analytical chemistry

pubs.acs.org/ac Article

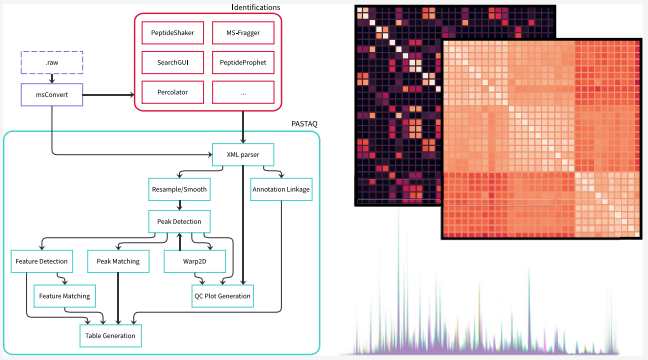
Pipelines and Systems for Threshold-Avoiding Quantification of LC–MS/MS Data

Alejandro Sánchez Brotons, Jonatan O. Eriksson, Marcel Kwiatkowski, Justina C. Wolters, Ido P. Kema, Andrei Barcaru, Folkert Kuipers, Stephan J. L. Bakker, Rainer Bischoff, Frank Suits, and Péter Horvatovich*

Cite This: *Anal. Chem.* 2021, 93, 11215–11224 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The accurate processing of complex liquid chromatography coupled to tandem mass spectrometry (LC–MS/MS) data from biological samples is a major challenge for metabolomics, proteomics, and related approaches. Here, we present the pipelines and systems for threshold-avoiding quantification (PASTAQ) LC–MS/MS preprocessing toolset, which allows highly accurate quantification of data-dependent acquisition LC–MS/MS datasets. PASTAQ performs compound quantification using single-stage (MS1) data and implements novel algorithms for high-performance and accurate quantification, retention time alignment, feature detection, and linking annotations from multiple identification engines. PASTAQ offers straightforward parameterization and automatic generation of quality control plots for data and preprocessing assessment. This design results in smaller variance when analyzing replicates of proteomes mixed with known ratios and allows the detection of peptides over a larger dynamic concentration range compared to widely used proteomics preprocessing tools. The performance of the pipeline is also demonstrated in a biological human serum dataset for the identification of gender-related proteins.



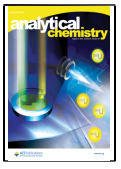
INTRODUCTION

Liquid chromatography coupled to tandem mass spectrometry (LC–MS/MS) is a powerful analytical technique for the quantitative profiling of proteins, peptides, and metabolites in complex biological samples. In the last decade, advances in instrumentation such as faster acquisition speed, improved sensitivity, and increased dynamic range have made LC–MS/MS the method of choice for routine analyses in clinical and life science applications,[1,2] as well as a widely used tool for biomarker discovery, quantitative protein and metabolite profiling, and drug screening.[3−6]

LC–MS/MS data are complex and require the use of sophisticated data preprocessing pipelines that allow extraction of quantitative and identification information of compounds. For this purpose, numerous tools and pipelines exist, both commercial and freely available ones, such as MaxQuant,[7] OpenMS,[8] and XCMS.[9] Some of these tools are used mainly for proteomics, while others are more flexible in their usage, with support for metabolomics or lipidomics analyses. In general, tools developed for label-free data-dependent acquisition (DDA) proteomics applications rely on the quantification of isotope clusters (features), which have been identified using MS/MS spectra to match the quantitative information of the same peptide across multiple samples. One limitation of this approach is that for DDA analyses, only a fraction of the most abundant compounds is selected for fragmentation. Furthermore, identifications might not be possible for all compounds due to the inherent stochasticity of the selection of precursor ions for fragmentation in the MS/MS sampling process, and thus, compounds present in the sample in a low concentration (or compounds that ionize poorly) will be less likely to be selected for fragmentation.

Chimeric spectra can also occur with typical DDA fragmentation windows of 0.4–2.0 m/z,[10] making the correct identification of peptides challenging. Single-stage (MS1) spectra in DDA LC–MS/MS data reflect a stable quantitative profile of compounds that can be detected by the mass spectrometer and are not subjected to a stochastic selection procedure such as DDA precursor selection for compound fragmentation. We argue that MS1 data offer the most stable information to perform comprehensive and accurate processing

Received: May 4, 2021
Accepted: July 26, 2021
Published: August 6, 2021



of DDA LC−MS/MS datasets. This approach relies only on the m/z and retention time information of the compounds' features and depends on accurate correction of any shifts existing in these separation domains.

Here, we present the pipelines and systems for threshold-avoiding quantification (PASTAQ), a set of tools and algorithms that can be used to preprocess and quantify compounds present in LC−MS/MS data, regardless of the existence of MS/MS spectra. PASTAQ is built on the algorithmic and workflow design of the threshold-avoiding proteomics pipeline,[11] which is focused on accurately processing single-stage LC−MS data at the isotope level. PASTAQ performs the quantification of LC−MS/MS compounds exclusively with MS1 information, and MS/MS based identification is used for annotation of matched MS1 features. PASTAQ includes an improved version of the Warp2D retention time alignment algorithm and allows assessment of alignment accuracy by evaluating the similarity of the chromatograms after retention time alignment. PASTAQ offers a comprehensive set of tools, a prebuilt DDA pipeline, and an easy to use graphical user interface (GUI) that enables the quantification and exploration of LC−MS/MS data tailored for metabolomics and proteomics analyses. Furthermore, PASTAQ allows the use of any identification engine that supports exporting identifications in the mzIdentML format,[12] including postprocessing tools such as PeptideShaker,[13] PeptideProphet,[14] ProteinProphet,[15] or Percolator[16] to adjust the false discovery rate (FDR).

To evaluate both the quantitative and computational performance of this pipeline, we used two different datasets. The first one consists of a proteome mixture of HeLa, yeast, and *Escherichia coli* (*E. coli*) at three different ratios: (A) 10:5:1, (B) 5:10:1, and (C) 1:5:10. The second dataset is composed of a HeLa matrix with an artificial concatemer protein[17] spiked at increasing ratios to cover 3 orders of magnitude of compound concentration and concomitant measured ion intensities. The first and second datasets are hereinafter referred to as the HYE dataset and the QconCAT dataset, respectively. Both datasets were acquired with a nano-LC system coupled to an Orbitrap QExactive Plus mass spectrometer. Additionally, we demonstrate the use of PASTAQ with publicly available serum LC−MS/MS datasets[18] to show its performance with a biologically relevant dataset.

