Journal of proteome research

pubs.acs.org/jpr

Technical Note

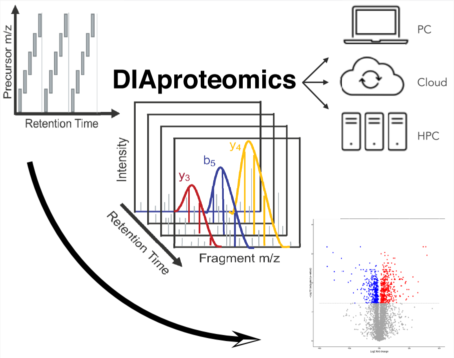
DIAproteomics: A Multifunctional Data Analysis Pipeline for Data-Independent Acquisition Proteomics and Peptidomics

Leon Bichmann,* Shubham Gupta, George Rosenberger, Leon Kuchenbecker, Timo Sachsenberg, Phil Ewels, Oliver Alka, Julianus Pfeuffer, Oliver Kohlbacher, and Hannes Röst

Cite This: J. Proteome Res. 2021, 20, 3758−3766

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Data-independent acquisition (DIA) is becoming a leading analysis method in biomedical mass spectrometry. The main advantages include greater reproducibility and sensitivity and a greater dynamic range compared with data-dependent acquisition (DDA). However, the data analysis is complex and often requires expert knowledge when dealing with large-scale data sets. Here we present DIAproteomics, a multifunctional, automated, high-throughput pipeline implemented in the Nextflow workflow management system that allows one to easily process proteomics and peptidomics DIA data sets on diverse compute infrastructures. The central components are well-established tools such as the OpenSwathWorkflow for the DIA spectral library search and PyProphet for the false discovery rate assessment. In addition, it provides options to generate spectral libraries from existing DDA data and to carry out the retention time and chromatogram alignment. The output includes annotated tables and diagnostic visualizations from the statistical postprocessing and computation of fold-changes across pairwise conditions, predefined in an experimental design. DIAproteomics is well documented open-source software and is available under a permissive license to the scientific community at https://www.openms.de/diaproteomics/.

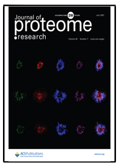
KEYWORDS: *data-independent acquisition, spectral library generation, automation, cloud computing, data processing, proteomics, peptidomics*

INTRODUCTION

Recently, data-independent acquisition (DIA) using sequential windowed acquisition of all theoretical fragment ion mass spectra (SWATH-MS)[1] has attracted much attention in the field of protein and peptide mass spectrometry due to its ability to overcome the shortcomings of the classical data-dependent (DDA) strategy.[2−11] Because of its outstanding performance in reproducibility and quantification, DIA is likely to become the state-of-the-art technology in clinical mass spectrometry (MS).[12] In addition, recent tailored applications of DIA have enabled new approaches for the chemoproteomic screening of drug targets[13] and the improved discovery of post-translational modifications.[14,15] The main advantages are its capacity to (1) acquire fragment spectra in a reproducible grid-based fashion over the entire mass and retention time range, (2) sample fragment spectra for nearly all precursor ions present in a sample, and (3) enable the tracing of elution profiles of fragments and integrate their quantities in a greater dynamic range.[16] Yet this comes at the cost of increased complexity of the acquired mass spectra due to the simultaneous fragmentation of multiple precursor ions, which requires appropriate methods for spectra identification.[17] Nonetheless, DIA has promising potential to achieve a greater identification rate and quantification range, higher reproducibility, and fewer missing values than DDA.

A key step to process DIA data is the generation of high-quality spectral libraries to identify the complex DIA spectra with higher sensitivity.[18] These spectral libraries can be derived from previously acquired DDA measurements by selectively annotating and storing peak intensities and other properties from confident peptide spectrum matches across multiple samples. Public repositories such as PRIDE,[19] the PeptideAtlas Project,[20] the SWATHAtlas,[21] and the SysteMHC Atlas[22] provide collections of aggregated spectral libraries from large DDA data sets such as the human proteome or spectral libraries of other species or specific contexts.[23] Alternatively, recently developed *in silico* methods that utilize advanced machine-learning strategies to predict peptide fragment intensities can be applied.[24−27] However, the library should match the settings of the instrument and the acquisition

Received: February 12, 2021
Published: June 21, 2021



method to which the respective DIA experiment will be compared, as different instruments, ionization methods, and corresponding parameters such as collision energies produce vastly different fragment spectra patterns. Finally, as an additional alternative, library-free approaches for the deconvolution of DIA data have been proposed to overcome the limitations and dependencies of spectral libraries.[15,17,28]

With increasing amounts of MS measurements recorded in both DDA and DIA acquisition modes deposited in publicly available data repositories,[19] there is a need for automated high-throughput data analysis pipelines. Because the parametrization of DIA search algorithms and the choice of a spectral library can strongly influence the analysis results, flexible and scalable software solutions for high-performance computing systems are required to provide ways to efficiently reprocess and compare analysis results using large amounts of existing data. This includes the automated generation of spectral libraries from available DDA measurements and the alignment of their transition retention times into the same space. Previously, multiple software solutions have been applied to process large-scale DIA data;[17,29−37] however, their application does not integrate the entire analysis workflow from spectral library generation to statistical output processing in a completely automated fashion, and it is not easily portable to various computerized systems.

We address this gap in available software solutions by introducing DIAproteomics, a complete, versatile, and high-throughput analysis pipeline for DIA proteomics and peptidomics MS measurements. It achieves a high degree of automation and scalability from single users to large high-performance computing (HPC) environments by integrating well-established tools such as the OpenSwathWorkflow[30] for DIA targeted extraction, provided through the OpenMS software toolbox for computational mass spectrometry.[38,39] The false discovery rate (FDR) is estimated by the PyProphet algorithm,[40] followed by chromatogram alignment as a postprocessing step using the DIAlignR software.[41] Moreover, its use provides the option to either specify a particular existing spectral library and retention time standards or generate the spectral library and select suitable pseudo-iRTs (internal retention time standards) from existing DDA measurements and search results. Ultimately, statistical postprocessing provided through MSstats[42] ensures reliable analysis results.

DIAproteomics is containerized and implemented using the workflow language Nextflow,[43] leveraging the capabilities of the powerful Nextflow execution engine to seamlessly run on single desktop computers and to scale up to large-scale HPC or cloud environments. As part of the nf-core repository for reproducible bioinformatics workflows,[44] it adheres to the corresponding strict standards. Ultimately, a web-browser-based user interface accompanies the workflow and allows easy-to-use parametrization and execution.

■ METHODS

Pipeline Architecture

DIAproteomics is an automated analysis pipeline that can be broadly partitioned into the following parts: optional spectral library and iRT generation from provided DDA data, optional spectral library merging and retention time (RT) alignment, DIA library search, FDR estimation, MS2 chromatogram alignment across runs, and output summarization (Figure 1). Each of these parts involves one or more required or optional steps within the workflow (Supporting Information, Table S1 and Figure S1). An experimental design needs to be provided in the form of an input sample sheet specifying the DDA and DIA samples, libraries, or iRT standards that should be coprocessed in one batch.

