Implementing the re-use of public DIA proteomics datasets: from the PRIDE database to Expression Atlas

Mathias Walzer[1], David García-Seisdedos[1], Ananth Prakash[1], Paul Brack[2], Peter Crowther[3], Robert L. Graham[4], Nancy George[1], Suhaib Mohammed[1], Pablo Moreno[1], Irene Papathedourou[1], Simon J. Hubbard[2], and Juan Antonio Vizcaíno[1]

[1]European Molecular Biology Laboratory, EMBL-European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, CB10 1SD, United Kingdom.
[2]Division of Evolution and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Oxford Road, Manchester M13 9PT, United Kingdom.
[3]Melandra Limited, 16 Brook Road, Urmston, Manchester M41 5RY, United Kingdom.
[4]School of Biological Sciences, Chlorine Gardens, Queen's University Belfast, Belfast, BT9 5DL, United Kingdom.
[*]corresponding author(s): Mathias Walzer (walzer@ebi.ac.uk), Dr. Juan Antonio Vizcaíno (juan@ebi.ac.uk)

ABSTRACT

Rising numbers of mass spectrometry proteomics datasets available in the public domain, increasingly include volumes generated from Data Independent Acquisition approaches, SWATH-MS in particular. Unlike Data Dependent Acquisition datasets, their re-use is limited, partially due to challenges in combination and use of free software for analysis in the non-specialist laboratory.

We introduce a (re-)analysis pipeline for SWATH-MS data available in the PRIDE database, which includes a harmonised combination of metadata annotation protocols, automated workflows for MS data, statistical analysis and results integration into the resource Expression Atlas. Automation is orchestrated with Nextflow, using containerised open analysis software tools, rendering the pipeline readily available, reproducible and easy to update. To demonstrate its utility, we reanalysed 10 public DIA datasets, 1,278 individual SWATH-MS runs, stored in PRIDE.

The robustness of the analysis was evaluated and compared to the results obtained in the original publications. The final results were exported into Expression Atlas, making quantitative results from SWATH-MS experiments more widely available and integrated with results from other reanalysed proteomics and transcriptomics datasets.

Introduction

In recent years, the public availability of mass spectrometry (MS)-based proteomics datasets has continued to increase dramatically, becoming a common and used scientific resource for the field. This has mirrored the growth and wide reuse that has happened previously in other related *omics* fields such as transcriptomics. The PRIDE (PRoteomics IDEntifications) database[1] at the European Bioinformatics Institute (EBI, https://www.ebi.ac.uk/pride/) is the largest proteomics data repository worldwide and is one of the founding members of the ProteomeXchange consortium[2]. During 2020 alone, PRIDE stored around 5,300 datasets, coming from a wide variety of species and different experimental approaches.

One of the main benefits of public data availability is that it enables experimental reproducibility and reuse, facilitating an independent assessment of the results described in the corresponding publications, representing an auditable route to (re)trace the source of key findings and supporting the discovery of new findings, as more advanced software becomes available. This huge increase in the amount of proteomics data in the public domain has indeed triggered a lot of possible data reuse activities[3,4] and new applications in the field, including among others, numerous meta-analysis studies, proteogenomics applications or the use of artificial intelligence approaches such as machine-learning[5,6]. Expression Atlas (https://www.ebi.ac.uk/gxa/home) is an added-value resource at the EBI that enables easy access to integrated information about gene and protein expression across species, tissues, cells, experimental conditions and diseases[7]. The EA 'bulk' Atlas has two sections: baseline and differential. In the context of public proteomics data re-use activities, protein expression results derived from Data Dependent Acquisition (DDA) reanalysis of human, mouse and cell-line public datasets have been already integrated into Expression Atlas[7,8]. The availability of such results in EA makes proteomics expression data integrated with transcriptomics information.

To date, only DDA-based proteomics datasets have been reanalysed and integrated in EA. However, an increasing fraction

of data deposited in public repositories is now coming from DIA (Data Independent Acquisition) approaches and from SWATH-MS (Sequential Window Acquisition of All Theoretical Mass Spectra) methods in particular. DIA methods differ from DDA techniques in that no narrow-window ion selection takes place in the instruments. As a result, fragment ion spectra of all technically available peptide ion species are produced. In the case of SWATH-MS, this is conducted in cycles of sequential windows of m/z values[9,10], creating permanent digital maps of the samples' peptides/proteins, which can potentially be re-interrogated over time[10–13], for instance with novel libraries including new peptides. This holds great promise for biomarker discovery and other applications, especially since DIA approaches can have distinct advantages for quantitative studies. For instance, high reproducibility between technical and biological replicates[14], and high intra-laboratory reproducibility, as well as between laboratories have been shown[15]. As a result, SWATH-MS methods can capture a comprehensive picture of the sample measured and have established themselves as a reproducible method for large-scale protein quantification.

In SWATH-MS, the m/z space per cycle is segmented into a set of acquisition windows, within which the analysis software attempts to detect signals from a library of expected peptides provided in a spectral library. Unfortunately, such an analysis is complex and not standardised, contributing to a very large number of parameter settings typically required to specify a given SWATH-MS data analysis pipeline. Depending on the instrument speed, target acquisition m/z range and gradient length, either 32 or 64 equally sized windows are often used but this can vary from study to study, and dynamically sized windows are also often used to uniformly cover a given proteome. Similarly, the methods used to detect, score and statistically validate the underlying peptide associated mass spectral features (i.e., co-eluting ion group signals from different transitions) vary between studies, and depend on the data analysis software used.

As mentioned above, most SWATH-MS data analyses use a targeted approach where peptides from a spectral library are directly queried from expected fragment ions (m/z, c) at a certain retention time (RT) range. These are usually empirically determined on the same or a similar instrumentation and settings, generating DDA data from pooled or representative samples. At present, despite some notable exceptions ([12,16–18]), sample/study specific target spectral libraries are rarely deposited alongside the original SWATH-MS data in public repositories. Additionally, different (custom) formats make them difficult to adapt for reanalysis. In 2014, a pan-human target library comprising 10,000 proteins was published[12], covering 50.9% of the annotated human proteins in UniProtKB/Swiss-Prot. These targets can be used more generally as their normalised RTs are reported and additional SWATH-MS studies can be normalised to the same scale, usually by the inclusion of known iRT peptides[19]. However, the use of a generic target library requires an appropriate statistical control of peptide and protein error rates[20].

This robustness of the overall SWATH-MS approach has been demonstrated several times, notably in a benchmarking comparison of DIA software tools which observed very similar qualitative and quantitative results[21]. Although such studies show that DIA methods are a powerful approach for reproducible quantification, as highlighted above, their re-use is at present very limited. This stems from the practicalities of deploying a data analysis pipeline, due to either the large amounts of data to be processed, the huge computational efforts needed, the intricate software settings necessary, or the uncertainty about whether the data from a particular study is adequate for reanalysis in a different research setting. In this study, we provide a robust open reanalysis pipeline for DIA data and demonstrate its utility to confidently reanalyse ten public human DIA datasets obtained from PRIDE. Furthermore, the resulting quantified proteins are subject to further statistical analysis and the resulting protein expression data is integrated into Expression Atlas for widely available access. To the best of our knowledge, this is the first time that public DIA data has been systematically reanalysed and integrated with gene expression information in a widely-used resource. We demonstrate, that with well-formed experimental design annotation and (known) iRT peptide spiked data, it is possible to reanalyse public DIA data in a generic context.

Results

Open analysis pipeline for DIA data

The developed reanalysis pipeline is available at https://gitlab.ebi.ac.uk/pride-reanalysis/dia-reanalysis. It represents a harmonised combination of metadata annotation protocols, open proteomics software tools for SWATH-MS data in computational workflows orchestrated by Nextflow[22] which can be run in multiple compute environments, and data integration procedures. Figure 1 illustrates the distinct parts of the overall pipeline process. The pipeline integrates all the necessary steps in SWATH-MS data analysis ranging from data curation, raw data conversion, through to statistical analysis and data integration procedures for submission into EA. Software for the different steps of each part of the pipeline are fully containerised (Table 3), making a version update to individual tools more straightforward. A new container including an updated version of a given tool can be provided through the same parameter file used to change parameter settings e.g., a FDR (False Discovery Rate) threshold for peak group detection.

To add flexibility, the pipeline was split into four parts. The first (Figure 1A), includes dataset curation and acquisition, and the second (Figure 1B) contains the upstream analysis of the SWATH-MS raw data, automated with a Nextflow workflow. The third (Figure 1C), provides downstream statistical analysis, and is also automated with Nextflow, and a fourth part (Figure 1D)

contains an internal workflow for the integration of results into Expression Atlas. The split has the advantage that the analysts can inspect intermediate results before initiating the following part of the pipeline. For example, the second part contains an extra step to produce Quality Control/Assurance (QC/QA) records for each MS run, which could be inspected prior to initiating subsequent steps of the pipeline. This was particularly helpful in assessing potentially incorrect iRT peptide detection in a newly reanalysed dataset. An example for such assessment with the resulting exclusion from further analysis can be found in the Supplementary Material (Supplementary Figure S1). Optional entry points for the SWATH-MS raw data analysis workflow were added for datasets already analysed with OpenSWATH and for pre-processed spectral libraries (Supplementary Figure S2). Further details on the pipeline methods and availability can be found in the 'Methods' and 'Code availability' sections, respectively.