METHODS

Preparation of Complex Proteome Samples. *HYE Dataset.* To generate complex proteome samples with known composition, a tryptic HeLa protein digest (Pierce HeLa Protein Digest Standard, Thermo Fisher Scientific, Dreieich, Germany), a tryptic yeast protein digest (Mass Spec-Compatible Yeast Extract, Promega, Walldorf, Germany), and an *E. coli* (*Escherichia coli*) tryptic protein digest (Waters, Manchester, UK) were used. For differential proteomics, three proteome mixtures (A, B, and C) were prepared, composed of the HeLa proteome, yeast proteome, and *E. coli* proteome. (A) 20 μg of HeLa digest (dissolved in 0.1% FA) was combined with 10 μg of yeast digest (dissolved in 0.1% FA) and 2 μg of *E. coli* digest (dissolved in 0.1% FA). (B) 10 μg of HeLa digest (dissolved in 0.1% FA) was combined with 20 μg of yeast digest (dissolved in 0.1% FA) and 2 μg of *E. coli* digest (dissolved in 0.1% FA). (C) 2 μg of HeLa digest (dissolved in 0.1% FA) was combined with 10 μg of yeast digest (dissolved in 0.1% FA) and 20 μg of *E. coli* digest (dissolved in 0.1% FA). The sample mixtures A, B, and C had a final concentration of 2 $\mu g{\cdot}\mu L^{-1}$ (dissolved in 0.1% FA).

QconCAT Dataset. The complex proteome samples with the spiked isotopically labelled standards covering the 3 orders of magnitude in the dynamic range were prepared with MS-compatible human protein extract (V694A, whole-cell protein extract prepared from human H562 cells, Promega), which was digested with an in-gel digestion protocol as described previously.[17] Digestates were resuspended in 0.1% FA at a final concentration of 1 $\mu g{\cdot}\mu L^{-1}$ and mixed 1:1 with isotopically labeled standard peptides. The isotopically labeled peptides were digested from the three concatamers targeting a specific set of mitochondrial proteins as described previously.[17] The standard peptides were added at increasing $\log_{10}(ng/\mu g)$ amounts from −3.0 up to 1.0. In $\log_{10}(ng\backslash\mu g)$ ng reflect the amount of spiked-in concatemer and μg the amount of the background cell line proteome.

LC−MS/MS Analysis. For LC−MS/MS analysis, 1 μL was injected on a nano-ultrapressure LC system (Dionex UltiMate 3000 RSLCnano pro flow, Thermo Scientific, Bremen, Germany) coupled via nano-electrospray ionization (ESI) to a quadrupole orbitrap mass spectrometer equipped with a nano-electrospray ion source (Orbitrap Q Exactive Plus, Thermo Scientific, Bremen, Germany). Chromatographic separation of the peptides was performed using a nano-LC column (Acclaim PepMapC100 C18, 75 μm × 50 cm, 2 μm, 100 Å, Dionex, buffer A: 0.1% v\v formic acid (FA), dissolved in H_2O, buffer b: 0.1% v \v FA, dissolved in acetonitrile). The peptides were loaded onto a trap column (μPrecolumn cartridge, Acclaim PepMap100 C18, 5 μm, 100 Å, 300 μm × 5 mm, Dionex) with a flow-rate of 20 $\mu L{\cdot}min^{-1}$ and 3% buffer B. The peptides were separated on the nano-LC column with a flow-rate of 300 $\mu L{\cdot}min^{-1}$ using a linear gradient from 3 to 30% buffer B in 85 min, followed by 30−50% buffer B in 5 min. The mass spectrometer was operated in the positive ion mode and DDA mode using a top-10 method. The MS spectra were acquired at a resolution of 70,000 at m/z 200 over a scan range of 350−1300 m/z with an automatic gain control (AGC) target of 10^6 ions and a maximum injection time of 50 ms. Peptide fragmentation was performed with higher-energy collision dissociation using a normalized collision energy of 28. The intensity threshold for ion selection was set at 2.0×10^4 with a charge exclusion of ≤1 and ≥7. The MS/MS spectra were acquired at a resolution of 17,500 at 200 m/z, an AGC target of 10^5 ions, and a maximum injection time of 50 ms, and the quadrupole isolation window set to 1.6 m/z. The same instrument was used for the QconCAT dataset with the following alterations compared to the HYE dataset: a total of 3 μL was injected, and the mass spectrometer was operated with a top-15 method, with a charge exclusion of ≤1 and ≥5. The raw data are available via ProteomeXchange with identifier PXD024584.

Preprocessing and Parameterization Details. *Format Conversion.* The raw Orbitrap data files were processed directly in MaxQuant (Version 1.6.10.43), but PASTAQ requires files in the mzXML[19] or mzML[20] format and SearchGUI[21] (Version 3.3.17) only works properly with centroided mgf[22] files. File conversion was performed with ProteoWizard's msConvert[23] (Version 3.0.18342-01b48c0f0), with a binary encoding precision of 64 bits. The mzXML conversion was performed without additional processing. For the mgf conversion, centroiding was performed using the vendor's peak picking algorithm included in msConvert.

Peptide and Protein Identification. For protein identification, UP000005640 (*Homo sapiens*, updated on August 21, 2019), UP000002311 (Baker's yeast, updated on July 25, 2019),

https://doi.org/10.1021/acs.analchem.1c01892
Anal. Chem. 2021, 93, 11215−11224

and UP000000625 (*E. coli*, strain K12, updated on August 21, 2019) Swissprot[24] protein sequences were used. The canonical sequence of each proteome was downloaded in the FASTA format from the UniProt Knowledgebase (UniProtKB) website on October 31, 2019.

Prior to identification, the FASTA files of the three proteomes were concatenated. This FASTA file was used directly in MaxQuant[7] and MSFragger[25] (Version 3.1.1). When using SearchGUI,[21] this FASTA file was modified with the addition of reverse decoy sequences using the built-in tools.

Fixed modification of cysteine by carbamidomethylation (C) and the variable modification of methionine through oxidation to the sulfoxide (M) were used for all datasets. The QconCAT dataset was searched with the additional variable modifications of ^{13}C+6 on lysine (K) and arginine (R) due to the presence of stable isotopes in these amino acids for this artificial protein.

SearchGUI was used to perform peptide and protein identification with the aforementioned FASTA file and PSM settings. The precursor mass tolerance was set to 5 ppm and the fragment ion mass tolerance to 0.02 Da. The search engines selected for identification were (1) X!Tandem,[26] (2) MS Amanda,[27] and (3) MS-GF+.[28] The SearchGUI results were unified into a consensus mzIdentML[12] identification file for each sample using PeptideShakerCLI (Version 1.16.42)[13] and less than 1% false discovery rate (FDR) at PSM, peptide, and protein levels.

HYE Analysis. We performed an exhaustive analysis of quantified isotopes, features, peptides, and protein groups obtained with PASTAQ and MaxQuant. When PASTAQ quantification was used, we used SearchGUI/PeptideShaker for identifications.

For both PASTAQ and MaxQuant, we assigned the corresponding "human", "yeast", or "*E. coli*" proteome to each of the identified clusters. In case a cluster contained a consensus identification sequence that could belong to more than one proteome, the said cluster was not considered for the analysis. Clusters without any linked identification information were assigned an "unknown" proteome. For consistency in the comparisons, only clusters within the retention time range between 1500 and 5700 s were considered since data outside of this region contain mostly contaminants or undigested proteins. In the case of MaxQuant, the data are clustered only at the peptide ("peptides.txt") and protein levels ("proteinGroups.txt"). Thus, to assess feature-level quantification, the results from the "evidence.txt" file were grouped by the combination of peptide sequences and charge states.

The following statistics are calculated overall (all samples) and per group (A, B, and C): (1) mean, (2) median, (3) standard deviation, (4) coefficient of variation (CV), (5) median of $\log_{10}$ transformed data, (6) standard deviation of $\log_{10}$ transformed data, (7) percentage of zero values, and (8) number of zero values. Afterward, the difference in median $\log_{10}$ data between sample groups A and C, A and B, and B and C was calculated.