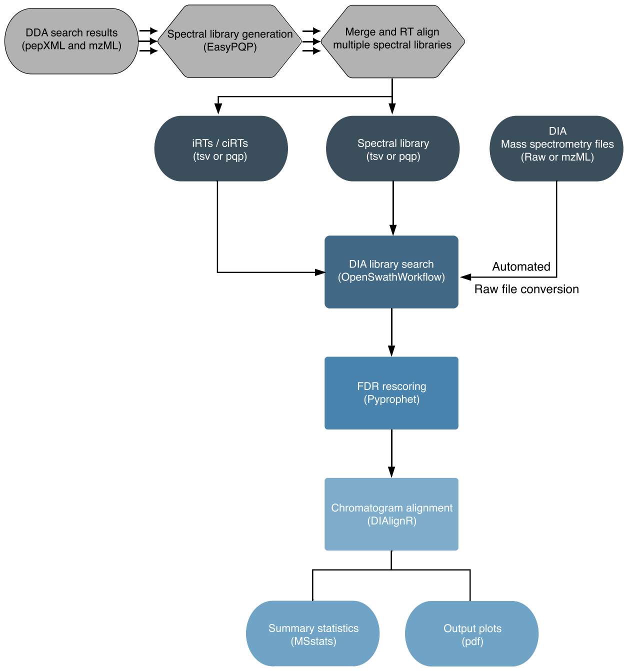


Figure 1. Simplified scheme of the DIAproteomics workflow. The input to the pipeline can be either spectral libraries and iRTs generated and combined from DDA raw data (optional, in gray) or an existing spectral library and internal retention time standards (iRTs). Next, targeted extraction is performed by searching the DIA-SWATH-MS raw files with the spectral library using the OpenSwathWorkflow. The false discovery rate (FDR) is assessed by subsequently applying PyProphet. Next, chromatograms are aligned using the DIAlignR software. Finally, the output is statistically postprocessed with MSstats and visualized.

Spectral Library Generation. In a first, optional step, the provided DDA raw MS measurements (Thermo Raw vendor format) are converted to the open, XML-based mzML format.[45] Next, the library is generated using EasyPQP (available at https://github.com/grosenberger/easypqp), which matches the provided search results (for example, in pepXML format) and the corresponding DDA raw measurements to annotate and store peptide transitions and their properties in a tab-separated table.[46] The library is transformed into an assay containing a specified number of transitions of b- and y-ions falling into a custom mass-to-charge range. Subsequently, decoy transitions that can be generated by OpenMS in multiple ways, such as reversed or shuffled, are added to the library. Finally, the generated library is exported in the peptide query parameter (pqp) SQLite-based data format. Optionally, all steps of the library and decoy generation can be skipped, and an existing library can be used instead.

Pseudo iRT Generation. If specified, a given number of highly confident peptide identifications spanning the entire RT range will be selected and exported to serve as iRT standards in the DIA library search step. This is important, for example, if no iRT standard kit was spiked into the samples before the DIA measurements. Selected iRTs will be exported in the peptide query parameter (pqp) SQLite-based data format. However, if provided, a set of user-defined iRTs can be used instead.

Spectral Library Merging. If multiple libraries per sample are provided, for example, when stemming from a set of technical replicates, then the libraries can be optionally merged and will then undergo a linear RT alignment onto the same reference. When merging is enabled, the best scoring peptide identification is kept in the library, omitting a lower scoring duplicate.

Spectral Library RT Alignment. When RT alignment is enabled, the multiple input spectral libraries will be pairwise aligned onto the same reference. This is achieved by computing a minimum spanning tree connecting all provided libraries by shared peptide overlap (Supporting Information, Figure S2). Hence the library having the highest overlap in shared peptides with all other libraries will be the central reference for the other libraries. Importantly, this strategy is also applicable when aligning onto the same reference very distant libraries that share no consensus peptide identifications among all libraries.[47] However, it requires peptides to be shared between all pairs of libraries, resulting in a connected tree.

DIA Spectral Library Search. In a first, optional step, the provided DIA raw MS measurements (Thermo RAW vendor format) may be converted to the mzML XML-based format. Next, DIA targeted extraction is carried out using the OpenSwathWorkflow, implemented within the OpenMS toolbox. The spectral library and iRT standards are used to individually search all input DIA raw measurements with a customizable parametrization. The SWATH windows can be determined from the data. Finally, extracted ion chromatograms (XICs) of the searched peptide transitions (mzML) are exported, and the output features and transition properties are stored in OpenSwathWorkflow files (osw).

False Discovery Rate Estimation. The OpenSwathWorkflow output files (osw) are merged sample-wise, as defined in the experimental design (sample sheet). The merged file is then scored using the PyProphet target-decoy FDR estimation procedure. Finally, the level of confidence, such as local transition-based or global peptide- or protein-level-based, can be defined.[40] The PyProphet scoring results will then be exported as a tab-separated table per DIA MS run, and the results will be visualized in a PDF report.

MS2 Chromatogram Alignment. As the last processing step, the extracted and scored MS2 chromatograms will be aligned using the DIAlignR software. This involves matching chromatograms between runs that can be aligned and integrating their transition areas. The sum of the integrated areas per peptide will be reported as peptide quantities in a TSV file. For this procedure, DIAlignR provides several FDR estimates that can be customized within the workflow to define cutoffs for transitions that should be excluded from matching between runs.[41,48] All DIAlignR intrinsic FDR measures are based on the prior PyProphet calculation. Hence, the statistical analysis is not changed throughout the procedure, but the values are used specifically to filter for groups of matched and unmatched peaks between runs.

Output Summarization. Depending on the user's choice, the output is summarized in a pairwise manner for each specified batch on the peptide or protein level using the MSstats postprocessing software.[42] For these comparisons, the biological and technical replicates and the differential conditions are taken into account according to the specifications in the input sample sheet. The quantities are derived based on the aggregated peptide-level intensity information resulting from the preceding chromatogram extraction and alignment steps. This input is read using the MSstats dataProcess function for preprocessing and quality control. (Runs with >50% missing values will be removed.) In the course of this, "highQuality" feature subsets keeping only informative features suitable for DIA SWATH data are used and summarized using the default robust estimation method "Tukey's median polish" (TMP), imputing missing values by the "accelerated failure model" (AFT).[49] In addition, the default equal median-based normalization method is applied. Finally, it is possible to export a number of diagnostic plots illustrating the peptide and protein identification results and their quantities and properties.

