery Rate
iBAQ: intensity-Based Absolute Quantification
IDF: Investigation Description Format
LFQ: Label-Free Quantification
ppb: parts per billion
SDRF: Sample and Data Relationship Format

References

1 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, doi:10.1074/mcp.O111.016717 (2012).

2 Jones, A. R., Deutsch, E. W. & Vizcaino, J. A. Is DIA proteomics data FAIR? Current data sharing practices, available bioinformatics infrastructure and recommendations for the future. *Proteomics* **23**, e2200014, doi:10.1002/pmic.202200014 (2023).

3 Xue, Z. *et al.* DPHL v.2: An updated and comprehensive DIA pan-human assay library for quantifying more than 14,000 proteins. *Patterns (N Y)* **4**, 100792, doi:10.1016/j.patter.2023.100792 (2023).

4 Perez-Riverol, Y. *et al.* The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* **50**, D543-D552, doi:10.1093/nar/gkab1038 (2022).

5 Deutsch, E. W. *et al.* The ProteomeXchange consortium at 10 years: 2023 update. *Nucleic Acids Res* **51**, D1539-D1548, doi:10.1093/nar/gkac1040 (2023).

6 Mann, M., Kumar, C., Zeng, W. F. & Strauss, M. T. Artificial intelligence for proteomics and biomarker discovery. *Cell Syst* **12**, 759-770, doi:10.1016/j.cels.2021.06.006 (2021).

7 Ochoa, D. *et al.* The functional landscape of the human phosphoproteome. *Nat Biotechnol* **38**, 365-373, doi:10.1038/s41587-019-0344-3 (2020).

8 Ramsbottom, K. A. *et al.* Meta-Analysis of Rice Phosphoproteomics Data to Understand Variation in Cell Signaling Across the Rice Pan-Genome. *J Proteome Res*, doi:10.1021/acs.jproteome.4c00187 (2024).

9 Martens, L. & Vizcaino, J. A. A Golden Age for Working with Public Proteomics Data. *Trends Biochem Sci* **42**, 333-341, doi:10.1016/j.tibs.2017.01.001 (2017).

10 George, N. *et al.* Expression Atlas update: insights from sequencing data at both bulk and single cell level. *Nucleic Acids Res* **52**, D107-D114, doi:10.1093/nar/gkad1021 (2024).

11 Prakash, A. *et al.* Integrated View of Baseline Protein Expression in Human Tissues. *J Proteome Res* **22**, 729-742, doi:10.1021/acs.jproteome.2c00406 (2023).

12 Wang, S. *et al.* Integrated view and comparative analysis of baseline protein expression in mouse and rat tissues. *PLoS Comput Biol* **18**, e1010174, doi:10.1371/journal.pcbi.1010174 (2022).

13 Wang, S. *et al.* Integrated Proteomics Analysis of Baseline Protein Expression in Pig Tissues. *J Proteome Res*, doi:10.1021/acs.jproteome.3c00741 (2024).

14 Jarnuczak, A. F. *et al.* An integrated landscape of protein expression in human cancer. *Sci Data* **8**, 115, doi:10.1038/s41597-021-00890-2 (2021).

15 Robles, J., Prakash, A., Vizcaino, J. A. & Casal, J. I. Integrated meta-analysis of colorectal cancer public proteomic datasets for biomarker discovery and validation. *PLoS Comput Biol* **20**, e1011828, doi:10.1371/journal.pcbi.1011828 (2024).

16 Walzer, M. *et al.* Implementing the reuse of public DIA proteomics datasets: from the PRIDE database to Expression Atlas. *Sci Data* **9**, 335, doi:10.1038/s41597-022-01380-9 (2022).

17 Rosenberger, G. *et al.* A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci Data* **1**, 140031, doi:10.1038/sdata.2014.31 (2014).

18 Samaras, P. *et al.* ProteomicsDB: a multi-omics and multi-organism resource for life science research. *Nucleic Acids Res* **48**, D1153-D1163, doi:10.1093/nar/gkz974 (2020).
19 Jiang, L. *et al.* A Quantitative Proteome Map of the Human Body. *Cell* **183**, 269-283 e219, doi:10.1016/j.cell.2020.08.036 (2020).
20 Athar, A. *et al.* ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res* **47**, D711-D715, doi:10.1093/nar/gky964 (2019).
21 Hulstaert, N. *et al.* ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion. *J Proteome Res* **19**, 537-542, doi:10.1021/acs.jproteome.9b00328 (2020).
22 Adusumilli, R. & Mallick, P. Data Conversion with ProteoWizard msConvert. *Methods Mol Biol* **1550**, 339-368, doi:10.1007/978-1-4939-6747-6_23 (2017).
23 Mellacheruvu, D. *et al.* The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat Methods* **10**, 730-736, doi:10.1038/nmeth.2557 (2013).
24 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, doi:10.1038/s41592-019-0638-x (2020).
25 Feng, X. D. *et al.* Using the entrapment sequence method as a standard to evaluate key steps of proteomics data analysis process. *BMC Genomics* **18**, 143, doi:10.1186/s12864-017-3491-2 (2017).
26 Yu, F. *et al.* Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. *Nat Commun* **14**, 4154, doi:10.1038/s41467-023-39869-5 (2023).
27 Yoo, A. B., Jette, M. A. & Grondona, M. SLURM: Simple Linux Utility for Resource Management. Springer Berlin Heidelberg, 44-60. doi:10.1007/10968987_3 (2003).
28 Moreno, P. *et al.* Expression Atlas update: gene and protein expression in multiple species. *Nucleic Acids Res* **50**, D129-D140, doi:10.1093/nar/gkab1030 (2022).
29 Miedema, S. S. M. *et al.* Distinct cell type-specific protein signatures in GRN and MAPT genetic subtypes of frontotemporal dementia. *Acta Neuropathol Commun* **10**, 100, doi:10.1186/s40478-022-01387-8 (2022).
30 Komor, M. A. *et al.* Molecular characterization of colorectal adenomas reveals POFUT1 as a candidate driver of tumor progression. *Int J Cancer* **146**, 1979-1992, doi:10.1002/ijc.32627 (2020).
31 Bourassa, S. *et al.* Evaluation of iTRAQ and SWATH-MS for the Quantification of Proteins Associated with Insulin Resistance in Human Duodenal Biopsy Samples. *PLoS One* **10**, e0125934, doi:10.1371/journal.pone.0125934 (2015).
32 Hou, G. *et al.* Biomarker Discovery and Verification of Esophageal Squamous Cell Carcinoma Using Integration of SWATH/MRM. *J Proteome Res* **14**, 3793-3803, doi:10.1021/acs.jproteome.5b00438 (2015).
33 Brandenburg, S. *et al.* Direct proteomic and high-resolution microscopy biopsy analysis identifies distinct ventricular fates in severe aortic stenosis. *J Mol Cell Cardiol* **173**, 1-15, doi:10.1016/j.yjmcc.2022.08.363 (2022).
34 Ng, C. K. Y. *et al.* Integrative proteogenomic characterization of hepatocellular carcinoma across etiologies and stages. *Nat Commun* **13**, 2436, doi:10.1038/s41467-022-29960-8 (2022).
35 Stewart, P. A. *et al.* Relative protein quantification and accessible biology in lung tumor proteomes from four LC-MS/MS discovery platforms. *Proteomics* **17**, doi:10.1002/pmic.201600300 (2017).