To explore the influence of intensity on the distribution of CVs, we generated 2D density plots of median $\log_{10}$ intensity versus CV % for each sample group (Figure S1).

Finally, the accuracy of the quantification was assessed by comparing the $\log_{10}$ ratio of the different groups as a factor of the intensity with the known expected values for each ratio: $\log_{10}$(A/C) versus $\log_{10}$(C), $\log_{10}$(A/B) versus $\log_{10}$(B), and $\log_{10}$(B/C) versus $\log_{10}$(C). For each of these groups, scatter plots of median $\log_{10}$ values were generated alongside a corresponding density plot (Figure S2). Similarly, the overall distribution of median $\log_{10}$ values in each proteome was also explored in the form of violin plots (Figure S3).

QconCAT Analysis. For the analysis of the QconCAT dataset, the goal was to study the effect of the dynamic range of compounds on the quantitative performance. We decided to focus on feature-level quantification. The applied QconCATs are artificial proteins created from concatenated peptides selected for the intended human and mouse targets. After digestion, the peptides only differ from the endogenous peptides by the presence of the isotope label that each peptide carries on the lysines or arginines.

To ensure that the features being compared are the same across pipelines, we selected only those features in which the peptide sequence and charge states were the same and for which the retention times and m/z are within ±50 s and ±0.05 m/z, respectively.

The wide range of 3 orders of magnitude of spiked-in relative amounts means that we could not assume that detected QconCAT compounds will follow a linear distribution throughout the entire range. Furthermore, to perform linear fitting between the $\log_{10}(\mathrm{ng}/\mu\mathrm{g})$ and $\log_{10}$ intensity values on any given feature cluster, we started by considering all the points in the cluster for the model fit and assessing the resulting R^2. If $R^2 < 0.98$, we iteratively removed low intensity values until the constraint is satisfied. To avoid spurious matches, we only considered features in which the aforementioned linear fit contains values in at least three spiked-in levels. If the fitting is successful, the CV and sum of squared errors (SSE) are calculated for all points in the fitted linear range. Only features that were successfully fitted in both PASTAQ and MaxQuant were considered for further analysis. An example of the fitted models for different compounds can be found in Figure S4. Additionally, we obtained the extracted ion chromatograms (XICs) of the monoisotopic peaks for all selected features after retention time alignment (Figure S5) with an m/z window of ±0.01.

THEORY

Preprocessing of DDA data comprises detecting isotopic peaks or peptide features, aligning the retention time of all chromatograms to a common reference, linking isotopic peaks and features with identification information, matching peaks/features across chromatograms, and generating quantitative tables amenable for statistical analysis.

The parameterization of PASTAQ is very intuitive to anyone familiar with LC−MS instrumentation settings and theory since it is based on the well-understood notions of peak widths and compound separation in chromatography and mass spectrometry. PASTAQ calculates the theoretical width of peaks in the retention time and m/z dimensions by considering the physics of the ion separation and detection process in different MS instruments[11] as well as chromatographic separation theory.[29] For example, peak width modeling in mass spectra is based on the ion separation equation of the mass analyzers, while chromatographic peak modeling is based on the assumption of a constant peak width, as predicted by the linear solvent strength theory for reverse-phase linear gradient elution.[29]

Another important feature of PASTAQ is the automatic generation of quality control plots, which allow us to assess the overall dataset similarity between samples, distribution of peak widths, as well as retention time and mass shifts. These quality control plots can be used to evaluate the quality of the

https://doi.org/10.1021/acs.analchem.1c01892
Anal. Chem. 2021, 93, 11215−11224

preprocessed data and identify issues with the acquired LC−MS/MS data or with the given parameterization (Figure S6). The default parameters were established to accurately process data acquired with the most common acquisition settings, and the user is only required to set three key parameters: the type of mass spectrometer used, the resolution at the reference m/z, and the average full width at half-maximum of chromatographic peaks. Many other parameters in the pipeline, such as the selection of regions of interest for peak detection, the level of smoothing, and the radius for feature/peak matching across chromatograms, are automatically derived from these parameters.

One of the goals of PASTAQ is to provide full data traceability from the beginning to end of the analysis. This allows for posthoc analyses and data exploration. For example, by tracing back all processing steps and intermediate results that comprise a quantified peptide, we can find each fragment ion spectrum associated with the isotopic peaks in all available files and display relevant information of retention time and mass shifts in all datasets. The spectra associated with a matched isotope or matched peptide feature can also be extracted from the raw data.

The core of PASTAQ is built as a C++ library, with bindings in the Python programing language. It can be used in all major operating systems (i.e., Windows, Mac, and Linux), including high-performance computing clusters. This allows PASTAQ to be easily integrated into existing workflows and LC−MS/MS analysis pipelines. Additionally, the Python bindings enable PASTAQ to be extended to suit the needs of different datasets and allow quick iteration and prototyping of new ideas, such as the generation of consensus spectra from matched and/or identified MS/MS spectra. The source code is available under a permissive open-source license (MIT) and are publicly available at https://pastaq.horvatovichlab.com.

This combination of features makes PASTAQ suitable for beginners and advanced users alike. Running the basic pipeline is simple, but more complex analyses can be performed using the Python bindings to derive further insights from the data.

Many of the steps necessary for LC−MS/MS data preprocessing can be computationally demanding. For instance, when considering a large number of samples, millions of isotopic peaks need to be quantified, aligned, and matched with suitable candidates throughout the entire dataset. Moreover, selecting the proper preprocessing parameters can be challenging due to the complexity of LC−MS/MS data, which often needs to be analyzed multiple times to test different identification or quantification parameters. PASTAQ circumvents this by separating the identification and quantification steps, and the main preprocessing algorithms try to take advantage of multicore processing when available.

The datasets presented here were analyzed on a workstation with an Intel(R) Core(TM) i7-8700K CPU running at 3.7 GHz and 64 GB of RAM memory running Linux 5.9.3. The data were stored in four Western Digital WD Blue 6TB hard drives in a RAID10 configuration under the BTRFS file system.

A brief description of the main preprocessing modules of the DDA pipeline is provided below, and a more detailed explanation can be found in Supporting Information. A diagram of how the different modules are connected in PASTAQ's DDA pipeline can be seen in Figure S7.

RESULTS

Peak Detection and Feature Detection. PASTAQ detects isotopic peaks by considering their three-dimensional nature in LC−MS/MS data. While mass spectra are smoothed in Orbitrap high-resolution mass spectrometers on the acquisition electronic board, which reduces the noise in the m/z dimension, noise is unaffected in the chromatographic dimension. Time-of-flight and quadrupole-based mass spectrometers tend to have a higher level of noise in both separation dimensions. The width of peaks in the m/z dimension is highly dependent on the instrument resolution at a given m/z, as well as the type of the mass spectrometer used. For accurate quantification, PASTAQ performs a simultaneous 2D Gaussian kernel smoothing and resampling of the spectra that maps the acquired raw data into a regular grid, keeping the number of sampling points per isotope peak constant throughout the entire m/z range independent of their width. This results in consistently sized peaks across the entire m/z range as well as reduced memory usage.