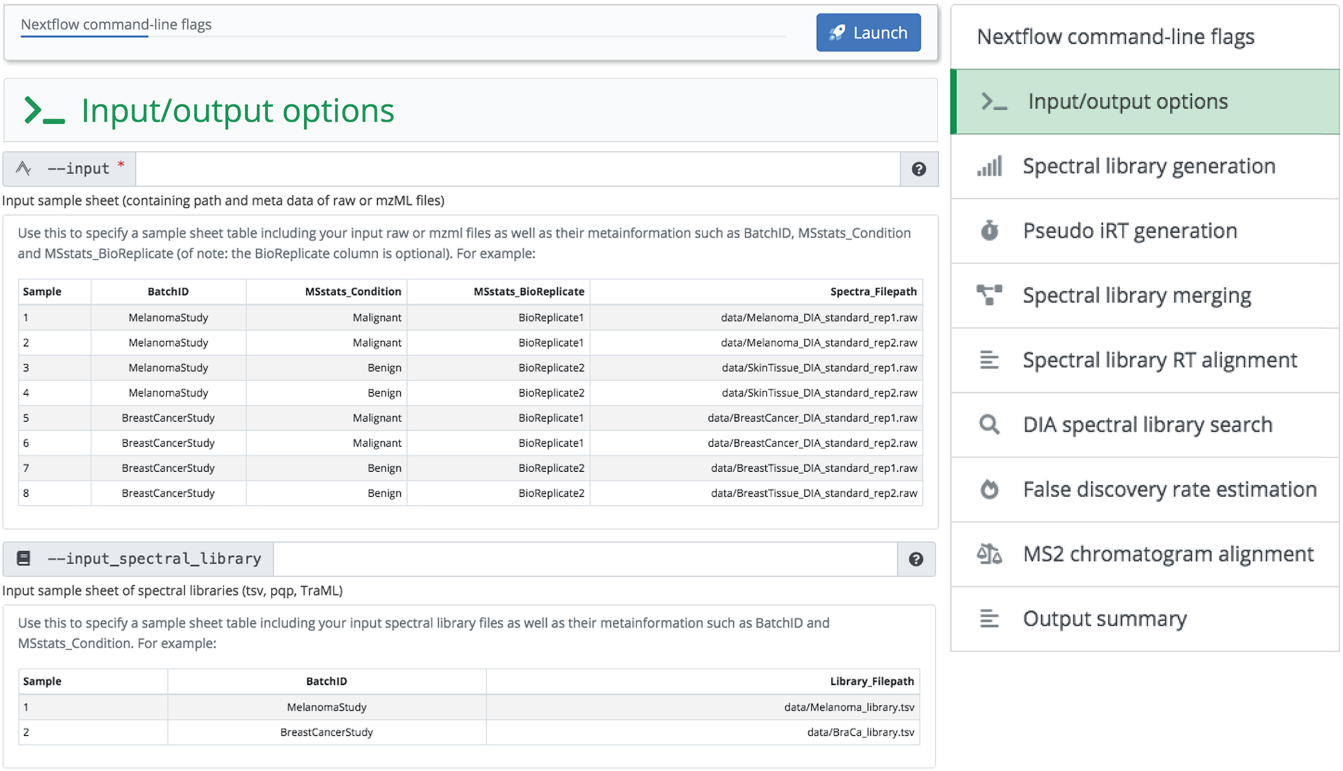
Implementation

The DIAproteomics pipeline is implemented in the Nextflow workflow programming language[43] based on the nf-core community template for reproducible bioinformatics workflows.[44] Support is provided for multiple functionalities, such as various container systems (e.g., docker, singularity, podman), environment management platforms (e.g., Conda), and user interfaces, as well as for execution on high-performance computing systems such as the Google cloud or Amazon Web Services (AWS). Each step of the workflow is executed as an independent process, allowing the efficient, parallel processing of large amounts of data.

Most of the inner functions and the file format handling is provided through the OpenMS v.2.5.0 toolbox for computational mass spectrometry.[39] Specifically, this includes the handling of spectral libraries, assay and decoy generation, and the implementation of the OpenSwathWorkflow.[30] Spectral library generation from DDA data is carried out by EasyPQP v.0.1.7. A customized Python v.3 script is executed to merge multiple libraries and to compute the minimum spanning tree for RT alignment using the module NetworkX v.2.4.[50] FDR estimation on merged OpenSwathWorkflow[30] output files is achieved using functionalities of PyProphet v.2.1.4,[40] and MS2 chromatogram alignment and the integration of peptide quantities are achieved using the "alignTargetedRuns" function of the DIAlignR software 1.2.0.[41] The "groupComparison" function using "highQuality" feature subsets within MSstats v. 3.20.1[42] is carried out to compute protein-level statistics and pairwise comparisons of protein fold-changes and significance across conditions. Finally, output visualizations are created using the R software libraries gplots and ggplot2.

Parametrization

The DIAproteomics workflow is highly flexible, and each execution step provides various parameters that can be customized for specific instrumental and experimental settings. An overview of the available parameters and a short description are provided at https://nf-co.re/diaproteomics. The default parametrization has been benchmarked multiple times in the past.[32,51] It involves spectral library assay generation with the six most intense b- and y-ion transitions falling into the precursor mass range of 400 to 1200 m/z and a fragment mass range of 350 to 2000 m/z. The default setting for the decoy transition generation is shuffling. The extraction of the MS1 precursor and the MS2 fragment transitions are carried out using a mass extraction window of 10 and 30 ppm, respectively, and an RT extraction window of 600 s for the


Nextflow command-line flags
Launch
Nextflow command-line flags
Input/output options
Spectral library generation
Pseudo iRT generation
Spectral library merging
Spectral library RT alignment
DIA spectral library search
False discovery rate estimation
MS2 chromatogram alignment
Output summary

Input/output options

--input *

Input sample sheet (containing path and meta data of raw or mzML files)

Use this to specify a sample sheet table including your input raw or mzml files as well as their metainformation such as BatchID, MSstats_Condition and MSstats_BioReplicate (of note: the BioReplicate column is optional). For example:

Sample	BatchID	MSstats_Condition	MSstats_BioReplicate	Spectra_Filepath
1	MelanomaStudy	Malignant	BioReplicate1	data/Melanoma_DIA_standard_rep1.raw
2	MelanomaStudy	Malignant	BioReplicate1	data/Melanoma_DIA_standard_rep2.raw
3	MelanomaStudy	Benign	BioReplicate2	data/SkinTissue_DIA_standard_rep1.raw
4	MelanomaStudy	Benign	BioReplicate2	data/SkinTissue_DIA_standard_rep2.raw
5	BreastCancerStudy	Malignant	BioReplicate1	data/BreastCancer_DIA_standard_rep1.raw
6	BreastCancerStudy	Malignant	BioReplicate1	data/BreastCancer_DIA_standard_rep2.raw
7	BreastCancerStudy	Benign	BioReplicate2	data/BreastTissue_DIA_standard_rep1.raw
8	BreastCancerStudy	Benign	BioReplicate2	data/BreastTissue_DIA_standard_rep2.raw

--input_spectral_library

Input sample sheet of spectral libraries (tsv, pqp, TraML)

Use this to specify a sample sheet table including your input spectral library files as well as their metainformation such as BatchID and MSstats_Condition. For example:

Sample	BatchID	Library_Filepath
1	MelanomaStudy	data/Melanoma_library.tsv
2	BreastCancerStudy	data/BraCa_library.tsv

Figure 2. Input/output options as available through the nf-core provided user interface. User-defined spreadsheets serve as input to the pipeline, defining the experimental design of raw files, the spectral libraries, their corresponding conditions, replicate meta information and batch identifiers (BatchID). Upon submission of the job, MS runs are grouped by their BatchID and coprocessed.

targeted extraction of the OpenSwathWorkflow. The FDR estimation is performed on a global protein level involving an LDA-based target-decoy separation. The MS2 chromatogram alignment requires transitions to satisfy several FDR thresholds. For global alignment, high-quality peaks are selected with the globalAlignmentFdr (set to 0.01) cutoff. A peak will only be matched across runs if at least one run has included it at an estimated FDR below of 0.01 (analyteFDR). It will then be compared to matching peaks in other runs below a higher maximum FDR threshold of 0.05 (MaxQueryFDR). This is an advantage over common strategies used in DDA to allow matching between runs because no FDR cutoff can be set for these approaches.

Reanalysis of Publicly Available Data Sets

A concise benchmark on the publicly available multicenter benchmark study data set by Navarro et al.[23] (PRIDE: PXD002952) was carried out using a human, *E. coli*, and yeast mixture HYE124. The ABsciex TripleToF 6600 and 64 variable SWATH window instrument setting was chosen, applying the default parametrization of DIAproteomic v1.1.0 and adjusting the precursor and fragment mass tolerances to 30 and 30 ppm, respectively. ABsciex wiff files were converted to mzML using the ProteoWizard msconvert software external to the pipeline.

Ultimately, to ensure the capability of the DIAproteomics pipeline v1.1.0 to process publicly available proteomics data sets, several HeLa cell line Thermo Orbitrap high-resolution MS runs from PRIDE project PXD003179[52] were reanalyzed using the default settings. The procedure was automated and integrated as a continuous integration full size test on AWS that can be actively run to verify the pipeline's functionality.