Reanalysis of public DIA datasets

Ten PRIDE public datasets, amounting to 1,278 individual SWATH-MS runs (Table 1), were reanalysed as explained in the 'Methods' section, using the EBI High Performance Computing (HPC) infrastructure. The CAL (Combined Assay Library) spectral library (SWATHAtlas accession SAL00031) was used for all except the "Plasma" dataset (for which a plasma-targeted library was preferred, see 'Methods' section).

The accumulated processing time for all datasets was around 19,000 CPU hours, which amounted to approximately 2,200 CPU hours per dataset, or 15 CPU hours per SWATH-MS run, on average.

We first compared the raw numbers of inferred proteins across three distinct levels of protein FDR: 5%, 1% and 0.1%, filtering on the q-values from the pyProphet FDR estimates for global protein level. Overall, the numbers of detected proteins were generally in the expected range for an analysis performed with mammalian cells/tissues, using the (generic) CAL target library and for protein FDR levels of 1% (Figure 2A). At a 1% cut-off for example, just under 3,000 quantified proteins were reported for PXD003497 (with 2,754) and for PXD004691 (with 2,872). Both datasets are prostate cancer sample datasets. On the higher side, PXD004589 (another prostate cancer sample dataset) yielded 3,703 detected proteins, PXD004873 (hepatocellular cancer) 3,530 proteins, and PXD010912 (human liver S9 fractions) 4,224 detected proteins. There were two datasets where fewer proteins were detected: 2,239 in PXD014194, a breast cancer dataset containing only tumour tissue measurements, and as expected, a much lower number were found in the plasma dataset PXD001064 (207 proteins). Additionally, two datasets showed many more detected proteins than the others: 5,946 in PXD014943, a lymphoma sample dataset, and 7,097 in PXD003539, the NCI-60 cell-line dataset containing samples that contributed to the design of the CAL. The FDR threshold of 1% on a global level appeared to be a good trade-off. Results for stricter and more generous cut-offs are shown in Figure 2A for comparison. We note that use of a stricter filter at 0.1% FDR proved impractical, as for the plasma dataset PXD001064 the analysis failed to complete successfully.

To check the broad reliability of the open reanalysis pipeline combined with a generic spectral library, we compared our results with the originally published ones. We further implemented a consistency filter to remove unreliable protein identifications, in common with best practice in many of the published studies. Proteins or protein groups with more than 50% of target features missing within the runs across a study group were removed. We note here that an informed and consistent comparison between the original (as published) and reanalysed protein numbers across all studies is essentially impossible, since among other reasons, the original reported studies have used slightly different consistency filter approaches or FDR thresholds. For 6 of the 10 studies considered, after reanalysis, the number of proteins was higher than those originally reported. In 4 of them, the numbers were markedly higher (Figure 2B). This is likely due to the protein numbers for PXD000672 and PXD004873 being originally reported at a 0.1% protein level FDR, at a 0.1% peptide level FDR for PXD004589. Additionally, PXD014149 was originally analysed with a target library of 3,284 proteins. In the case of PXD010912 we assume fewer proteins were present in the original spectral library (not reported) as well, as the authors reported that only 580 proteins were quantified via DDA. The Supplementary Material to the original publication of PXD003539 offers 6,556 protein groups with missing values quantified, a much closer value to the results of the reanalysis.

After re-analysis, the protein numbers of 4 studies were lower than those originally published, around 300 proteins less with the exception of dataset PXD003497, where a difference of just under 1,000 proteins was found. Interestingly, the study was originally analysed with CAL, albeit customised with additional DDA measurements, obtaining 6,800 proteins. However, the authors subsequently reduced their set of proteins to 3,700 applying a more stringent consistency filter (at least two concordant proteotypic peptides[23]), which is more in line with our reanalysis.

All protein expression results are available via Expression Atlas (for URLs see Table 1). Source numbers are available in Supplementary Table S3, intermediate results are available through the pipeline repository (see Supplementary Table S4).

Baseline expression analysis - Technical validation

To validate the internal consistency of the reanalysed results, we calculated the coefficient of variation (CV) from the results within the respective study groups' technical replicates and in total, comparing them to the original CVs, where available. Three of the original datasets included technical replicates and in the corresponding publication the authors reported protein

intensities at a MS-run level, which are necessary to calculate the CV (Table 2). One publication reported only sample averages, and one included only pooled-sample replicates, preventing a more comprehensive CV comparison. The overall median CV was generally below 21% and the CV values obtained from the reanalyses were similar to the originally published CVs.

The group-wise comparison showed a similar picture to the overall median CV comparison (Figure 3), although remaining in a very similar range, and were in one instance slightly smaller (Figure 3E). Again, this demonstrates the consistency of the reanalysis pipeline to reproduce the global characteristics of measurement variance observed in the original studies. To further compare the results of the reanalyses with the originally published results, we analysed the MS run-wise correlation of previously published protein abundances versus the equivalent results of the reanalyses for the technical replicates (Figure 4). The reanalysis results showed a high correlation with the original studies, though markedly closer to the regression line at higher abundance and a wider distribution towards the lower intensities – as it might be expected. The Pearson product-moment correlation coefficients (R) ranged between 0.9 and 0.98 with a p-value ≤ 0.001, demonstrating a high concordance between the original and the reanalysed protein quantitative data overall, with matched technical reproducibility.

However, when comparing the correlation of the reanalysis results with the original results on a per-MS run pairing (reanalysis versus original results) basis, correlations between matched protein expression values were less pronounced (R values between 0.52 and 0.84, see Supplementary Figure S3). This is consistent with the challenges in matching the exact protein identification parameters and highlights the inherent difficulties in implementing DIA reanalyses in general.

Differential Expression Analysis

Additionally, we also performed a downstream differential expression analysis for the three datasets that reported such values in the original publications: datasets PXD000672, PXD004691 and PXD014943. The programming language R[24] and MSStats[25] were used for the differential statistical analysis (see details in the 'Methods' section). For consistency, we used default MSstats settings (except for the use of median normalisation and the 'top3' protein inference method). Regardless, substantial differences were found when compared with the original studies, though the differentially expressed proteins were in the same general order.

We calculated the correlation of originally reported fold-changes and with our reanalysis fold-changes (see Supplementary Figure S4). There was a substantial overlap of quantified protein expression values between the originally reported results and the reanalysed ones, ranging from 2,992 proteins (out of 4,572 protein fold-changes) in the reanalysis of the PXD014943 eDLBCL-PCNSL contrast, to 863 proteins (out of 1,868 protein fold-changes) in the reanalysis of the PXD000672 benign-ccRCC contrast. The higher ratio of overlap (of the calculated fold changes to overlapping) in PXD014943 also shows the highest correlation R value at 0.84, whereas the lowest also shows the smallest correlation R value of 0.52. This illustrates that the protein signals detected were generally in the same order, yet showing a varying amount of overlap and correlation. This however also hinted at an increased number of lower strength protein signals, that differed between analyses to a greater extent, being picked up as quantified proteins increase, and in turn reducing the overall concordance. Accordingly, when drawing a cut-off for adjusted p-values (< 0.05) and fold-changes (> 2), the overlap between the original and the reanalysed results was reduced further. For the PXD014943 eDLBCL-PCNSL contrast, we found 118 proteins to be significantly differentially expressed overlapping in 21 proteins with the original study's 97 proteins significantly expressed (see Supplementary Table S1). The overlap of proteins of significance between original results and the reanalysed results was more pronounced in the PXD000672 benign-ccRCC contrast, with 59 of the 262 proteins from the reanalysis matching with the originally reported 613 proteins (see Supplementary Table S1). We also encountered an instance of no overlap in PXD004691 normal(ff)-PrC(ff) contrast from the 3 proteins of significance originally reported and the 8 proteins detected from the reanalysis. The corresponding protein expression values were exported into Expression Atlas as explained in the 'Methods' section. See details about the corresponding datasets in Table 1. Of note, the analysis with the MSstats default protein inference setting of 'all' (instead of 'top3') resulted in a far fewer number significantly differentially expressed proteins. Here, the intensity correlations between originally reported protein abundance values and the reanalyses (see Supplementary Figure S4) revealed an unexpected lack of correspondence.

Discussion

The popularity of DIA approaches in the proteomics field is increasing. However, data reuse of this type of public datasets has been so far very scarce, apart from cases where the main aim was the benchmarking of software (for example [26,27]). This is despite one of the principal advantages of SWATH-MS, namely that it represents a permanent and comprehensive digital signature of a proteome that can be analysed again, with new software or new libraries. Since it is clearly desirable for this to become a reality, systematic reanalyses as presented here need to be conducted to establish good practice and to provide a common reference framework. However, the complexity of a SWATH-MS study data analysis can be overwhelming. A common, ready-made pipeline lowers the entry threshold for the analysis of unseen datasets – both factors which motivated this study. Hence, in this paper, we report, to the best of our knowledge, the first systematic effort to enable the reanalysis and

integration of the results of this type of studies. The implemented open reanalysis pipeline produces overall robust results and provides sufficient flexibility for users to adjust to different reanalysis scenarios.