The smoothed grid is searched for the local maxima, and a novel, fast method for noniterative 3D Gaussian fitting is used to minimize the error of the Gaussian model applied to the raw data using the location of peak maxima in the smoothed grid. This results in superior quantitative performance while being computationally efficient, which makes PASTAQ suitable for processing large datasets in a reasonable time. This process results in a list of modeled peaks for each chromatogram, describing their m/z and retention time location as well as their height and width in the respective dimensions. In our work, PASTAQ detected an average of 550,393 isotopic peaks for the chromatograms of the HYE dataset and 398,521 for the QconCAT dataset.

For this first stage of the pipeline, it took an average of 17 s to parse 6 GB profile mzXML files for MS1 spectra and 33 s for MS2. Peak detection was performed in around 10 s for resampling and 4 s for peak detection, 2D Gaussian fitting and quantification per file. The total mzXML parsing time for the 30 samples in the HYE dataset, including saving the detected raw spectra to disk in binary format, was 32 min. The peak detection procedure finished in less than 40 min.

It is often desirable to group peaks belonging to the same isotopic envelope into "features". To achieve this, PASTAQ uses the previously obtained peak lists to generate undirected graphs, in which the peaks are tentatively linked if they are within a close retention time range and their m/z location difference corresponds to 1.0033 divided by the candidate charge state. The range of candidate charge states is set by the user and should be typically between 1 and 8. The tolerance for retention time and m/z is set as a unit of peak width (sigma) in the respective dimensions. Features are formed by using these graphs to find the best matching isotopic patterns to the appropriate Averagine model.[30] By using this approach, an average of 150,793 features was obtained for the HYE dataset and 116,519 for the QconCAT dataset. This corresponds to respective averages of 3.42 and 3.65 isotopes per detected feature. Despite the inherent complexity of the task, feature detection was performed in around 4 s per file and less than 5 min for all the 30 samples in the HYE dataset.

Retention Time Alignment. Aligning the retention time of all chromatograms in a dataset is a crucial step to be able to match compounds between samples using predominantly MS1-based approach. For this purpose, PASTAQ uses an improved version of the Warp2D algorithm.[31] This method allows the alignment of two chromatograms by maximizing the similarity function of their respective peak lists based on the sum of overlapping 2D Gaussian peaks. An extra benefit of this approach is that the calculation of similarity values across all

https://doi.org/10.1021/acs.analchem.1c01892
Anal. Chem. 2021, 93, 11215−11224

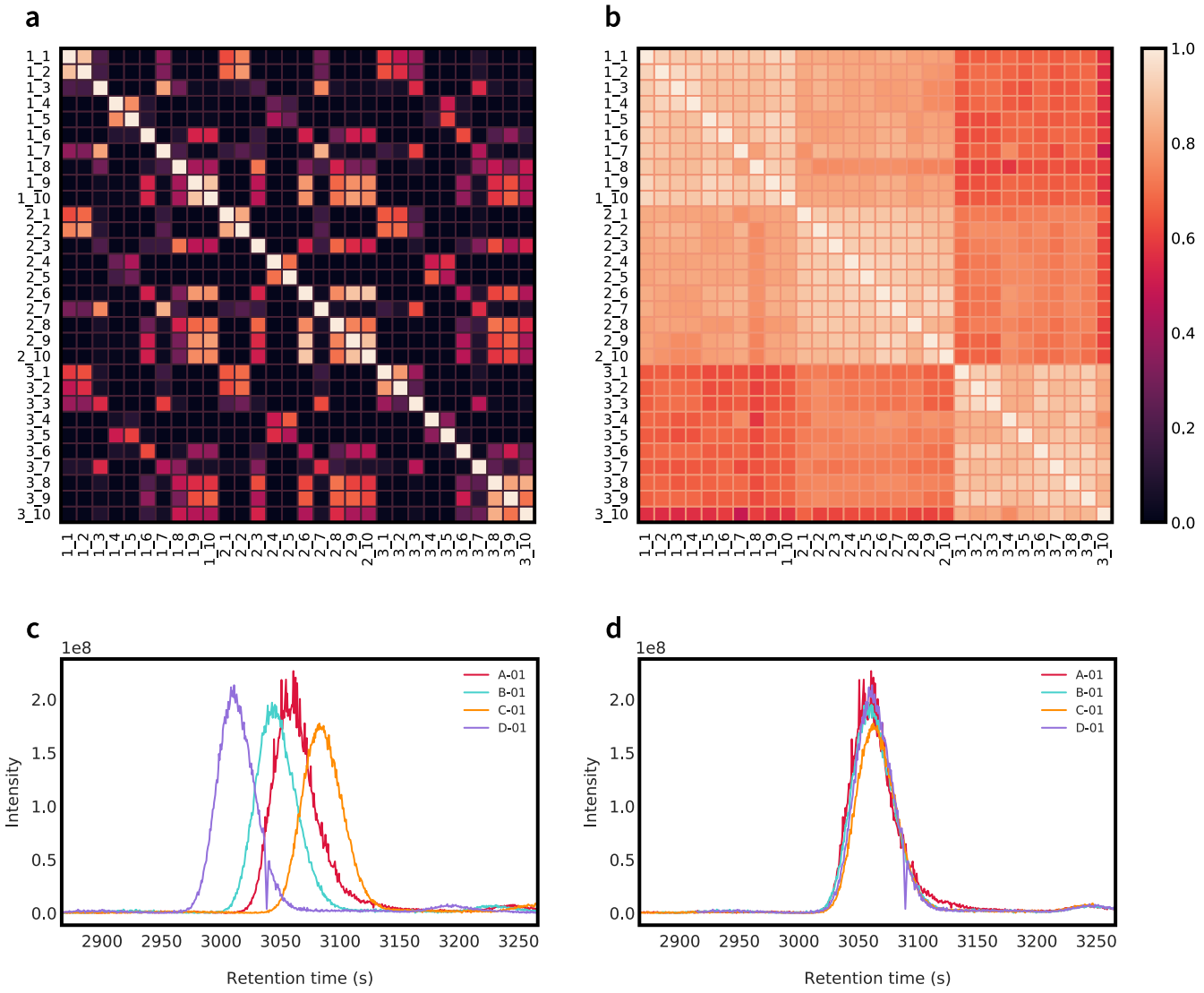


Figure 1. Effect of the retention time alignment on the similarity matrix of the HYE dataset (top) and the extracted ion chromatogram of an isotopic peak from the QconCAT dataset (bottom), before (a,c) and after (b,d) alignment.

samples allows the identification of potential outliers. For example, in Figure 1a,b, the similarity matrix before and after alignment shows a remarkable improvement after Warp2D, where the replicates of the three different ratios of the HYE dataset are clearly differentiated.

Performing retention time alignment of two samples takes roughly 12 s with the default parameters. If a given sample is used as a reference, the 30 replicates of the HYE dataset can be aligned in less than 6 min. If no reference is given for retention time alignment, an exhaustive search is performed in order to select the best reference to maximize the average similarity for all samples. This optional step can become the most computationally expensive part of the pipeline for large datasets as the time required will increase proportionally to the square of the number of samples in the dataset. In the case of large datasets, it is possible to skip this step or to select a subset of samples on which alignment between all pairs can be performed. For the HYE dataset, the exhaustive reference search took 86 min until completion.