RESULTS AND DISCUSSION

Facilitation of Large-Scale DIA-SWATH-MS Analysis

DIAproteomics is a versatile analysis pipeline for the processing of large-scale proteomics and peptidomics DIA-SWATH-MS runs. Because its implementation is based on the nf-core template for reproducible bioinformatics workflows, DIAproteomics provides a web-based browser interface that can be customized. (See the Supporting Information for a short tutorial on how to get started and which software prerequisites are required.) It allows one to get an overview of and to adjust the grouping of the available parameters into several categories and to document their functions in short to longer expandable descriptions. After launching, the execution of the pipeline runs on the local system, and the progress can be monitored in the command line or using the Nextflow Tower functionality. Several sample sheets that annotate the batch identifiers and conditions for each sample as defined by the experimental design serve as input to the pipeline (Figure 2).

Whenever possible, each step of the pipeline is executed and individually submitted for processing by the computing infrastructure. In this way, the processing of multiple large batches of files can be efficiently parallelized. On the contrary, if steps allow the combination of multiple files, then the workflow groups the files according to the experimental design and coprocesses them. This occurs, for instance, when merging and aligning multiple spectral libraries or when carrying out a global FDR estimation on merged DIA search results.

Depending on how the parameters are set within the major categories, the input and output files may vary. Most importantly, it can be defined whether one or multiple existing

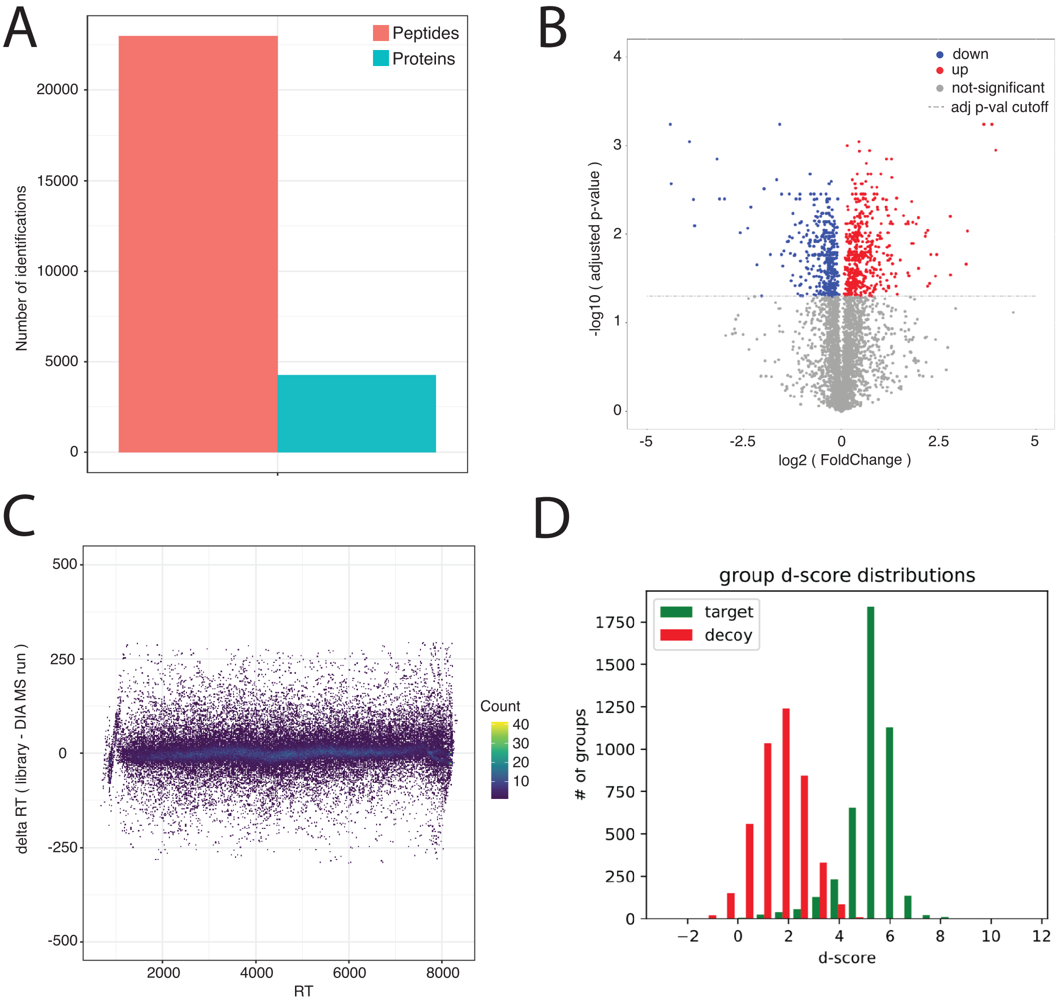


Figure 3. Several diagnostic visualizations of the DIAproteomics output can be generated. The results of the reprocessing of the publicly available data set PRIDE: PXD003179 are shown here as an example. (A) Peptide and protein identification counts. (B) Volcano plot of differentially regulated proteins (red up, blue down) proteins across conditions. (C) Deviation of the spectral library and the MS run in retention time (RT) over the entire RT range. (D) Target and decoy *d*-score distribution as computed by PyProphet to assess the false discovery rate (FDR).

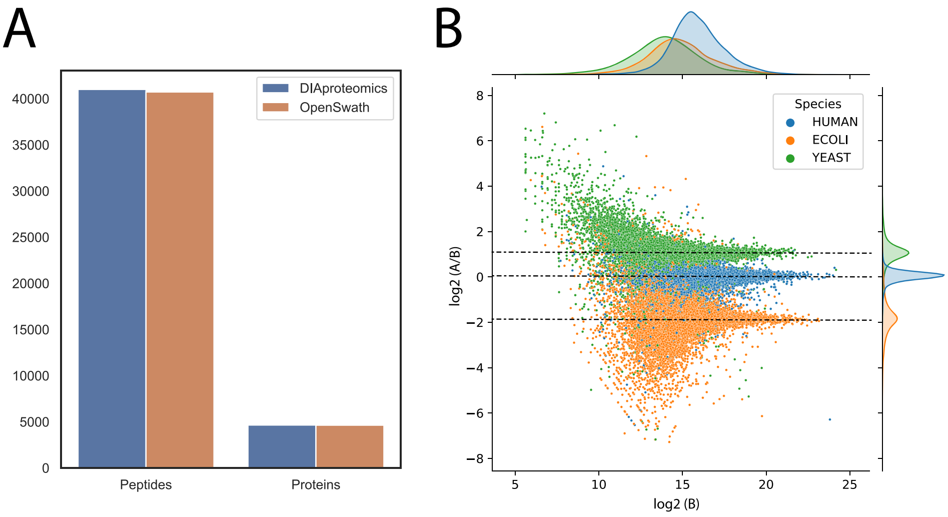


Figure 4. Comparing the performance of DIAproteomics with the OpenSWATH benchmark results on the ABSciex TripleToF 6600, 64 variable window setup (Navarro et al. (2017), PRIDE: PXD002952[32]). (A) Number of peptide and protein identifications retrieved using the DIAproteomics workflow compared with the reported numbers in the benchmark. (B) Quantitative protein log2 fold changes of human (ratio: 1:1), *E. coli* (ratio: 1:4), and yeast (ratio: 2:1) organism mixtures.