Due to the differing experimental designs, the downstream data analysis was hard to formalise in a generic analysis workflow. However, a good compromise could be reached by designing the statistical analysis part of the pipeline using MSstats' dataProcess and further data adjustments such as normalisation and filtering done using R. From a technical perspective, the additionally implemented entry points to the first Nextflow workflow (Figure 1B) for pre-processed data (e.g., OpenSWATH output files) proved to be advantageous, since analyses that follow the initial SWATH-MS data analysis step often needed to be tested with different parameter settings. The Nextflow configuration of iteratively repeating failed jobs with an increased memory request was also found to be advantageous. The most influential variables on memory requirements for the analysis depended on vastly differing settings between the individual studies, including liquid Chromatography (LC) gradient length, sample content, and of course the study size. HPC compute environments appear as the logical choice for data (re-)analysis efforts without clear resource requirement boundaries. However, research institutions' collaborations with private cloud systems are becoming increasingly popular (e.g. OpenStack). SWATH-MS data analysis in the cloud, as showcased by Peters et al.[28], can result in greatly increased compute costs, when applied to a broad basis of inputs but without the close knowledge of the dataset as the data producing lab would have - like with the datasets chosen in this study. DIA analyses are also inherently disk-intensive processes due to the size of the MS runs and the spread of analysis features throughout the data (window sequences). This, too, is an often-overlooked contributor to compute cost, especially in cloud compute environments, where storage and volume of uploaded/downloaded data are usually billed separately. The interrogation of *.wiff/.scan* files is, as with any other proprietary raw file formats, bound to many constraints. Inspection either needs proprietary software (e.g. Microsoft Windows), conversion into an open format which takes considerable time and disk space or is otherwise not compatible with an automated high throughput.

During the initial selection of public datasets from PRIDE for reanalysis, we realised that a great proportion did not have (completely) paired *.wiff/.scan* files, rendering the unpaired samples unusable. At that point, complete pairing was made mandatory when submitting new SWATH-MS datasets to PRIDE. We hope this will help reanalysis efforts in the future, since it is essential information reflecting the experimental design as used in the original publication. Mapping raw file names in PRIDE to the samples in the original publication was done manually and it constituted one of the most time-consuming steps in this work. In the context of the activities of the Proteomics Standards Initiative, we are trying to address this issue by developing a standard file format called SDRF-Proteomics (Sample and Data Relationship Format-Proteomics) file (as part of the file format MAGE-TAB-Proteomics) for capturing the experimental design in proteomics experiments[1], and we have started working in the related tooling to facilitate the creation of these files. It is important to highlight that submission of SDRF-Proteomics files is already supported by PRIDE, although it is optional at the time of writing. The authors encourage data submitters to align and improve efforts to match metadata referenced in a publication with the underlying data and make the experimental design an explicit part of any publication. In any case, concrete and consistent name-and-group mapping tables are already of great help to any reanalysis or replication effort and increase the transparency of any original publication and therefore its value to the community. The annotation of the used iRT peptides, too, is crucial to DIA data (re-)analysis efforts. Indeed, detection of the iRT peptides is essential to their success. We consider this part of quality assurance an important step during measurement acquisition or the reanalysis process.

We deployed the CAL target library in combination with a 1% FDR level threshold for global peptide and protein quantitation as an appropriate common setting for our reanalyses. This sometimes generated relatively lower numbers of proteins quantified (when compared to the results from custom-made target libraries used in the original publications). This could be attributed to the limited number of target (peptide fragment ion) features available for some proteins in CAL. For example, the filtered CAL used for the plasma dataset, reduced the number of detected proteins by over half compared to the original study. Hence, we consider the reduction of the target space for more specialised samples such as biofluids like plasma to warrant further customisation to the analysis procedure, as opposed to the 'generic' application of CAL to the rest of the datasets. However, the trade-off merits between comprehensive libraries and specialised ones are beyond the scope of this initial study, and we consider the "core" of UniProtKB/Swiss-Prot proteins present in CAL is appropriate for a large proportion of human tissue and cell line studies. Other possibilities for future reanalysis efforts could be the use of "library-free" approaches[29–31] and/or in silico predicted libraries[32,33]. These approaches are starting to get more popular in the field, however when this study was designed, we chose the more established open data analysis approach, which remains the most frequently used approach to date.

As noted, the reanalysis pipeline produced different number of proteins than the original published studies. There is a vast choice of parameter settings to be made and many technical details are interconnected throughout steps of the process. This generates a dependency on subsequent steps, resulting in a high sensitivity to minor fluctuations in the parameters which can be propagated down the pipeline to the final stages of protein quantitation and differential expression, as shown. It should also be noted that not all aspects of an original study analysis setup could be reflected because of either the limited information available or the infeasibility to integrate it into a single generic automated workflow. Hence, differences in methodology, parameters, and

representation of experimental design can be attributed to the result deviations. This study's aim was the reanalysis of multiple studies using a common, open and reproducible analysis pipeline.

Regardless, there was a good level of agreement between replicates within each pipeline, as measured by CV values, demonstrating that a robust quantification methodology was followed. However, even with high protein detection numbers and a good agreement between technical replicates and calculated CV values close to the original study, the outcome of protein inference and subsequent differential analyses can vary substantially, which underlines the importance of, and the sensitivity to, parameter selection choices when handling DIA data. Particularly, we noted the impact that the method of protein inference chosen in MSstats had on the consistency filtered protein lists and the final differential quantification results (Supplementary Figure S4, Supplementary Tables S3, S2).

Nonetheless, there is also basic agreement between the analyses on fold-change level, with a sufficient overlap and preservation of general trends. However, with a further restriction to discrete sets of proteins classed as significant, overlaps were reduced and differences widened. SWATH-MS study data analysis remains a complex set of tasks and increasing differences from subtle changes must be expected with each added layer of complexity, a classic paradigm in differential *omics* analyses.

Our reanalysis results haves been made available via Expression Atlas, thereby offering a convenient route to integrate DDA and DIA proteomics expression data with transcriptomics data in the same resource. We hope this will help popularise proteomics approaches in general and DIA approaches in particular, especially for non-proteomics researchers. Finally, the integration of qualitative and quantitative data from complementary *omics* techniques is, in our view, vital for having a broadened understanding of the underlying biological processes in different organisms.

Conclusions

To summarise, we have built an open, versatile and modular reanalysis pipeline for DIA proteomics data, that presents a harmonised combination of open SWATH-MS and statistical analysis tools. We selected and manually annotated ten publicly available datasets from PRIDE, and successfully reanalysed them without using custom-built target libraries. We showed the feasibility of a general reanalysis approach for SWATH-MS datasets and highlighted the shortcomings when compared to a customised spectral library target approach. We made the final quantitative reanalysis results available via Expression Atlas to further data integration from different *omics* sources. The analysis pipeline represents a resource in its own right, providing researchers a streamlined basis of established tools. Importantly, this provides a unified consistent reanalysis pipeline to support comparative analyses of datasets from different sources.

Methods

Selection of public DIA proteomics datasets and manual curation

After performing an initial selection of publicly available SWATH-MS datasets from PRIDE1 (all measured from human samples with SCIEX instruments), 10 datasets were selected with a preference for studies with technical replicates. Further criteria for reanalysis inclusion were the completeness of data in terms of the necessary files and data types, including both *.wiff* and *.scan* files, existing annotation of the used iRT peptides, and further MS run information, to be able to reconstruct the study analysis design. In all cases, each dataset was manually curated to extract the main analysis characteristics, the processing parameters, the experimental design and sample characteristics. The biological metadata for each dataset was captured in a SDRF file. Then, the raw data files from the selected 10 datasets (Table 1) were downloaded and used as input for the reanalysis (Figure 1A). The main characteristics of the selected 10 datasets are summarised below.

HCCpct (PXD00487322)

This dataset consisted of 76 PCT-SWATH runs from 19 HCC patients' hepatectomies. Overall, 38 tissue samples (pairs of benign and tumorous tissues) were prepared, spiked with iRT peptides and measured with a 5600 TripleTOF instrument over a 45-min LC gradient, in technical replicates. The CAL library was used for the original analysis, resulting in 2,579 quantified proteins.

DigitalKidney (PXD00067213)

This kidney dataset consisted of 48 SWATH-MS runs: 18 kidney tissue samples were processed as benign and tumorous pairs from 9 patients with renal cell carcinoma (RCC), 6 of which classified as clear-cell RCC (ccRCC), two were classified as papillary RCC (pRCC) and one as chromophobe RCC. Additionally, 4 digests of human kidney tissue were processed in triplicate (twelve aliquots). The processing included spiking with iRT peptides. Data acquisition was performed on a 5600 TripleTOF mass spectrometer over a 120-min LC gradient. The custom target library used in the original analysis contained targets from 41,542 proteotypic peptides, from 4,624 proteins, compiled using DDA analyses of kidney tissues. Using an FDR of 0.1% at a precursor level, 1,632 unique proteins were originally quantified.