36 Trilla-Fuertes, L. *et al.* Identification of Carcinogenesis and Tumor Progression Processes in Pancreatic Ductal Adenocarcinoma Using High-Throughput Proteomics. *Cancers (Basel)* **14**, doi:10.3390/cancers14102414 (2022).
37 Koch, M. *et al.* Quantitative Proteomics Identifies Reduced NRF2 Activity and Mitochondrial Dysfunction in Atopic Dermatitis. *J Invest Dermatol* **143**, 220-231 e227, doi:10.1016/j.jid.2022.08.048 (2023).
38 Martinez-Aguilar, J., Clifton-Bligh, R. & Molloy, M. P. Proteomics of thyroid tumours provides new insights into their molecular composition and changes associated with malignancy. *Sci Rep* **6**, 23660, doi:10.1038/srep23660 (2016).
39 Mol, M. O. *et al.* Proteomics of the dentate gyrus reveals semantic dementia specific molecular pathology. *Acta Neuropathol Commun* **10**, 190, doi:10.1186/s40478-022-01499-1 (2022).
40 De Marchi, T. *et al.* Proteogenomic Workflow Reveals Molecular Phenotypes Related to Breast Cancer Mammographic Appearance. *J Proteome Res* **20**, 2983-3001, doi:10.1021/acs.jproteome.1c00243 (2021).
41 Solberg, J. B. K. *et al.* The Proteome of Hand Eczema Assessed by Tape Stripping. *J Invest Dermatol* **143**, 1559-1568 e1555, doi:10.1016/j.jid.2022.12.024 (2023).
42 Doering, T. M. *et al.* The muscle proteome reflects changes in mitochondrial function, cellular stress and proteolysis after 14 days of unilateral lower limb immobilization in active young men. *PLoS One* **17**, e0273925, doi:10.1371/journal.pone.0273925 (2022).
43 Gotti, C. *et al.* Extensive and Accurate Benchmarking of DIA Acquisition Methods and Software Tools Using a Complex Proteomic Standard. *J Proteome Res* **20**, 4801-4814, doi:10.1021/acs.jproteome.1c00490 (2021).
44 Frohlich, K. *et al.* Benchmarking of analysis strategies for data-independent acquisition proteomics using a large-scale dataset comprising inter-patient heterogeneity. *Nat Commun* **13**, 2622, doi:10.1038/s41467-022-30094-0 (2022).
45 Staes, A. *et al.* Benefit of In Silico Predicted Spectral Libraries in Data-Independent Acquisition Data Analysis Workflows. *J Proteome Res* **23**, 2078-2089, doi:10.1021/acs.jproteome.4c00048 (2024).
46 Rost, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol* **32**, 219-223, doi:10.1038/nbt.2841 (2014).
47 Martinez-Val, A., Bekker-Jensen, D. B., Hogrebe, A. & Olsen, J. V. Data Processing and Analysis for DIA-Based Phosphoproteomics Using Spectronaut. *Methods Mol Biol* **2361**, 95-107, doi:10.1007/978-1-0716-1641-3_6 (2021).
48 Gupta, S., Sing, J. C. & Rost, H. L. Achieving quantitative reproducibility in label-free multisite DIA experiments through multirun alignment. *Commun Biol* **6**, 1101, doi:10.1038/s42003-023-05437-2 (2023).
49 Allen, C. *et al.* nf-encyclopedia: A Cloud-Ready Pipeline for Chromatogram Library Data-Independent Acquisition Proteomics Workflows. *J Proteome Res* **22**, 2743-2749, doi:10.1021/acs.jproteome.2c00613 (2023).
50 Dai, C. *et al.* quantms: a cloud-based pipeline for quantitative proteomics enables the reanalysis of public proteomics data. *Nat Methods*, doi:10.1038/s41592-024-02343-1 (2024).