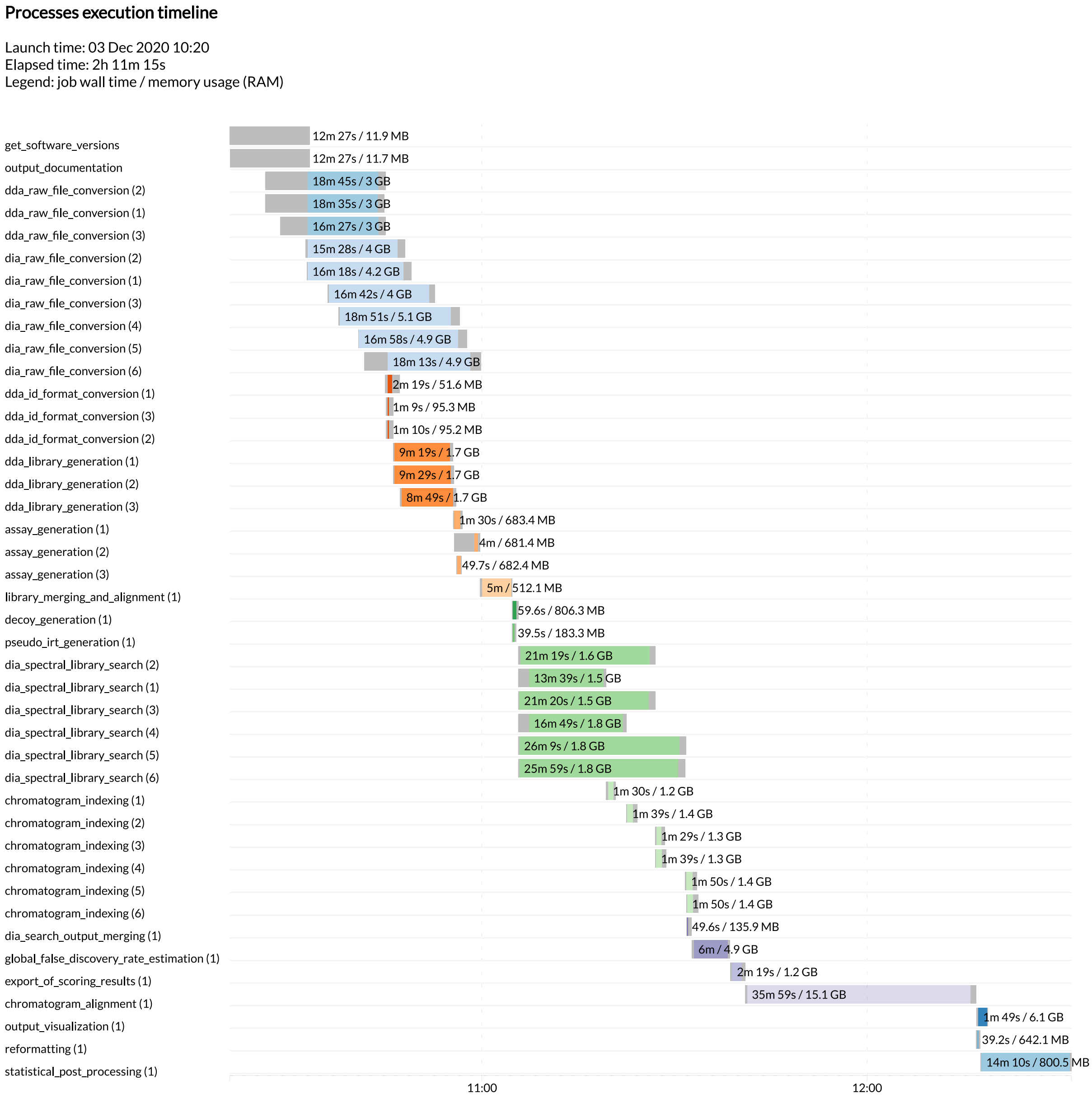


Figure 5. Detailed overview of the runtimes and memory usage of all integrated steps of the DIAproteomics pipeline when processing the PRIDE data set PXD003179 on the Amazon Web Services cloud infrastructure. The gray portions of the plot indicate non-process-specific task scheduling wait times (e.g., cluster job queue, download of the data set, container image or termination cleanup, and file unstaging). Each process is labeled with a different color, and runs resulting form the same process are colored in the same way.

spectral libraries should be used or whether the spectral libraries should be generated from matching DDA raw files and peptide identification results. Yet many more settings for each of the parameter categories are available and can be tailored to specific problem settings and MS instrument requirements.

Statistical Postprocessing and Diagnostic Output Visualization

The output of the DIAproteomics pipeline is by default a set of tables as well as illustrations summarizing the peptide or protein amount and quantity and scoring results. Moreover, important intermediate results such as the generated libraries, the output of the DIA spectral library search, and the XICs are reported. Most importantly, the detailed target-decoy score distribution results and their visualizations, as exported from PyProphet, are deposited in the output directory. The MSstats postprocessing software is run on the determined peptide or protein quantities. This results in the statistically sound estimation of pairwise fold changes and their significance across the conditions defined in the experimental design that are as well visualized in comparative plots such as a Volcano visualization. In addition, more diagnostic plots can be generated listing the number of peptides and proteins identified, their properties, such as the charge distribution, and the RT deviation between the spectral library and the DIA measurement to assess the performance of the iRT alignment (Figure 3). Finally, if specified, a heatmap of peptide quantities and missing values across all DIA MS runs is exported.

Reproduction of OpenSWATH benchmark performance

The performance of the DIAproteomics was assessed in a concise benchmark by reanalyzing the raw data and comparing with the results of a multicenter benchmark study.[23] As a result, we were able to reproduce approximately the same number of peptide (DIAproteomics: 41 006, OpenSWATH: 40 728) and protein (DIAproteomics: 4651, OpenSWATH: 4636) identifications as that retrieved in the original publication in 2017. In addition, the protein log-fold changes of the human (1:1), *E. coli* (1:4), and yeast (2:1) mixtures at defined ratios could be reproduced (Figure 4). The decrease in the accuracy and the increase in the variance of protein fold changes from the highest to the lowest intensity tertile reflect the uncertainty of quantification and background subtraction at this range and were also observed in the original benchmark study.

Runtime Considerations

The runtime of the DIAproteomics workflow depends on its parametrization. For example, if spectral library generation from DDA data is chosen, additional analysis steps are carried out in contrast to running the workflow with a pre-existing input spectral library. Moreover, the number of samples and batches that are analyzed in one submission vastly influence the required runtime. In addition, the employed computational infrastructure and the number of available cores determine the amount of parallel threads that can be simultaneously run. We assessed the computational runtime and required resources by making use of the AWS cloud infrastructure and the German Network for Bioinformatics Infrastructure (de.NBI) cloud HPC node with 28 cores and 64 GB. The analysis of six DIA-SWATH-MS runs applying a library and pseudo iRTs generated from three DDA MS runs was carried out in ∼2 h and 10 min using AWS (Figure 5), in contrast with 33.4 CPU hours of sequential analysis time.

■ CONCLUSIONS

In this work, we present DIAproteomics, a flexible computational workflow to automatically process large-scale DIA-SWATH-MS-based proteomics and peptidomics studies on diverse computational systems. It combines all steps including the optional generation of spectral libraries from DDA data and the essential DIA library search, FDR estimation, and chromatogram alignment. The implementation and sharing of the workflow as part of the nf-core initiative for reproducible bioinformatics research provides an easy-to-use user interface as well as reproducible, well-tested analysis. Finally, the results retrieved using the DIAproteomics workflow were compared to a thorough benchmark study that assessed the OpenSWATH workflow in 2017, and an equally good performance was reproduced. The DIAproteomics pipeline is provided for free to the science community with the purpose of enabling easier access to as well as automated and reproducible analysis of DIA-SWATH-MS-based proteome research.

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.1c00123.