BankPrCa (accession PXD00469123)

This dataset contained 224 SWATH-MS runs coming from prostate cancer (PrCa) biopsies. The tissues samples were either fresh frozen (FF) or formalin-fixed-paraffin-embedded before pressure cycling technology (PCT) was used. All samples were spiked with iRT peptides (Biognosys19) before injection. Data acquisition was performed on a 5600 TripleTOF mass spectrometer with a 30-min LC gradient, with technical replicates. DIA analysis was performed with a custom library (70,981 peptide precursors from 6,686 UniProtKB/Swiss-Prot proteins) compiled from unfractionated prostate tissue digests (PCT) and measured in DDA mode on the same type of instrument, over a 2-h gradient. The original analysis resulted in 3,030 detected proteins and a median CV of 16.2%.

WyIm (PXD01494323)

Another PCT-SWATH dataset was included, this from diffuse large B-cell lymphoma (DLBCL). The samples were prepared from FFPE and spiked with iRT peptides before MS data measurement. The dataset contained 113 runs, acquired on a 6600 TripleTOF mass spectrometer with a 60-min LC gradient, in technical duplicates. The original analysis was conducted using a custom target library and resulted in 5,769 proteins detected.

PrCaRegions (PXD00349724)

This dataset included 60 PCT-SWATH-MS runs coming from prostatectomies of three individuals. The tissue samples were reported to include acinar prostate tumours and a ductal prostate tumour. Prepared samples were spiked with iRT peptides. Data acquisition was performed on a 5600 TripleTOF mass spectrometer with a 120-min LC gradient, with technical replicates for each selected region of the prostatectomies. The target library used for the original analysis was compiled by combining targets from prostate tissue digest runs and the DDA files from the pan-human combined assay library (CAL). In the original analysis, 6,873 proteins were reported to be consistently quantified, 3,700 proteins of those with a high correlation. The dataset included additional SWATH-MS runs of 36 human liver S9 fractions (HLS9) samples and three pooled human liver microsomes (HLM) acquired with the same instrumentation but using a 45-min LC gradient.

PrCaNetwork (PXD00458925)

This dataset contained 210 SWATH-MS runs from PrCa tissue. The tissue material was sourced from 39 individuals. The sample preparation was performed with PCT and spiked with 10% iRT peptides. Data acquisition was performed on a 5600 TripleTOF mass spectrometer, with a 120-min LC gradient with technical replicates. The analysis target library was custom built from 422 DDA MS runs of prostate tissue samples. Overall, 2,371 quantified proteins were reported.

BraCaOFLM4 (PXD01419426)

This dataset contained 52 breast cancer (BrCa) samples, consisting of 13 estrogen receptor and/or progesterone receptor (ERPR) positive cases, 13 Her2 (Her) positive, 13 estrogen/progesterone/her2 (ERPRHer) positive, and 13 triple negative (TN) breast tumours, and a cohort of 20 ductal carcinoma in situ (DCIS) samples. Overall, 145 SWATH-MS runs were prepared from FFPE samples, measured on a 5600+ TripleTOF mass spectrometer, using a 95-min LC gradient. A custom library containing 3,200 proteins was used in the analysis, and resulted in 1,313 detected proteins, of which 1,064 were consistently quantified.

NCI-60 (PXD00353927)

This dataset titled "Quantitative Proteome Landscape of the NCI-60 Cancer Cell Lines" consisted of a cell line panel that contained 60 cell lines coming from 9 different tissue types. Each cell line was measured in technical duplicates. The 120 MS runs were acquired with the PCT-SWATH method on a 5600 TripleTOF mass spectrometer with a 120-min LC gradient. A custom target library from DDA MS runs obtained from the same samples (containing 86,209 proteotypic peptide precursors in 8,056 proteotypic UniProtKB/Swiss-Prot proteins) was created, resulting in 3,171 consistently detected proteins at a <1% peptide and protein FDR level (with all proteins detected in technical replicates).

Plasma (PXD00106428)

For data with more complex expected variation, a twin study design dataset including longitudinal factors, was also selected. It contained 240 SWATH-MS runs measured on a 5600+ TripleTOF mass spectrometer with a 120-min LC gradient. Of these, 234 were measured from blood sample pairs of 72 monozygotic and 44 dizygotic twins, ranging from 38 to 74 years of age. An additional 6 MS runs were measured from pooled samples as technical replicates. The sampling was done at two different time points: 5.2 ± 1.4 years apart. A custom target library was created for the original analysis, with targets from 652 proteins shotgun proteomics sequencing of mixed plasma samples and CAL, resulting in targets representing 1,667 unique plasma proteins. Originally, an average of 425 proteins were identified, 342 of which were consistently quantified in each twin sample at a 1% protein FDR.

DIATPA (PXD01091229)

This human liver dataset consisted of overall 43 SWATH-MS runs. From these 43 runs, 36 were measured from individual HLS9 samples sourced from 17 males and 19 females, aged 23-81 years, 3 MS runs of pooled HLS9 samples from the same cohort, and 3 MS runs from pooled human liver microsomes (HLM) from 100 males and 100 females, aged 11–83 years. Analysis was done on a 5600+ TripleTOF mass spectrometer with a 90-min LC gradient. A custom target library was constructed from DDA runs of the cohort of 36 samples. At a 1% protein FDR, on average, 1,250 proteins were quantified per each HLS9 sample.

SWATH-MS data analysis

The data analysis workflows were constructed using Nextflow. This choice allowed the data processing to be executed either in single-computer mode, on HPC clusters, or on cloud computing platforms. The SWATH-MS data analysis workflow steps (Figure 1B,C) can be broken down into raw data conversion, QC/QA and SWATH-MS window processing, OpenSWATH target generation and data analysis, FDR analysis and multi-run alignment with PyProphet[20] and TRIC[34] (Figure 1B), followed by a separate workflow for statistical analysis performed with R and MSstats (Figure 1C), concluded by result inspection and upload to Expression Atlas via custom submission scripts (Figure 1D). The workflows included in the pipeline were built for top-to-bottom analysis, starting with conversion of the raw instrument data, followed by statistical data processing with MSstats. The Nextflow workflows are split after the multi-run alignment step and equipped with optional entry points for pre-processed data from OpenSWATH[9] (*.osw* files) to enable more flexible compute settings. All analysis software was containerised either from available software releases or built-for-purpose to ensure a well-defined compute environment and software compatibility. All container 'recipes' are included in the GitHub repository (https://gitlab.ebi.ac.uk/pride-reanalysis/dia-reanalysis/) and can be rebuilt for local use (see Table 3).

The inputs (Table 4) for the Nextflow pipeline consisted of the SWATH-MS runs as a collection of *.wiff/.scan* files, together with a descriptor file in *.TraML* format, detailing the iRT peptides used. As target input, the pipeline consumes a target library file in an OpenSWATH conformant *.tsv* format. In case tool parameters needed changing, an additional *.yaml* file can be provided with updated parameters settings. If a target library in *.pqp* format has been created already or the analysis of processed *.osw* files has already been conducted with different downstream parameters (FDR thresholds), these can be used as alternative entry points in the Nextflow workflow. For the last Nextflow workflow, the study design in an MSstats conformant format must also be provided.

Target library and FDR control and statistical analysis

As target library input for the analysis, we used the combined assay library (CAL12), which was originally envisioned as a generic large-scale human assay library to support human SWATH-MS studies. It consists of 1,164,312 transitions identifying 139,449 proteotypic peptides and 10,316 proteins, considering only non-redundant entries of UniProtKB/Swiss-Prot. It therefore supports the detection and quantification of 50.9% of all human proteins, as annotated by UniProtKB/Swiss-Prot. It was generated by combining the results from 331 measurements of fractions from different sample types with a technical variation between replicates of below 20% for the quantified signals at a precursor level. For the "Plasma" dataset, CAL was filtered for blood/plasma proteins (as annotated in UniProtKB/Swiss-Prot) to reflect the changed base assumption of potentially present proteins. Protein identification was performed with OpenSWATH and PyProphet as described in [20,35]. PyProphet was used to calculate the confidence scores on peptide level (global context) and protein level (global context) in parallel at the given FDR thresholds. Decoys were generated with the OpenSWATH tool OpenSwathDecoyGenerator (using default parameters) and the target library assay was customised to the input dataset's SWATH-MS window setting with OpenSWATHAssayGenerator.