Peak Matching and Feature Matching. Matching peaks and features belonging to the same compounds across multiple samples is necessary for assessing relative differences in intensity. In PASTAQ, this task is greatly simplified thanks to the robust and accurate retention time alignment algorithm used in the pipeline. The process consists of comparing the retention time and m/z locations within a tolerance range dictated by a given number of sigmas in the retention time and m/z dimensions. The matching is the same for peaks and features, but in the case of features, the monoisotopic m/z is used and only features that share the same charge state (as determined by the feature matching procedure in PASTAQ) are considered. At the end of this process, a list of clustered isotopic peaks and features is obtained.

To reduce the effect of noise on the data, an optional filtering step will keep the clusters in which at least a minimum percentage of samples from any of the groups under study contain detected values. Matching based exclusively on the MS1 location allows consistent matching of all detected MS1 peaks or features, independent of their identification. This approach avoids the common risk of identification-based methods, namely, using two different types of matching for identified and unidentified peaks, where the latter is often based on identification transfer. The MS/MS-based annotation of MS1 peaks in individual samples allows us to assess consistency of

https://doi.org/10.1021/acs.analchem.1c01892
Anal. Chem. 2021, 93, 11215−11224

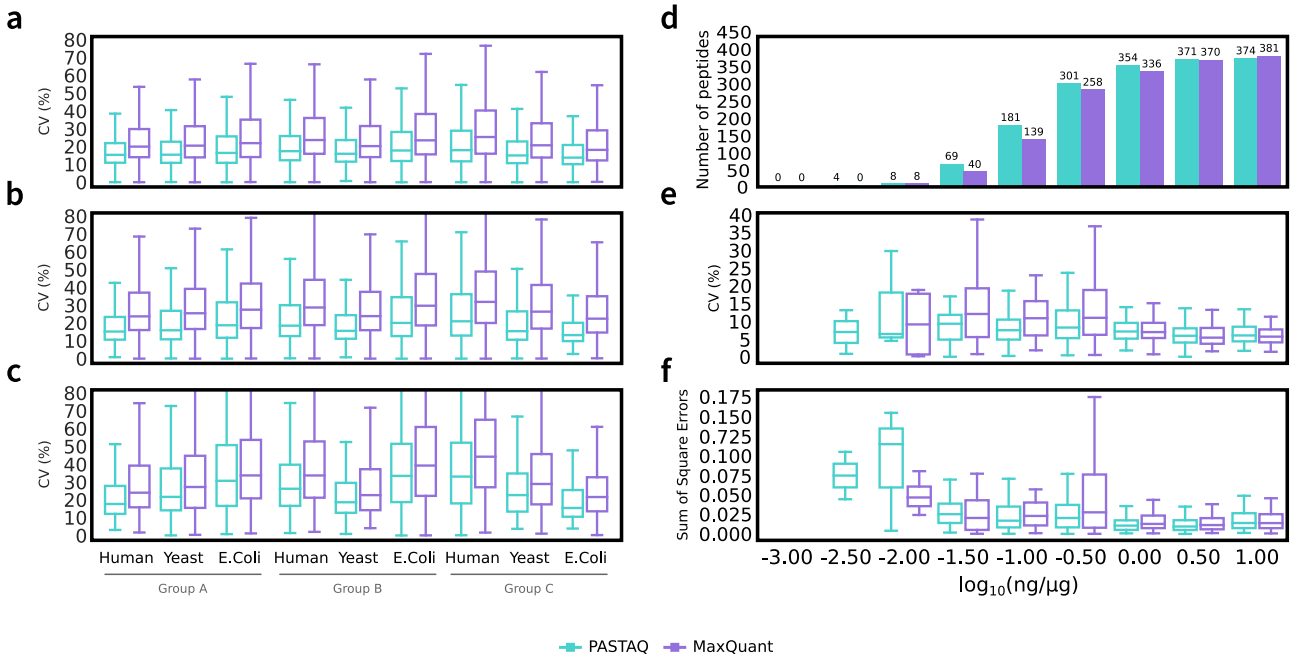


Figure 2. (a−c) Comparison of the CV (%) between PASTAQ quantification with PeptideShaker's consensus identifications and MaxQuant for (a) features, (b) peptides, and (c) proteins in the HYE dataset. (d−f) Evaluation of (d) number, (e) CV, and (f) sum of square errors of quantified peptide for each $\log_{10}$ spiked-in level in the QconCAT dataset.

identification annotation in matched clusters and for the creation of a consensus identification for each cluster. With our matching algorithm, the matching of 639,543 isotopic peak clusters in the HYE dataset took 3 min and 30 s and 65 s for 83,295 feature clusters. Similarly, we obtained 828,801 isotopic peaks and 229,307 feature clusters for the QconCAT dataset.

MS1-Based Quantification, Combined with Threshold-Avoiding Peak Rejection, Boosts the Accuracy, Precision, and Number of Quantified Compounds. Analysis of the chosen datasets mainly focuses on the quantitative performance of PASTAQ compared with the widely used MaxQuant software. The first quantitative metric assessed was the CV of each of the 10 replicates in each group for all measured peptides in the HYE dataset. The CV is calculated for features, peptides, as well as protein groups. For MaxQuant, we used the "evidence.txt" file for assessment of features, the "peptides.txt" file for peptide level quantification, and the "proteinGroups.txt" file for protein group quantification. The combination of PASTAQ quantification with PeptideShaker consensus identifications resulted in lower CV between replicates, as indicated in Table S1. A more detailed overview of the analysis process can be found in Supporting Information. Furthermore, when comparing the distribution of the CV for all quantified features, peptides, and proteins, PASTAQ has a smaller interquartile range. As shown in Figure 2a−c and Table S1, this applies for features as well as protein groups, but the difference is smaller for the latter, likely due to the error inherent to the aggregation of peptides to proteins, which relates to the so-called protein inference problem.[32]

To assess how well the different pipelines perform to detect and quantify peptides spiked across 3 orders of magnitude, we used the QconCAT dataset. The features reported in the "evidence.txt" file from MaxQuant were matched with the corresponding features from PASTAQ using their associated peptide sequence and charge state, excluding ambiguous identifications. For each of the matched features between PASTAQ and MaxQuant, a linear fitting was performed over the $\log_{10}$ transformed intensity versus the nanograms of the spiked protein over micrograms of the total protein (ng/μg) to select the spiked-in levels that fall within the linear range (Figure S4), as described in more detail in Supporting Information. This resulted in 118 features, for which the CV, SSE of the linear fit, and number of features present in replicates at each spiked-in level were calculated. While both pipelines performed similarly in groups with higher concatemer amount (1.00−10.00 ng/μg) (±5% difference in the number of quantified peptides), we observe an increase of 16.7−72.5% in the number of quantified peptides with PASTAQ compared to MaxQuant (using the match-between-runs option) when the spiked-in amount is below 1 ng/μg. All spiked-in features show similar or smaller CV and SSE in PASTAQ, as shown in Figure 2d−f and Table 1.