> Description S1. Pipeline installation tutorial. Table S1. Details of all steps in the Nextflow workflow implementation. Figure S1. Command line execution report of the workflow. Figure S2. Pairwise RT alignment option for merging multiple spectral libraries (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Leon Bichmann − *Department of Computer Science, Applied Bioinformatics, University of Tübingen, Tübingen 72076, Germany; Institute for Cell Biology, Department of Immunology, University of Tübingen, Tübingen 72076, Germany;* orcid.org/0000-0001-7135-0073; Email: leon.bichmann@uni-tuebingen.de

Authors

Shubham Gupta − *Donnelly Center for Biomolecular Research, University of Toronto, Toronto, Ontario ON M5S 3E1, Canada*

George Rosenberger − *Department of Systems Biology, Columbia University, New York 10032, United States*

Leon Kuchenbecker − *Department of Computer Science, Applied Bioinformatics, University of Tübingen, Tübingen 72076, Germany*

Timo Sachsenberg − *Department of Computer Science, Applied Bioinformatics, University of Tübingen, Tübingen 72076, Germany;* orcid.org/0000-0002-2833-6070

Phil Ewels − *Science for Life Laboratory (SciLifeLab), Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden*

Oliver Alka − *Department of Computer Science, Applied Bioinformatics, University of Tübingen, Tübingen 72076, Germany*

Julianus Pfeuffer − *Department of Computer Science, Applied Bioinformatics, University of Tübingen, Tübingen 72076, Germany; Institute for Informatics, Freie Universität Berlin, Berlin 14195, Germany; Zuse Institute Berlin, Berlin 14195, Germany;* orcid.org/0000-0001-8948-9209

Oliver Kohlbacher − *Department of Computer Science, Applied Bioinformatics, University of Tübingen, Tübingen 72076, Germany; Institute for Biological and Medical Informatics, University of Tübingen, Tübingen 72076, Germany; Institute for Translational Bioinformatics, University Hospital Tübingen, Tübingen 72076, Germany;* orcid.org/0000-0003-1739-4598

Hannes Röst − *Donnelly Center for Biomolecular Research, University of Toronto, Toronto, Ontario ON M5S 3E1, Canada;* orcid.org/0000-0003-3500-8152

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.1c00123

Author Contributions

L.B. constructed the pipeline, carried out the data analysis, and wrote the paper. S.G. created DIAlignR and G.R. created EasyPQP, two software tools that are essential components of the pipeline, and both supported their integration and debugging into the workflow. L.K., T.S., P.E., J.P., and O.A. assisted in reviewing the source code and suggested architecture and parameter changes. O.K. and H.R. were involved in the study design. All authors discussed and commented on the manuscript.

Notes

The authors declare no competing financial interest.
The workflow is freely available under an open-source license as Nextflow implementation in the nf-core bioinformatics workflow repository: https://www.openms.de/diaproteomics/.

Detailed documentation regarding the parameters and pipeline output can be found at: https://nf-co.re/diaproteomics

■ ACKNOWLEDGMENTS

This work was supported by the German Ministry for Research and Education (BMBF) as part of the German Network for Bioinformatics infrastructure (FKZ: 31A535A) and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2180-390900677. In addition, the work was initiated through a travel stipend by the Boehringer Ingelheim Fonds for basic research in medicine and supported by the Chan Zuckerberg Initiative program "Essential Open-Source Software for Science (EOSS)". We thank all nf-core and OpenMS team members, in particular, Phil Ewels and Gisela Gabernet, for supporting the development and debugging of the pipeline as well as for the provision of the template. In addition, we thank the Quantitative Biology Center in Tübingen (QBiC) for hosting a productive software developer meeting.

■ ABBREVIATIONS

LC-MS/MS, liquid chromatography−mass spectrometry; MS, mass spectrometry; DIA, data-independent acquisition; DDA, data-dependent acquisition; FDR, false discovery rate; XIC, extracted ion chromatogram; HPC, high-performance computing

■ REFERENCES

(1) Gillet, L. C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. 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* **2012**, *11* (6), No. O111.016717.

(2) Hu, A.; Noble, W. S.; Wolf-Yadlin, A. Technical Advances in Proteomics: New Developments in Data-Independent Acquisition. *F1000Research* **2016**, *5*, No. 419.

(3) Doerr, A. DIA Mass Spectrometry. *Nat. Methods* **2015**, *12* (1), 35−35.

(4) Bouchal, P.; Schubert, O. T.; Faktor, J.; Capkova, L.; Imrichova, H.; Zoufalova, K.; Paralova, V.; Hrstka, R.; Liu, Y.; Ebhardt, H. A.; Budinska, E.; Nenutil, R.; Aebersold, R. Breast Cancer Classification Based on Proteotypes Obtained by SWATH Mass Spectrometry. *Cell Rep.* **2019**, *28* (3), 832−843. e7.

(5) Poulos, R. C.; Hains, P. G.; Shah, R.; Lucas, N.; Xavier, D.; Manda, S. S.; Anees, A.; Koh, J. M. S.; Mahboob, S.; Wittman, M.; Williams, S. G.; Sykes, E. K.; Hecker, M.; Dausmann, M.; Wouters, M. A.; Ashman, K.; Yang, J.; Wild, P. J.; deFazio, A.; Balleine, R. L.; Tully, B.; Aebersold, R.; Speed, T. P.; Liu, Y.; Reddel, R. R.; Robinson, P. J.; Zhong, Q. Strategies to Enable Large-Scale Proteomics for Reproducible Research. *Nat. Commun.* **2020**, *11* (1), 3793.

(6) Krasny, L.; Huang, P. H. Data-Independent Acquisition Mass Spectrometry (DIA-MS) for Proteomic Applications in Oncology. *Mol. Omics* **2021**, *17* (1), 29−42.

(7) Caron, E.; Espona, L.; Kowalewski, D. J.; Schuster, H.; Ternette, N.; Alpízar, A.; Schittenhelm, R. B.; Ramarathinam, S. H.; Lindestam Arlehamn, C. S.; Chiek Koh, C.; Gillet, L. C.; Rabsteyn, A.; Navarro, P.; Kim, S.; Lam, H.; Sturm, T.; Marcilla, M.; Sette, A.; Campbell, D. S.; Deutsch, E. W.; Moritz, R. L.; Purcell, A. W.; Rammensee, H.-G.; Stevanovic, S.; Aebersold, R. An Open-Source Computational and Data Resource to Analyze Digital Maps of Immunopeptidomes. *eLife* **2015**, *4*, No. 07661.

(8) Ritz, D.; Kinzi, J.; Neri, D.; Fugmann, T. Data-Independent Acquisition of HLA Class I Peptidomes on the Q Exactive Mass Spectrometer Platform. *Proteomics* **2017**, *17* (19), No. 1700177.

(9) Saidi, M.; Kamali, S.; Beaudry, F. Neuropeptidomics: Comparison of Parallel Reaction Monitoring and Data-Independent Acquisition for the Analysis of Neuropeptides Using High-Resolution Mass Spectrometry. *Biomed. Chromatogr.* **2019**, *33* (7), e4523.

(10) Lin, L.; Zheng, J.; Zheng, F.; Cai, Z.; Yu, Q. Advancing Serum Peptidomic Profiling by Data-Independent Acquisition for Clear-Cell Renal Cell Carcinoma Detection and Biomarker Discovery. *J. Proteomics* **2020**, *215*, 103671.

(11) Pak, H.; Michaux, J.; Huber, F.; Chong, C.; Stevenson, B. J.; Müller, M.; Coukos, G.; Bassani-Sternberg, M. Sensitive Immunopeptidomics by Leveraging Available Large-Scale Multi-HLA Spectral Libraries, Data-Independent Acquisition and MS/MS Prediction. *Molecular & Cellular Proteomics* **2021**, *20*, No. 100080.