The proteomic analysis were performed at distinct levels (5%, 1% and 0.1%) of peptide and protein FDR, and TRIC was used to align the SWATH-MS runs detected features. The results were then processed with MSstats for protein inference and quantitative analysis. As protein inference method, the 'top3' method was chosen (for the impact of choosing the default 'all' protein inference method, see the 'Results' section and Supplementary Figure S4). The abundance values were median normalised and log_2 transformed. Normalisation using MSstats methods was replaced by a more conservative median normalisation approach. Further, a consistency filter was also applied to filter out proteins with $\leq 50\%$ of all target features of the protein detected over the respective study groups' MS runs. If the study included multiple study groups and protein expression contrasts of biological interest, additionally, a differential expression analysis was performed, and fold-change contrasts were calculated (see Table 1, "Analysis type" column) between study groups. Between the respective groups, the protein log_2 fold-change was computed from the mean of each group's expression value and the significance level controlled ($\alpha = 0.05$, R software, Welch t-test, Benjamini & Hochberg).

Technical validation

The resulting protein abundances amongst technical replicates were used to calculate the CV for the datasets. The median CV was calculated per study group and subject, from the median CV for protein abundances in each pair of technical replications and compared to the initially reported CV (Figure 3). For each reanalysed dataset containing technical replicates, the correlation

of each study's pairs of technical replicates was also investigated (Figure 4). Where available, we additionally analysed the correlation of originally reported protein abundances against our reanalysis results. The correlations were measured with the Pearson product-moment correlation coefficient.

Integration of the results in Expression Atlas

The data integration into Expression Atlas was performed on the results filtered with an FDR threshold of 1% from global context peptide and protein FDR. The MyGene.info R client (version 1.24.0, Ensembl 99/GRCh38) was used to map UniProtKB/Swiss-Prot protein accessions to Ensembl gene IDs. Protein groups with mappings to more than one Ensembl gene ID or decoys and targets without mappings were removed. The filtered and median normalised quantitative results per MS run were integrated into Expression Atlas. In the case of differential datasets, the log_2 fold-changes with $-log_{10}$ p-values (adjusted) were also submitted with information on the compared sample groups instead. The corresponding annotated SDRF files are also available. Expression Atlas URLs for each dataset are indicated in Table 1.

Data availability

All input data was downloaded from PRIDE. Experimental design annotations (*.sdrf, .txt*) and reanalysis outputs are available in the reanalysis repository under the respective *PXD* folder. Table 1 contains the collected information on the datasets, Table 5 information on the result availability, Supplementary Table S4 intermediate result availability.

Code availability

The complete open reanalysis pipeline description and documentation, workflows, container recipes, and custom code and visualisation scripts, as well as parameter input files are available through the GitHub repository at https://gitlab.ebi.ac.uk/pride-reanalysis/dia-reanalysis/.

Acknowledgements

First of all, we would like to thank all data submitters who made their datasets available in PRIDE. This work has been funded by the BBSRC grant "Proteomics-DIA" [grant numbers BB/P024599/1 and BB/P024424/1]. The authors would also like to thank H. Roest for his help in the implementation of the analysis pipeline. JAV, AP and DGS would also like to acknowledge The Wellcome Trust [grant number 208391/Z/17/Z] and EMBL-core funding.

Author contributions statement

M.W., A.P., and D.G.S. designed the analysis pipeline, A.P., D.G.S. and M.W. analysed the datasets and integrated the results into ExpressionAtlas. P.B, P.C., R.L.G. S.J.H, and J.A.V. consulted on the pipeline development, N.G., S.M., P.M., and I.P. contributed to the integration of the data in ExpressionAtlas. S.J.H, R.L.G. and J.A.V. acquired the funding. M.W, S.J.H., and J.A.V. wrote the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare that there is no conflict of interest.

Figures & Tables

Table 1. Main characteristics of the selected public DIA datasets for data reanalysis. Further details can be found in the 'Data availability' section.

Dataset Identifier	Analysis Type	Short Name	Dataset Size (MS runs)	Technical Replicates	Expression Atlas Accession Number
PXD004873[36]	Baseline	HCCpct	76	Available	E-PROT-69
PXD000672[13]	Differential	Digital Kidney	48	Not available	E-PROT-59
PXD004691[37]	Differential	Bank PrCa	224	Available	E-PROT-68
PXD014943[37]	Differential	Wylm	113	Not available	E-PROT-67
PXD003497[23]	Baseline	PrCa Regions	60	Available	E-PROT-66
PXD004589[38]	Baseline	PrCa Network	210	Not available	E-PROT-70
PXD014194[39]	Baseline	BraCa OFLM4	145	Available	E-PROT-72
PXD003539[40]	Baseline	NCI60	120	Not available	E-PROT-73
PXD001064[41]	Baseline	Plasma	240	Pooled sample replicates	E-PROT-60
PXD010912[42]	Baseline	DIATPA	42	Not available	E-PROT-71

Table 2. Median CV values for technical replicates, for the reanalysed results and originally published data, respectively.

Dataset identifier	Reanalysed data	Original data
PXD014194	16.8	19.0
PXD004873	7.14	6.05
PXD003497	20.8	20.3

Table 3. Overview of the containers and software versions used in the open data analysis pipeline.

Step	Name	URL or DockerHub handle	Version
Conversion from raw file to mzML	wiffConverter	sciex/wiffconverter:0.7	0.7.0
QC/QA	yamato	https://github.com/PaulBrack/Yamato/releases/download/v1.0.4/release-linux-x64.zip	1.0.4
Window management	python scripts	https://gitlab.ebi.ac.uk/pride-reanalysis/dia-reanalysis/	1.0.0
OpenSWATH	OpenSWATH 9	openswath/Openswath:0.1.2	2.4.0 (git 868546e)
	PyProphet 41		2.0.dev1 (git ddcedac)
	TRIC 42	msproteomicstools	0.8.0 (git eeed765)
Post-processing	R 43	https://gitlab.ebi.ac.uk/pride-reanalysis/dia-reanalysis/	4.0.3
	MSstats 30		3.22.0
	MyGene.info 45		1.24.0, Ensembl 99/GRCh38

Table 4. Types of input to the DIA reanalysis pipeline implemented in Nextflow.

Analysis pipeline inputs
An iRT file in OpenSWATH conformant *.TraML* format
A collection of *.wiff/.scan* files (alternatively pre-processed *.osw* files)
Target library file in OpenSWATH conformant *.tsv* format (alternatively a prepared *.pqp* file)
Study design annotation in MSstats conformant format and SDRF format (*.txt/.sdrf*)
Parameter *.yaml* file (in case the defaults need to be changed)
Output upstream Nextflow
FDR filtered alignment results of transition quantifications (*.tsv*)
QC records for each MS run (*.json*)
Output downstream Nextflow
MSstats analysis object (*.rda*)
Dataset analysis report (*.pdf*)
Prefiltered Expression Atlas upload input (*.tsv*)

Table 5. Detailed information on the datasets included in the reanalysis. *The current URLs correspond to the development instance of Expression Atlas. In the next Expression Atlas release, these datasets will be moved to the production instance. At that point links will be updated to the corresponding URLs starting by https://www.ebi.ac.uk/gxa/experiments/

Dataset identifier	Originally reported protein quantification	Condition contrasts available	Condition common Proteins	Expression Atlas URL*
PXD004873	Per MS run	-	-	https://wwwdev.ebi.ac.uk/gxa/experiments/E-PROT-69/Results
PXD000672	Per MS run	2	975; 1022	https://wwwdev.ebi.ac.uk/gxa/experiments/E-PROT-59/Results
PXD004691	Per sample	2	1435; 991	https://wwwdev.ebi.ac.uk/gxa/experiments/E-PROT-68/Results
PXD014943	Per sample	2	3019; 1080	https://wwwdev.ebi.ac.uk/gxa/experiments/E-PROT-67/Results
PXD003497	Per MS run	-	-	https://wwwdev.ebi.ac.uk/gxa/experiments/E-PROT-66/Results
PXD004589	Per MS run	-	-	https://wwwdev.ebi.ac.uk/gxa/experiments/E-PROT-70/Results
PXD014194	Per MS run	-	-	https://wwwdev.ebi.ac.uk/gxa/experiments/E-PROT-72/Results
PXD003539	-	-	-	https://wwwdev.ebi.ac.uk/gxa/experiments/E-PROT-73/Results
PXD001064	-	-	-	https://wwwdev.ebi.ac.uk/gxa/experiments/E-PROT-60/Results
PXD010912	-	-	-	https://wwwdev.ebi.ac.uk/gxa/experiments/E-PROT-71/Results