The threshold-avoiding methodology in PASTAQ is key to the larger number of peptides detected at the low spiked-in levels since faint peaks are retained when they appear consistently across samples, increasing the faint true positives, while more

Table 1. Difference in the Number of Detected Peptides between PASTAQ and MaxQuant for the Different Spiked-In Levels in the QconCAT Dataset

spiked-in amount		num. peptides		
ng/μg	$\log_{10}$(ng/μg)	PASTAQ	MaxQuant	difference (%)
10.0000	1.0	374	381	−1.85
3.16230	0.5	371	370	0.27
1.00000	0.0	354	336	5.36
0.31623	−0.5	301	258	16.67
0.10000	−1.0	181	139	30.22
0.03162	−1.5	69	40	72.50
0.01000	−2.0	8	8	0.00
0.00316	−2.5	4	0	NA
0.00100	−3.0	0	0	NA

https://doi.org/10.1021/acs.analchem.1c01892
Anal. Chem. 2021, 93, 11215−11224

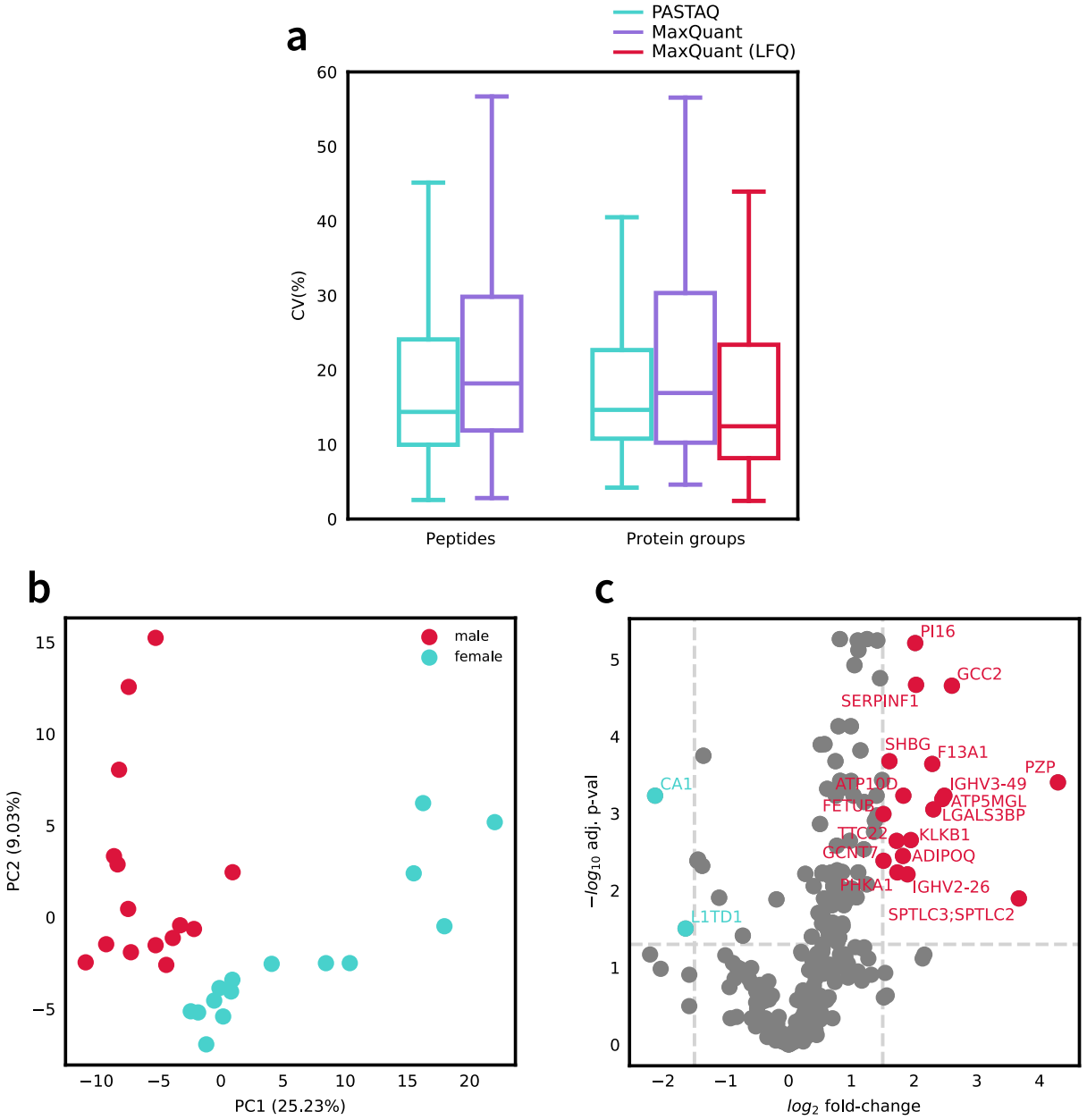


Figure 3. Evaluation of PASTAQ performance on human serum datasets: (a) distribution of CVs on the peptide and protein group quantification levels for the 15 technical replicate dataset. (b) PCA of the male−female dataset, showing a clear separation between groups. (c) Volcano plot showing selected gender-related proteins for the male−female dataset based on the adjusted *p*-values from Welch's *t*-test and the $\log_2$ fold-change.

abundant peaks that appear less consistently are rejected, decreasing the strong false positives.

Enhanced Detection of Faint Peaks Enables Identification of a Larger Number of Quantified Proteins in Biological Datasets. To show the performance of PASTAQ for preprocessing biological LC−MS/MS data, we have used two label-free LC−MS/MS proteomics datasets of human serum acquired with a short gradient for blood-based biomarker profiling.[18] First, we focused on evaluating the reproducibility (CV %) and number of protein groups using a set of 15 technical replicates from a male subject. For these analyses, we used PASTAQ with the MSFragger identification engine, which resulted in 448 protein groups quantified in more than 60% of the replicates (Table S2). The data processed with MaxQuant with the match-between-runs option enabled contain 278 quantified protein groups in more than 60% of the replicates. The median CV of the quantified protein groups with PASTAQ was 14.6%, which is between the values obtained with the original MaxQuant quantification (16.9%) and after MaxLFQ[33] normalization (12.4%). The standard deviation of the CV values is of 10.8% for PASTAQ, approximately 2 times lower than MaxQuant (23.9%) or MaxLFQ (19.5%), indicating the consistency of the quantification by PASTAQ when analyzing serum sample replicates (Figure 3a).