(12) Meyer, J. G.; Schilling, B. Clinical Applications of Quantitative Proteomics Using Targeted and Untargeted Data-Independent Acquisition Techniques. *Expert Rev. Proteomics* **2017**, *14* (5), 419−429.

(13) Piazza, I.; Beaton, N.; Bruderer, R.; Knobloch, T.; Barbisan, C.; Chandat, L.; Sudau, A.; Siepe, I.; Rinner, O.; de Souza, N.; Picotti, P.; Reiter, L. A Machine Learning-Based Chemoproteomic Approach to Identify Drug Targets and Binding Sites in Complex Proteomes. *Nat. Commun.* **2020**, *11* (1), 4200.

(14) Meyer, J. G.; Mukkamalla, S.; Steen, H.; Nesvizhskii, A. I.; Gibson, B. W.; Schilling, B. PIQED: Automated Identification and Quantification of Protein Modifications from DIA-MS Data. *Nat. Methods* **2017**, *14* (7), 646−647.

(15) Bekker-Jensen, D. B.; Bernhardt, O. M.; Hogrebe, A.; Martinez-Val, A.; Verbeke, L.; Gandhi, T.; Kelstrup, C. D.; Reiter, L.; Olsen, J. V. Rapid and Site-Specific Deep Phosphoproteome Profiling by Data-Independent Acquisition without the Need for Spectral Libraries. *Nat. Commun.* **2020**, *11* (1), 787.

(16) Canterbury, J. D.; Merrihew, G. E.; Goodlett, D. R.; MacCoss, M. J.; Shaffer, S. A. Comparison of Data Acquisition Strategies on Quadrupole Ion Trap Instrumentation for Shotgun Proteomics. *J. Am. Soc. Mass Spectrom.* **2014**, *25* (12), 2048−2059.

(17) Tsou, C.-C.; Avtonomov, D.; Larsen, B.; Tucholska, M.; Choi, H.; Gingras, A.-C.; Nesvizhskii, A. I. DIA-Umpire: Comprehensive Computational Framework for Data-Independent Acquisition Proteomics. *Nat. Methods* **2015**, *12* (3), 258−264.

(18) Schubert, O. T.; Gillet, L. C.; Collins, B. C.; Navarro, P.; Rosenberger, G.; Wolski, W. E.; Lam, H.; Amodei, D.; Mallick, P.; MacLean, B.; Aebersold, R. Building High-Quality Assay Libraries for Targeted Analysis of SWATH MS Data. *Nat. Protoc.* **2015**, *10* (3), 426−441.

(19) Vizcaíno, J. A.; Csordas, A.; del-Toro, N.; Dianes, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; Xu, Q.-W.; Wang, R.; Hermjakob, H. 2016 Update of the PRIDE Database and Its Related Tools. *Nucleic Acids Res.* **2016**, *44* (D1), D447−456.

(20) Desiere, F.; Deutsch, E. W.; King, N. L.; Nesvizhskii, A. I.; Mallick, P.; Eng, J.; Chen, S.; Eddes, J.; Loevenich, S. N.; Aebersold, R. The PeptideAtlas Project. *Nucleic Acids Res.* **2006**, *34* (suppl_1), D655−D658.

(21) Rosenberger, G.; Koh, C. C.; Guo, T.; Röst, H. L.; Kouvonen, P.; Collins, B. C.; Heusel, M.; Liu, Y.; Caron, E.; Vichalkovski, A.; Faini, M.; Schubert, O. T.; Faridi, P.; Ebhardt, H. A.; Matondo, M.; Lam, H.; Bader, S. L.; Campbell, D. S.; Deutsch, E. W.; Moritz, R. L.; Tate, S.; Aebersold, R. A Repository of Assays to Quantify 10,000 Human Proteins by SWATH-MS. *Sci. Data* **2014**, *1* (1), 140031.

(22) Shao, W.; Pedrioli, P. G. A.; Wolski, W.; Scurtescu, C.; Schmid, E.; Vizcaíno, J. A.; Courcelles, M.; Schuster, H.; Kowalewski, D.; Marino, F.; Arlehamn, C. S. L.; Vaughan, K.; Peters, B.; Sette, A.; Ottenhoff, T. H. M.; Meijgaarden, K. E.; Nieuwenhuizen, N.; Kaufmann, S. H. E.; Schlapbach, R.; Castle, J. C.; Nesvizhskii, A. I.; Nielsen, M.; Deutsch, E. W.; Campbell, D. S.; Moritz, R. L.; Zubarev, R. A.; Ytterberg, A. J.; Purcell, A. W.; Marcilla, M.; Paradela, A.; Wang, Q.; Costello, C. E.; Ternette, N.; van Veelen, P. A.; van Els, C. A. C. M.; Heck, A. J. R.; de Souza, G. A.; Sollid, L. M.; Admon, A.; Stevanovic, S.; Rammensee, H.-G.; Thibault, P.; Perreault, C.; Bassani-Sternberg, M.; Aebersold, R.; Caron, E. The SysteMHC Atlas Project. *Nucleic Acids Res.* **2018**, *46* (D1), D1237−D1247.

(23) Noble, W. S. Mass Spectrometrists Should Search Only for Peptides They Care About. *Nat. Methods* **2015**, *12* (7), 605−608.

(24) Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A.; Reimer, U.; Ehrlich, H.-C.; Aiche, S.; Kuster, B.; Wilhelm, M. Prosit: Proteome-Wide Prediction of Peptide Tandem Mass Spectra by Deep Learning. *Nat. Methods* **2019**, *16* (6), 509−518.

(25) Gabriels, R.; Martens, L.; Degroeve, S. Updated MS^{2}PIP Web Server Delivers Fast and Accurate MS^2 Peak Intensity Prediction for Multiple Fragmentation Methods, Instruments and Labeling Techniques. *Nucleic Acids Res.* **2019**, *47* (W1), W295−W299.

(26) Tiwary, S.; Levy, R.; Gutenbrunner, P.; Salinas Soto, F.; Palaniappan, K. K.; Deming, L.; Berndl, M.; Brant, A.; Cimermancic, P.; Cox, J. High-Quality MS/MS Spectrum Prediction for Data-Dependent and Data-Independent Acquisition Data Analysis. *Nat. Methods* **2019**, *16* (6), 519−525.

(27) Van Puyvelde, B.; Willems, S.; Gabriels, R.; Daled, S.; De Clerck, L.; Van de Casteele, S.; Staes, A.; Impens, F.; Deforce, D.; Martens, L.; Degroeve, S.; Dhaenens, M. Front Cover: Removing the Hidden Data Dependency of DIA with Predicted Spectral Libraries. *Proteomics* **2020**, *20* (3−4), 2070021.

(28) Ting, Y. S.; Egertson, J. D.; Bollinger, J. G.; Searle, B. C.; Payne, S. H.; Noble, W. S.; MacCoss, M. J. PECAN: Library-Free Peptide Detection for Data-Independent Acquisition Tandem Mass Spectrometry Data. *Nat. Methods* **2017**, *14* (9), 903−908.

(29) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; Eng, J. K.; Martin, D. B.; Nesvizhskii, A.; Aebersold, R. A Guided Tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10* (6), 1150−1159.

(30) Röst, H. L.; Rosenberger, G.; Navarro, P.; Gillet, L.; Miladinović, S. M.; Schubert, O. T.; Wolski, W.; Collins, B. C.; Malmström, J.; Malmström, L.; Aebersold, R. OpenSWATH Enables Automated, Targeted Analysis of Data-Independent Acquisition MS Data. *Nat. Biotechnol.* **2014**, *32* (3), 219−223.