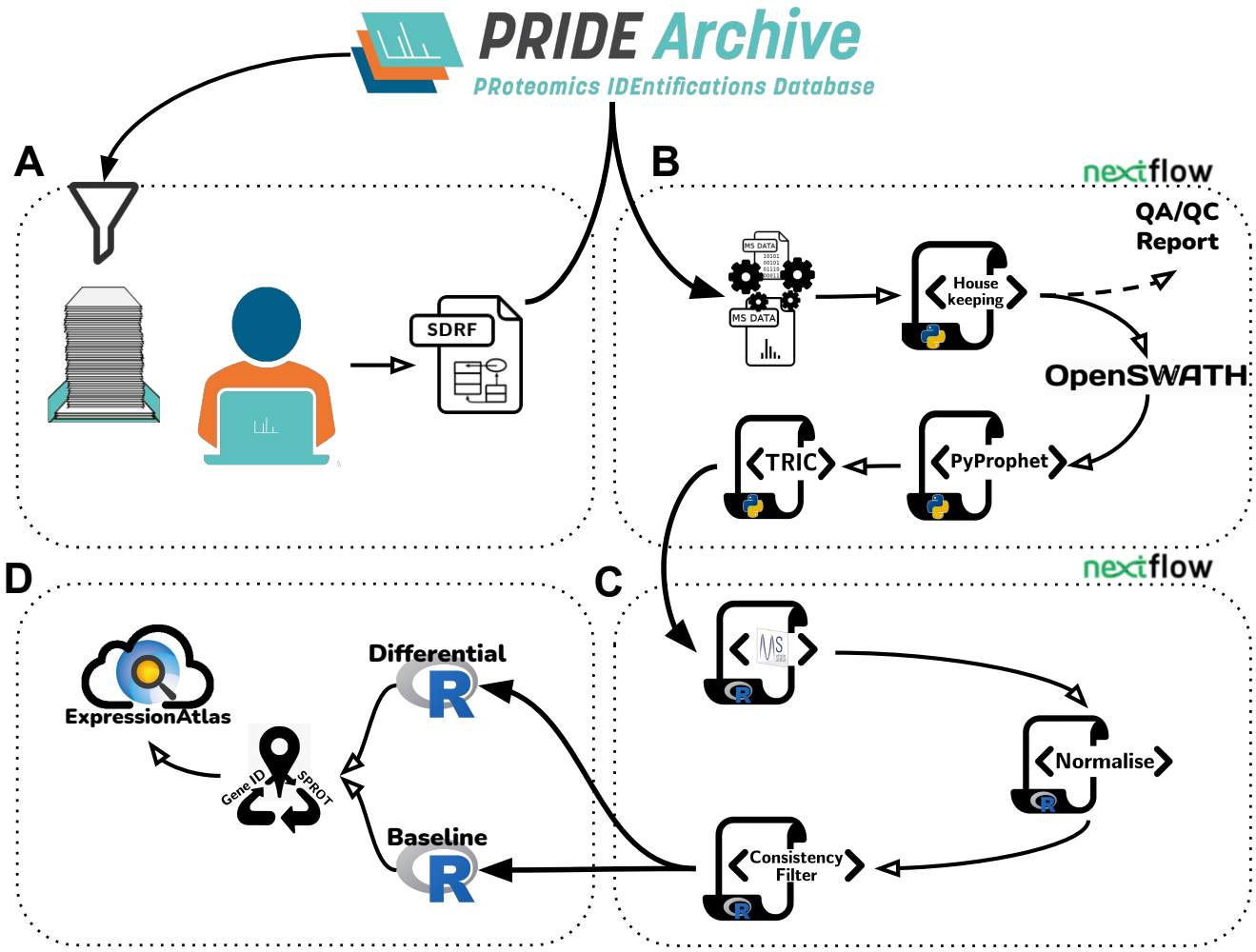

Figure 1. Graphical representation of the DIA data reanalysis pipeline, consisting of 4 parts. A) Data curation: Metadata annotation protocols and dataset acquisition. B) SWATH-MS data analysis: Nextflow workflow including steps ranging from data conversion, SWATH-MS window management, data quality assessment and control (QA/QC), OpenSWATH analysis, FDR calculation to measurement alignment. C) Statistical analysis: Nextflow workflow for MSstats analysis, normalisation and result filtering. D) Data integration: Data preparation, accession mapping and integration into Expression Atlas.

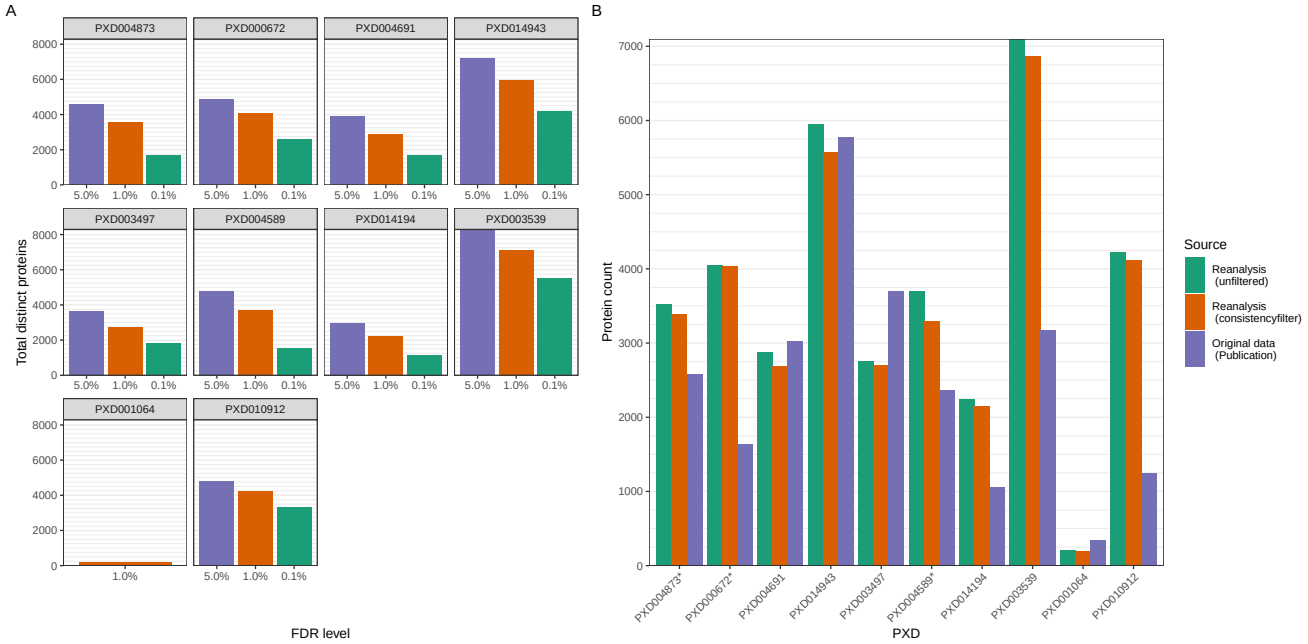

Figure 2. A) Number of detected proteins per dataset at different FDR levels in the data reanalysis. B) Protein detection results after 1% protein FDR threshold filtering. Original data refers to the respective publication's mentioned protein numbers, reported at 1% protein FDR unless indicated otherwise. Reanalysis numbers are provided unfiltered and with the consistency filter applied (at least 50% of all protein's peptide fragment targets have to be detected within a study group).*Proteins coming from datasets PXD000672 and PXD004873 were reported in the original publication at a 0.1% protein level FDR only. In the case of dataset PXD004589 at 0.1% peptide level FDR was reported.