LC−MS/MS data are often used for biomarker discovery. To demonstrate the suitability of PASTAQ for this task, we processed the datasets from 5 male and 5 female subjects (in triplicate) from the aforementioned work of Geyer et al.[18] The data processing parameterization and 60% filtering with PASTAQ were the same as in the previous example, yielding 291 protein groups. We performed a principal-component analysis (PCA), in which a difference between the male and

https://doi.org/10.1021/acs.analchem.1c01892
Anal. Chem. 2021, 93, 11215−11224

female groups can be clearly observed (Figure 3b). A $\log_2$ transformation was applied to the data to account for its log−normal distribution. In addition, Welch's *t*-test was performed on each protein group and the resulting *p*-values were adjusted for multiple testing using the Benjamini−Hochberg[34] method, resulting in 100 significant protein groups (adjusted $p < 0.05$). The statistical results are visualized as a volcano plot in Figure 3c. We focused on protein groups with a $\log_2$ fold-change of more than 1.5, yielding several proteins already documented in the literature, including two of the reported significant proteins in the original work of Geyer et al.:[18] pregnancy-zone protein (PZP) and sex hormone-binding globulin (SHBG). Furthermore, we found fetuin-B (FETUB), phospholipid-transporting ATPase VD (ATP10D), and tetratricopeptide repeat protein 22 (TTC22) to be different between males and females. The level of FETUB is known to follow the menstrual cycle and correlates with women's fertility,[35] while TTC22 and ATP10D are known to be highly expressed in female tissues based on the human protein atlas.[36−38]

DISCUSSION

In this work, we show how the innovative design of the preprocessing and quantification methods of PASTAQ lead to the detection of a large number of low-intensity signals from LC−MS/MS data while offering excellent reproducibility and accurate quantification for low abundant peptides and proteins. This is due to the combined use of smoothed and raw data for peak detection and quantification, the avoidance of early thresholding, which enlarges the dynamic quantification range, the use of overlapping peak volumes in WARP2D retention time alignment to automatically and accurately align multiple chromatograms, as well as the exclusive MS1-based peak matching. Quality metrics such as exhaustive pairwise similarity matrices before and after alignment may be used to assess the accuracy of data processing. Linking preprocessing parameters to peak widths in the mass and retention time domains allows straightforward parameterization. Furthermore, the speed of the preprocessing algorithms used in PASTAQ allows rapid iteration of different parameters and/or identification engines when analyzing large datasets, thus enabling an efficient parameter optimization in order to obtain higher quantification accuracy.

The stochasticity of fragmentation in DDA LC−MS/MS can lead to situations where no identification information is available for the detected features in the data. This has important consequences for experiments that emphasize the search for new biomarkers since the information in the entire dataset can be used to select a list of candidates even when no identification is available. PASTAQ allows us to collect all MS/MS spectra linked to matched peaks or features and allows manual interpretation of individual or consensus spectra in order to elucidate the compound to which they belong. This is particularly relevant for metabolomics and lipidomics datasets, where identification remains a considerable challenge.

In this paper, we highlighted the use of PASTAQ for preprocessing DDA proteomics data, but the given tools are not limited to this domain. Data-independent acquisition (DIA) methods attempt to address the stochasticity issue by performing comprehensive fragmentation of the entire mass range using consecutive and large precursor selection windows. This method has proven to be more effective in extracting a large number of identifications from the data,[39] but it has some disadvantages, including the increased complexity of data analysis due to heavily convoluted MS/MS spectra that require more elaborate approaches for data processing and quantification. We are actively working toward the implementation of novel DIA quantifications methods in PASTAQ, as well as its usage with metabolomics and lipidomics datasets.

In summary, PASTAQ represents a step forward in LC−MS/MS data preprocessing. The straightforward set of tools in its suite allow the user to extract additional and more reliable biological insights from the data and identify potential sources of error, enabling faster analysis cycles and providing the necessary tools for the in-depth exploration of LC−MS/MS datasets.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.1c01892.

Detailed description of PASTAQ's algorithms: peak detection and isotope quantification, 2D Gaussian fitting using least-squares and SVD, retention time alignment, linkage of peaks to MS/MS events and identifications, feature detection, matching isotopic peaks and features across samples, peptide and protein group quantification, quantitative table generation, quality control plots, Supplementary figures: 2D density distribution of CVs versus intensity, scatterplot of log(A/C) versus log(C) of detected features, violin plot with log distribution of intensity ratios, example of linear fitting of two QconCAT features, XIC for two features at different spiked-in levels, example of automatically generated quality control plots, diagram of PASTAQ's main modules for the DDA pipeline, comparison of the distribution of CVs between PASTAQ and MaxQuant, number of quantified protein groups, and distribution of CVs for the biological dataset (PDF)

AUTHOR INFORMATION

Corresponding Author

Péter Horvatovich − *Department of Analytical Biochemistry, Groningen Research Institute of Pharmacy, University of Groningen, 9713 AV Groningen, The Netherlands;* orcid.org/0000-0003-2218-1140; Email: p.l.horvatovich@rug.nl

Authors

Alejandro Sánchez Brotons − *Department of Analytical Biochemistry, Groningen Research Institute of Pharmacy, University of Groningen, 9713 AV Groningen, The Netherlands*

Jonatan O. Eriksson − *Department of Biomedical Engineering, Lund University, 221 84 Lund, Sweden*

Marcel Kwiatkowski − *Department of Analytical Biochemistry, Groningen Research Institute of Pharmacy, University of Groningen, 9713 AV Groningen, The Netherlands; Functional Proteo-Metabolomics, Department of Biochemistry, University of Innsbruck, A-6020 Innsbruck, Austria;* orcid.org/0000-0002-5804-6031

Justina C. Wolters − *Department of Pediatrics, University Medical Center Groningen, University of Groningen, 9713 GZ Groningen, The Netherlands*

Ido P. Kema − *Department of Laboratory Medicine, University Medical Center Groningen, University of Groningen, 9700 RB*

Groningen, The Netherlands; orcid.org/0000-0003-1166-6169

Andrei Barcaru − *Department of Analytical Biochemistry, Groningen Research Institute of Pharmacy, University of Groningen, 9713 AV Groningen, The Netherlands*

Folkert Kuipers − *Department of Pediatrics, University Medical Center Groningen, University of Groningen, 9713 GZ Groningen, The Netherlands; Department of Laboratory Medicine, University Medical Center Groningen, University of Groningen, 9700 RB Groningen, The Netherlands*

Stephan J. L. Bakker − *Department of Internal Medicine, Division of Nephrology, University Medical Center Groningen, University of Groningen, 9713 GZ Groningen, The Netherlands*

Rainer Bischoff − *Department of Analytical Biochemistry, Groningen Research Institute of Pharmacy, University of Groningen, 9713 AV Groningen, The Netherlands;* orcid.org/0000-0001-9849-0121

Frank Suits − *IBM Research—Australia, Southbank 3006 Victoria, Australia*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.1c01892

Author Contributions

P.H., F.S. and R.B. designed and supervised the study; A.S.B., P.H., F.S. and A.B. contributed to the design of computational algorithms, A.S.B. designed and implemented PASTAQ; J.O.E. performed independent testing of PASTAQ; M.K. and J.C.W. performed sample preparation and acquisition of LC−MS/MS datasets; A.S.B., S.J.L.B., I.K., F.K., M.K., R.B., and F.S., discussed assessment and quality metrics and contributed to interpretation of the results; A.S.B. and P.H. wrote the paper with contributions from all authors.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was part of the Netherlands X-omics Initiative and partially funded by NWO, project 184.034.019.