(31) Boekel, J.; Chilton, J. M.; Cooke, I. R.; Horvatovich, P. L.; Jagtap, P. D.; Käll, L.; Lehtiö, J.; Lukasse, P.; Moerland, P. D.; Griffin, T. J. Multi-Omic Data Analysis Using Galaxy. *Nat. Biotechnol.* **2015**, *33* (2), 137−139.

(32) Navarro, P.; Kuharev, J.; Gillet, L. C.; Bernhardt, O. M.; MacLean, B.; Röst, H. L.; Tate, S. A.; Tsou, C.-C.; Reiter, L.; Distler, U.; Rosenberger, G.; Perez-Riverol, Y.; Nesvizhskii, A. I.; Aebersold, R.; Tenzer, S. A Multi-Center Study Benchmarks Software Tools for Label-Free Proteome Quantification. *Nat. Biotechnol.* **2016**, *34* (11), 1130−1136.

(33) Searle, B. C.; Pino, L. K.; Egertson, J. D.; Ting, Y. S.; Lawrence, R. T.; MacLean, B. X.; Villén, J.; MacCoss, M. J. Chromatogram Libraries Improve Peptide Detection and Quantification by Data Independent Acquisition Mass Spectrometry. *Nat. Commun.* **2018**, *9* (1), 5128.

(34) Chen, C.-T.; Ko, C.-L.; Choong, W.-K.; Wang, J.-H.; Hsu, W.-L.; Sung, T.-Y. WinProphet: A User-Friendly Pipeline Management System for Proteomics Data Analysis Based on Trans-Proteomic Pipeline. *Anal. Chem.* **2019**, *91* (15), 9403−9406.

(35) Pino, L. K.; Searle, B. C.; Bollinger, J. G.; Nunn, B.; MacLean, B.; MacCoss, M. J. The Skyline Ecosystem: Informatics for Quantitative Mass Spectrometry Proteomics. *Mass Spectrom. Rev.* **2020**, *39* (3), 229−244.

(36) Wang, D.; Gan, G.; Chen, X.; Zhong, C.-Q. QuantPipe: A User-Friendly Pipeline Software Tool for DIA Data Analysis Based on the OpenSWATH-PyProphet-TRIC Workflow. *J. Proteome Res.* **2021**, *20* (1), 1096−1102.

(37) Li, C.; Gao, M.; Yang, W.; Zhong, C.; Yu, R. Diamond: A Multi-Modal DIA Mass Spectrometry Data Processing Pipeline. *Bioinformatics* **2021**, *37*, No. 265.

(38) Alka, O.; Sachsenberg, T.; Bichmann, L.; Pfeuffer, J.; Weisser, H.; Wein, S.; Netz, E.; Rurik, M.; Kohlbacher, O.; Rost, H. OpenMS for Open Source Analysis of Mass Spectrometric Data. *PeerJ Preprints* **2019**, e27766v1 DOI: 10.7287/peerj.preprints.27766v1.

(39) Röst, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmström, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O. OpenMS: A Flexible Open-Source Software Platform for Mass Spectrometry Data Analysis. *Nat. Methods* **2016**, *13* (9), 741.

(40) Rosenberger, G.; Bludau, I.; Schmitt, U.; Heusel, M.; Hunter, C.; Liu, Y.; MacCoss, M. J.; MacLean, B. X.; Nesvizhskii, A. I.; Pedrioli, P. G. A.; Reiter, L.; Röst, H. L.; Tate, S.; Ting, Y. S.; Collins, B. C.; Aebersold, R. Statistical Control of Peptide and Protein Error Rates in Large-Scale Targeted DIA Analyses. *Nat. Methods* **2017**, *14* (9), 921−927.

(41) Gupta, S.; Ahadi, S.; Zhou, W.; Röst, H. DIAlignR Provides Precise Retention Time Alignment Across Distant Runs in DIA and Targeted Proteomics. *Molecular & Cellular Proteomics* **2019**, *18* (4), 806−817.

(42) Choi, M.; Chang, C.-Y.; Clough, T.; Broudy, D.; Killeen, T.; MacLean, B.; Vitek, O. MSstats: An R Package for Statistical Analysis of Quantitative Mass Spectrometry-Based Proteomic Experiments. *Bioinformatics* **2014**, *30* (17), 2524−2526.

(43) Di Tommaso, P.; Chatzou, M.; Floden, E. W.; Barja, P. P.; Palumbo, E.; Notredame, C. Nextflow Enables Reproducible Computational Workflows. *Nat. Biotechnol.* **2017**, *35*, 316−319.

(44) Ewels, P. A.; Peltzer, A.; Fillinger, S.; Patel, H.; Alneberg, J.; Wilm, A.; Garcia, M. U.; Di Tommaso, P.; Nahnsen, S. The Nf-Core Framework for Community-Curated Bioinformatics Pipelines. *Nat. Biotechnol.* **2020**, *38* (3), 276−278.

(45) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Römpp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.; Souda, P.; Hermjakob, H.; Binz, P.-A.; Deutsch, E. W. MzML—a Community Standard for Mass Spectrometry Data. *Mol. Cell Proteomics* **2011**, *10* (1), No. R110.000133.

(46) Yu, F.; Haynes, S. E.; Teo, G. C.; Avtonomov, D. M.; Polasky, D. A.; Nesvizhskii, A. I. Fast Quantitative Analysis of TimsTOF PASEF Data with MSFragger and IonQuant. *Molecular & Cellular Proteomics* **2020**, *19*, 1575.

(47) Röst, H. L.; Liu, Y.; D'Agostino, G.; Zanella, M.; Navarro, P.; Rosenberger, G.; Collins, B. C.; Gillet, L.; Testa, G.; Malmström, L.; Aebersold, R. TRIC: An Automated Alignment Strategy for Reproducible Protein Quantification in Targeted Proteomics. *Nat. Methods* **2016**, *13* (9), 777−783.

(48) Gupta, S.; Röst, H. Automated Workflow For Peptide-Level Quantitation From DIA/ SWATH-MS Data. *bioRxiv* **2020**, DOI: 10.1101/2020.01.21.914788.

(49) Taylor, S. L.; Leiserowitz, G. S.; Kim, K. Accounting for Undetected Compounds in Statistical Analyses of Mass Spectrometry 'omic Studies. *Stat. Appl. Genet. Mol. Biol.* **2013**, *12* (6), 703−722.

(50) Hagberg, A.; Schult, D.; Swart, P. Exploring Network Structure, Dynamics, and Function using NetworkX, In *Proceedings of the 7th Python in Science Conference (SciPy 2008)*; Varoquaux, G., Vaught, T., Millman, J., Eds., 2008; pp 11−15. http://conference.scipy.org/proceedings/SciPy2008/paper_2/ (accessed 2020-11-10).

(51) Gotti, C.; Roux-Dalvai, F.; Joly-Beauparlant, C.; Leclercq, M.; Mangnier, L.; Droit, A. Extensive and Accurate Benchmarking of DIA Acquisition Methods and Software Tools Using a Complex Proteomic Standard. *bioRxiv* **2020**, DOI: 10.1101/2020.11.03.365585.

(52) Tsou, C.-C.; Tsai, C.-F.; Teo, G.; Chen, Y.-J.; Nesvizhskii, A. I. Untargeted, Spectral Library-Free Analysis of Data Independent Acquisition Proteomics Data Generated Using Orbitrap Mass Spectrometers. *Proteomics* **2016**, *16* (15−16), 2257−2271.