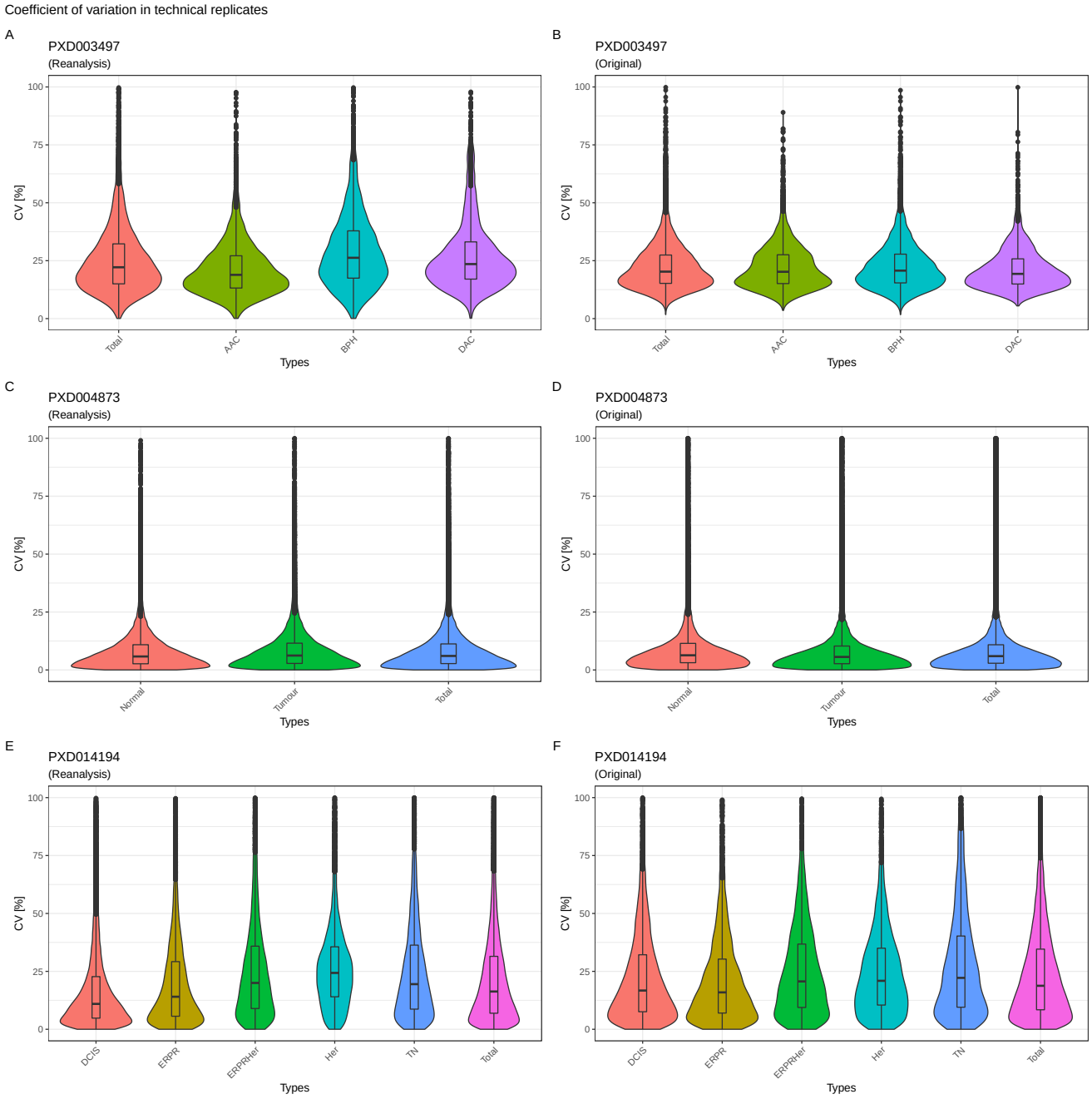

Figure 3. Violin-plots showing the results of the group-wise CV comparisons: A) PXD003497 reanalysis; B) PXD003497 original data; C) PXD004873 reanalysis; D) PXD004873 original data; E) PXD014194 reanalysis; F) PXD014194 original data. As it can be seen from the similar size and shapes of the violin-plots, the CVs across the datasets are largely concordant.

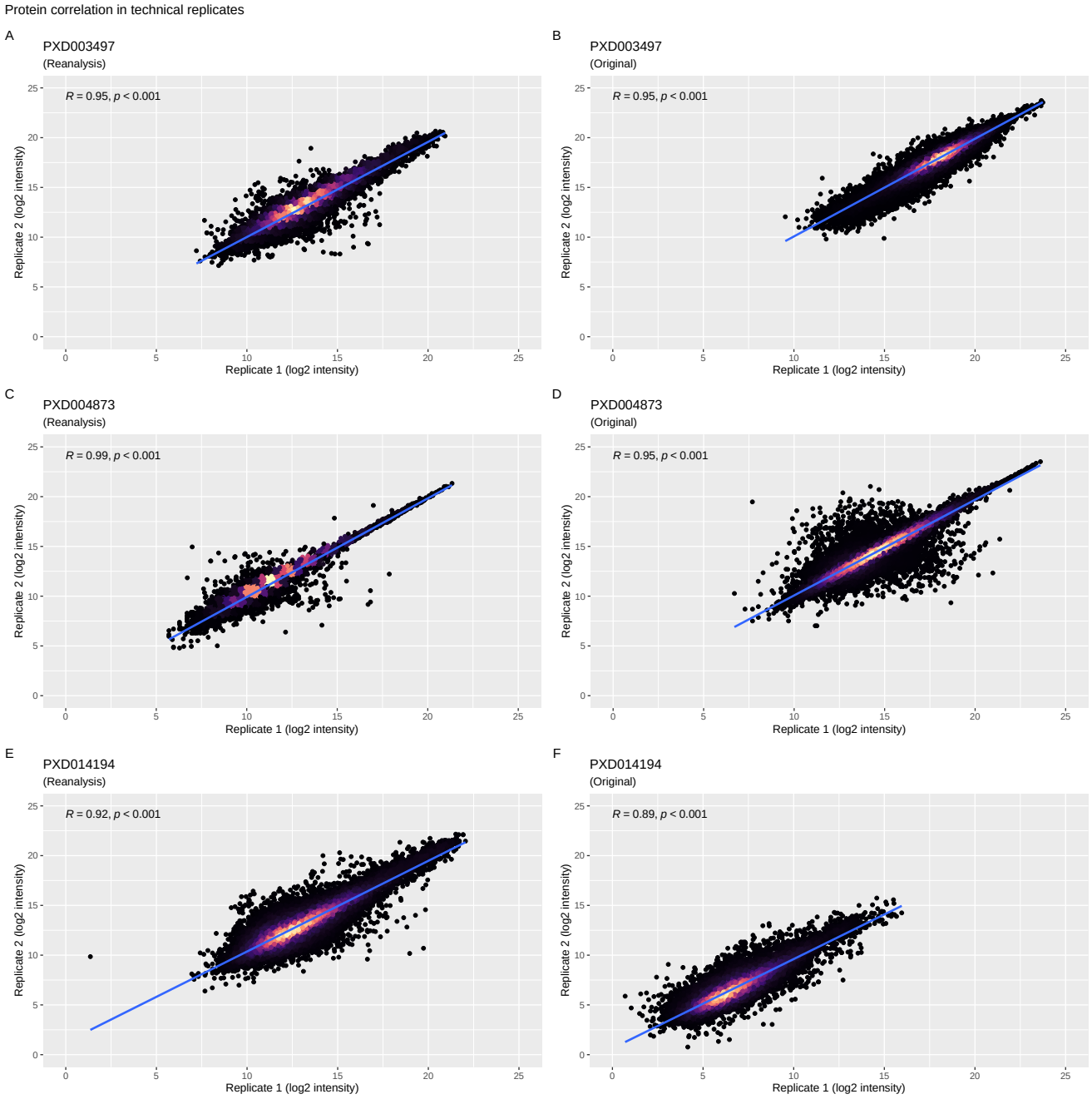

Figure 4. Correlation analysis of reported log2 protein intensities from technical replicate pairs: A) PXD003497 reanalysis; B) PXD003497 original data; C) PXD004873 reanalysis; D) PXD004873 original data; E) PXD014194 reanalysis; F) PXD014194 original data. The first items of pairs are on the x-axis and second items are on the y-axis. Each point represents a protein. The point density is indicated by the colour gradient, with black showing the lowest density. The higher the density the lighter the colour becomes.