■ REFERENCES

(1) Wiener, M. C.; Sachs, J. R.; Deyanova, E. G.; Yates, N. A. *Anal. Chem.* **2004**, *76*, 6085−6096.
(2) Levin, Y.; Schwarz, E.; Wang, L.; Leweke, F. M.; Bahn, S. *J. Sep. Sci.* **2007**, *30*, 2198−2203.
(3) McCormack, A. L.; Schieltz, D. M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J. R. *Anal. Chem.* **1997**, *69*, 767−776.
(4) Maurer, H. H. *Clin. Biochem.* **2005**, *38*, 310−318.
(5) Lee, M. S.; Kerns, E. H. *Mass Spectrom. Rev.* **1999**, *18*, 187−279.
(6) Fernie, A. R.; Trethewey, R. N.; Krotzky, A. J.; Willmitzer, L. *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 763−769.
(7) Cox, J.; Mann, M. *Nat. Biotechnol.* **2008**, *26*, 1367−1372.
(8) Weisser, H.; Nahnsen, S.; Grossmann, J.; Nilse, L.; Quandt, A.; Brauer, H.; Sturm, M.; Kenar, E.; Kohlbacher, O.; Aebersold, R.; Malmström, L. *J. Proteome Res.* **2013**, *12*, 1628−1644.
(9) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−787.
(10) Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-Arendt, K.; Ahn, N. G.; Old, W. M. *J. Proteome Res.* **2010**, *9*, 4152−4160.
(11) Suits, F.; Hoekman, B.; Rosenling, T.; Bischoff, R.; Horvatovich, P. *Anal. Chem.* **2011**, *83*, 7786−7794.
(12) Mayer, G.; Proteom-center, M.; Bochum, R.-u.; Eisenacher, M. *mzIdentML: Exchange Format for Peptides and Proteins Identified from Mass Spectra*, 2011, pp 67.report
(13) Vaudel, M.; Burkhart, J. M.; Zahedi, R. P.; Oveland, E.; Berven, F. S.; Sickmann, A.; Martens, L.; Barsnes, H. *Nat. Biotechnol.* **2015**, *33*, 22−24.
(14) Keller, A. D.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of protein identifications made by MS/MS and database search. *Proceedings 50th ASMS Conference on Mass Spectrometry and Allied Topics*, 2002; Vol. *74*, pp 37−38.
(15) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 4646−4658.
(16) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. *Nat. Methods* **2007**, *4*, 923−925.
(17) Wolters, J. C.; Ciapaite, J.; Van Eunen, K.; Niezen-Koning, K. E.; Matton, A.; Porte, R. J.; Horvatovich, P.; Bakker, B. M.; Bischoff, R.; Permentier, H. P. *J. Proteome Res.* **2016**, *15*, 3204−3213.
(18) Geyer, P. E.; Kulak, N. A.; Pichler, G.; Holdt, L. M.; Teupser, D.; Mann, M. *Cell Syst.* **2016**, *2*, 185−195.
(19) Pedrioli, P. G. A.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R. *Nat. Biotechnol.* **2004**, *22*, 1459.
(20) 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. *Mol. Cell. Proteomics* **2011**, *10*, R110.000133.
(21) Barsnes, H.; Vaudel, M. *J. Proteome Res.* **2018**, *17*, 2552−2555.
(22) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.
(23) Chambers, M. C.; MacLean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M.-Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. *Nat. Biotechnol.* **2012**, *30*, 918−920.
(24) Consortium TU. *Nucleic Acids Res.* **2021**, *49*, D480−D489.
(25) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. *Nat. Methods* **2017**, *14*, 513−520.
(26) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20*, 1466−1467.
(27) Dorfer, V.; Pichler, P.; Stranzl, T.; Stadlmann, J.; Taus, T.; Winkler, S.; Mechtler, K. *J. Proteome Res.* **2014**, *13*, 3679−3684.
(28) Kim, S.; Pevzner, P. A. *Nat. Commun.* **2014**, *5*, 5277.
(29) Blumberg, L. M. *Chromatographia* **2014**, *77*, 189−197.
(30) Senko, M. W.; Beu, S. C.; Mclafferty, F. W. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 229.
(31) Suits, F.; Lepre, J.; Du, P.; Bischoff, R.; Horvatovich, P. *Anal. Chem.* **2008**, *80*, 3095−3104.
(32) Nesvizhskii, A. I.; Aebersold, R. *Mol. Cell. Proteomics* **2005**, *4*, 1419−1440.
(33) Cox, J.; Hein, M. Y.; Luber, C. A.; Paron, I.; Nagaraj, N.; Mann, M. *Mol. Cell. Proteomics* **2014**, *13*, 2513−2526.
(34) Benjamini, Y.; Hochberg, Y. *J. Roy. Stat. Soc. B* **1995**, *57*, 289−300.
(35) Fang, L.; Hu, X.; Cui, L.; Lv, P.; Ma, X.; Ye, Y. *J. Assist. Reprod. Genet.* **2019**, *36*, 1101−1107.
(36) Uhlen, M.; Fagerberg, L.; Hallström, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, Å.; Kampf, C.; Sjöstedt, E.; Asplund, A.; Olsson, I. M.; Edlund, K.; Lundberg, E.; Navani, S.; Szigyarto, C. A. K.; Odeberg, J.; Djureinovic, D.; Takanen, J. O.; Hober, S.; Alm, T.; Edqvist, P. H.; Berling, H.; Tegel, H.; Mulder, J.; Rockberg, J.; Nilsson, P.; Schwenk, J. M.; Hamsten, M.; Von Feilitzen, K.; Forsberg, M.; Persson, L.; Johansson, F.; Zwahlen, M.; Von Heijne, G.; Nielsen, J.; Pontén, F. *Science* **2015**, *347*, 1260419.

(37) Uhlen, M.; Zhang, C.; Lee, S.; Sjöstedt, E.; Fagerberg, L.; Bidkhori, G.; Benfeitas, R.; Arif, M.; Liu, Z.; Edfors, F.; Sanli, K.; Von Feilitzen, K.; Oksvold, P.; Lundberg, E.; Hober, S.; Nilsson, P.; Mattsson, J.; Schwenk, J. M.; Brunnström, H.; Glimelius, B.; Sjöblom, T.; Edqvist, P. H.; Djureinovic, D.; Micke, P.; Lindskog, C.; Mardinoglu, A.; Ponten, F. *Science* **2017**, *357*, No. eaan2507.
(38) Thul, P. J.; Akesson, L.; Wiking, M.; Mahdessian, D.; Geladaki, A.; Ait Blal, H.; Alm, T.; Asplund, A.; Björk, L.; Breckels, L. M.; Bäckström, A.; Danielsson, F.; Fagerberg, L.; Fall, J.; Gatto, L.; Gnann, C.; Hober, S.; Hjelmare, M.; Johansson, F.; Lee, S.; Lindskog, C.; Mulder, J.; Mulvey, C. M.; Nilsson, P.; Oksvold, P.; Rockberg, J.; Schutten, R.; Schwenk, J. M.; Sivertsson, A.; Sjöstedt, E.; Skogs, M.; Stadler, C.; Sullivan, D. P.; Tegel, H.; Winsnes, C.; Zhang, C.; Zwahlen, M.; Mardinoglu, A.; Pontén, F.; Von Feilitzen, K.; Lilley, K. S.; Uhlén, M.; Lundberg, E. *Science* **2017**, *356*, No. eaal3321.
(39) Gillet, L. C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. *Mol. Cell. Proteomics* **2012**, *11*, O111.016717.

https://doi.org/10.1021/acs.analchem.1c01892
Anal. Chem. 2021, 93, 11215−11224