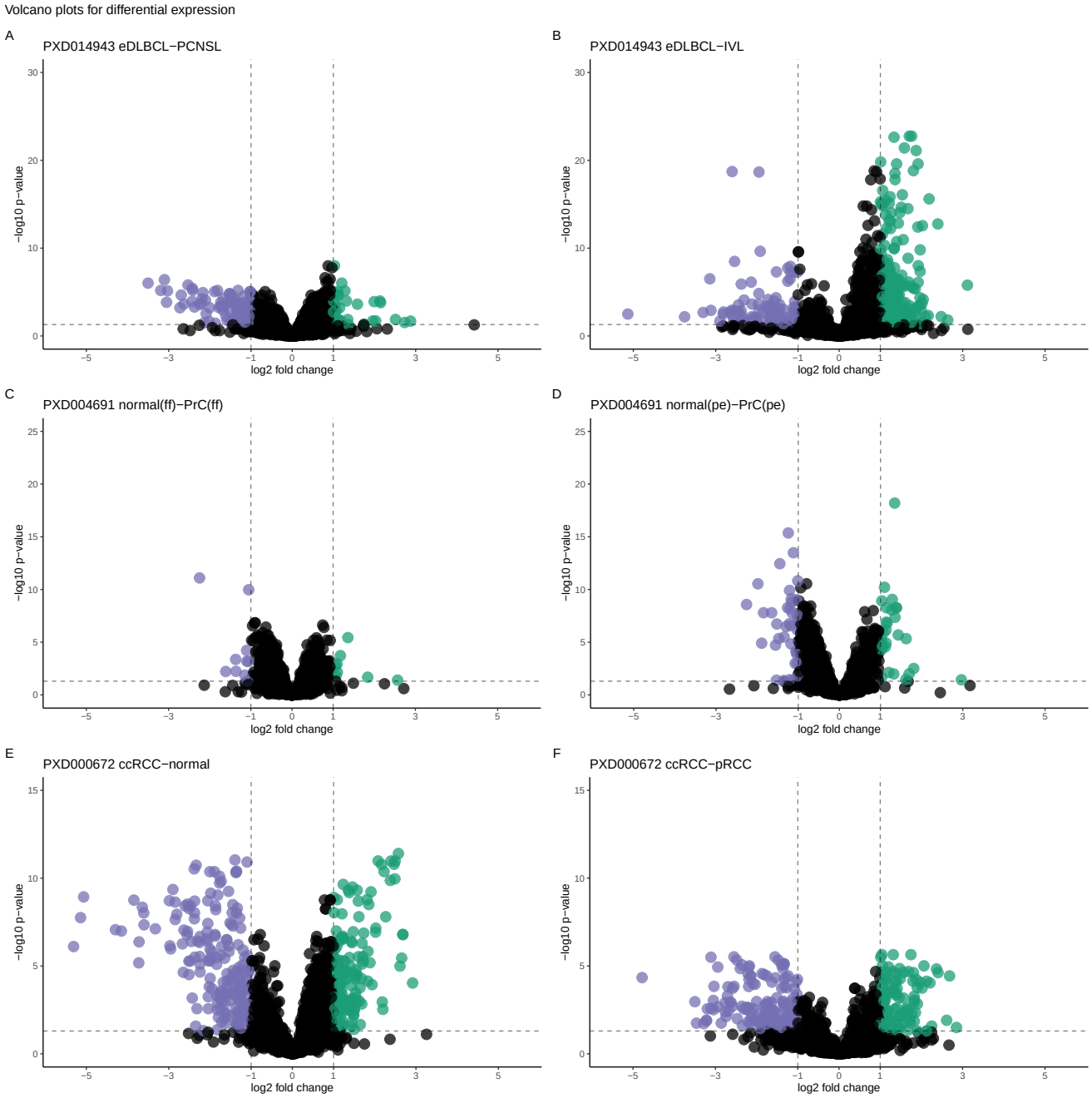

Figure 5. Volcano plots corresponding to the differential expression analysis for dataset PXD014943: A) extranodal diffuse large B-cell lymphoma (eDLBCL) versus primary central nervous system lymphoma (PCNSL); B) intravascular lymphoma (IVL) versus eDLBCL. For dataset PXD004691: C) normal tissue (fresh frozen) versus PrC (fresh frozen); D) normal tissue (paraffin embedded) versus tumour tissue (paraffin embedded). For dataset PXD000672: E) benign tissue samples versus clear cell RCC; F) clear cell RCC versus paillary RCC. The FC compared are represented by points on the plot. Significant FC proteins are colour indicated, dashed lines indicate the fold-change cutoff of 2 and the (adjusted) p-value cutoff at 0.05.

References

1. Perez-Riverol, Y. & for Mass Spectrometry, E. B. C. Toward a sample metadata standard in public proteomics repositories. *J. Proteome Res.* **19**, 3906–3909, 10.1021/acs.jproteome.0c00376 (2020).
2. Deutsch, E. W. *et al.* The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. *Nucleic Acids Res.* **48**, D1145–D1152, 10.1093/nar/gkz984 (2020).
3. Vaudel, M. *et al.* Exploring the potential of public proteomics data. *Proteomics* **16**, 214–225, 10.1002/pmic.201500295 (2016).
4. Martens, L. & Vizcaíno, J. A. A golden age for working with public proteomics data. *Trends Biochem. Sci.* **42**, 333–341, 10.1016/j.tibs.2017.01.001 (2017).
5. Ochoa, D. *et al.* The functional landscape of the human phosphoproteome. *Nat. Biotechnol.* **38**, 365–373, 10.1038/s41587-019-0344-3 (2020).
6. Bouwmeester, R., Gabriels, R., Van Den Bossche, T., Martens, L. & Degroeve, S. The age of data-driven proteomics: How machine learning enables novel workflows. *Proteomics* **20**, e1900351, 10.1002/pmic.201900351 (2020).
7. Papatheodorou, I. *et al.* Expression atlas update: from tissues to single cells. *Nucleic Acids Res.* **48**, D77–D83, 10.1093/nar/gkz947 (2020).
8. Jarnuczak, A. F. *et al.* An integrated landscape of protein expression in human cancer. *Sci. data* **8**, 115, 10.1038/s41597-021-00890-2 (2021).
9. Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **32**, 219–223, 10.1038/nbt.2841 (2014).
10. 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, 10.1074/mcp.O111.016717 (2012).
11. Ludwig, C. *et al.* Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **14**, e8126, 10.15252/msb.20178126 (2018).
12. Rosenberger, G. *et al.* A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. data* **1**, 140031, 10.1038/sdata.2014.31 (2014).
13. Guo, T. *et al.* Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nat. Medicine* **21**, 407–413, 10.1038/nm.3807 (2015).
14. Selevsek, N. *et al.* Reproducible and consistent quantification of the saccharomyces cerevisiae proteome by SWATH-mass spectrometry. *Mol. & Cell. Proteomics* **14**, 739–749, 10.1074/mcp.M113.035550 (2015).
15. Collins, B. C. *et al.* Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat. Commun.* **8**, 291, 10.1038/s41467-017-00249-5 (2017).
16. Deutsch, E. W., Lam, H. & Aebersold, R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Reports* **9**, 429–434, 10.1038/embor.2008.56 (2008).
17. Bouchal, P. *et al.* Breast cancer classification based on proteotypes obtained by SWATH mass spectrometry. *Cell reports* **28**, 832–843.e7, 10.1016/j.celrep.2019.06.046 (2019).
18. Weerakoon, H. *et al.* A primary human t-cell spectral library to facilitate large scale quantitative t-cell proteomics. *Sci. data* **7**, 412, 10.1038/s41597-020-00744-3 (2020).
19. Escher, C. *et al.* Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* **12**, 1111–1121, 10.1002/pmic.201100463 (2012).
20. Rosenberger, G. *et al.* Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat. Methods* **14**, 921–927, 10.1038/nmeth.4398 (2017).
21. Navarro, P. *et al.* A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* **34**, 1130–1136, 10.1038/nbt.3685 (2016).
22. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319, 10.1038/nbt.3820 (2017).
23. Guo, T. *et al.* Multi-region proteome analysis quantifies spatial heterogeneity of prostate tissue biomarkers. *Life Sci. Alliance* **1**, 10.26508/lsa.201800042 (2018).

24. Team, R. C. R: A language and environment for statistical computing (2020).

25. Choi, M. *et al.* MSstats: an r package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **30**, 2524–2526, 10.1093/bioinformatics/btu305 (2014).

26. Tiwary, S. *et al.* High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat. Methods* **16**, 519–525, 10.1038/s41592-019-0427-6 (2019).

27. 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, 10.1038/s41592-019-0638-x (2020).

28. Peters, S., Hains, P. G., Lucas, N., Robinson, P. J. & Tully, B. A case study and methodology for openswath parameter optimization using the procan90 data set and 45810 computational analysis runs. *J. Proteome Res.* **18**, 1019–1031, 10.1021/acs.jproteome.8b00709 (2019).

29. Tsou, C.-C. *et al.* DIA-umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12**, 258–64, 7 p following 264, 10.1038/nmeth.3255 (2015).

30. Li, Y. *et al.* Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. *Nat. Methods* **12**, 1105–1106, 10.1038/nmeth.3593 (2015).

31. Mehta, D., Scandola, S. & Uhrig, R. G. BoxCar and library-free data-independent acquisition substantially improve the depth, range, and completeness of label-free quantitative proteomics in arabidopsis. *BioRxiv* 10.1101/2020.11.07.372276 (2021).

32. Yang, Y. *et al.* In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nat. Commun.* **11**, 146, 10.1038/s41467-019-13866-z (2020).

33. Van Puyvelde, B. *et al.* Removing the hidden data dependency of DIA with predicted spectral libraries. *Proteomics* **20**, e1900306, 10.1002/pmic.201900306 (2020).

34. Röst, H. L. *et al.* TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nat. Methods* **13**, 777–783, 10.1038/nmeth.3954 (2016).

35. Röst, H. L., Aebersold, R. & Schubert, O. T. Automated SWATH data analysis using targeted extraction of ion chromatograms. *Methods Mol. Biol.* **1550**, 289–307, 10.1007/978-1-4939-6747-6_20 (2017).

36. Zhu, Y. *et al.* High-throughput proteomic analysis of FFPE tissue samples facilitates tumor stratification. *Mol. Oncol.* **13**, 2305–2328, 10.1002/1878-0261.12570 (2019).

37. Zhu, Y. *et al.* Identification of protein abundance changes in hepatocellular carcinoma tissues using PCT-SWATH. *Proteomics. Clin. Appl.* **13**, e1700179, 10.1002/prca.201700179 (2019).

38. Charmpi, K. *et al.* Convergent network effects along the axis of gene expression during prostate cancer progression. *Genome Biol.* **21**, 302, 10.1186/s13059-020-02188-9 (2020).

39. Valo, I. *et al.* OLFM4 expression in ductal carcinoma in situ and in invasive breast cancer cohorts by a SWATH-based proteomic approach. *Proteomics* **19**, e1800446, 10.1002/pmic.201800446 (2019).

40. Guo, T. *et al.* Quantitative proteome landscape of the NCI-60 cancer cell lines. *iScience* **21**, 664–680, 10.1016/j.isci.2019.10.059 (2019).

41. Liu, Y. *et al.* Quantitative variability of 342 plasma proteins in a human twin population. *Mol. Syst. Biol.* **11**, 786, 10.15252/msb.20145728 (2015).

42. He, B., Shi, J., Wang, X., Jiang, H. & Zhu, H.-J. Label-free absolute protein quantification with data-independent acquisition. *J. Proteomics* **200**, 51–59, 10.1016/j.jprot.2019.03.005 (2019).