

Article

Not peer-reviewed version

Observations from the proteomics bench

[Simone König](#)*, [Karin Schork](#), [Martin Eisenacher](#)

Posted Date: 5 January 2024

doi: 10.20944/preprints202401.0493.v1

Keywords: mass spectrometry; bioinformatics; proteomics; false-positives



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Article

Observations from the Proteomics Bench

Simone König^{1,*}, Karin Schork^{2,3,4} and Martin Eisenacher^{2,3,4}

¹ IZKF Core Unit Proteomics, University of Münster, 48149 Münster, Germany; koenigs@uni-muenster.de (S.K.)

² Medizinisches Proteom-Center, Medical Faculty, Ruhr-University Bochum, 44801 Bochum, Germany

³ Center for Protein Diagnostics (PRODI), Medical Proteome Analysis, Ruhr-University Bochum, 44801 Bochum, Germany; Karin.schork@rub.de (K. S.), martin.eisenacher@rub.de (M. E.)

⁴ Ruhr University Bochum, Medical Faculty, Core Unit for Bioinformatics (CUBiMed.RUB)

* Correspondence: koenigs@uni-muenster.de; Tel.: 0049-251 8357164

Abstract: Many challenges in proteomics result from the high-throughput nature of the experiments. The paper first presents pre-analytical problems, which still occur, although the call for standardisation in omics has been ongoing for many years. This article also discusses aspects, which affect bioinformatic analysis, based on three sets of reference data measured with different orbitrap instruments. Despite continuous advances in mass spectrometer technology as well as analysis software, data set-wise quality control is still necessary and decoy-based estimation, although challenged by modern instruments, should be utilized. We draw attention to the fact that numerous young researchers perceive proteomics as a mature, readily applicable technology. However, it is important to emphasize that the maximum potential of the technology can only be realized by an educated handling of its limitations.

Keywords: mass spectrometry; bioinformatics; proteomics; false-positives

1. Introduction

The journal *Proteomes* is celebrating its 10 year anniversary and it has become one of the established proteomics journals. When the Joint Editor-In-Chief asked us to contribute to the anniversary issue “with thoughts and reflections of what proteomics has achieved through its history and especially the last ten years”, we did not hesitate very long. The Editors wanted “honest opinions that challenge some of the dogma that persists in the proteomics field”, and we do have a number of those. Over the years, several researchers have pointed out problems in proteomics (see, [1–3] and references therein) and it is not our goal to repeat them. We would like to, however, present topics and difficulties, which we still observe in our daily routines despite many years of great technical advances in the field. To that end, a mass spectrometry (MS) practitioner and bioinformaticians working in the proteomics field joined forces with insights from the analysis of both Q-TOF and orbitrap instrument data. We do not offer a comprehensive overview, but rather focus on a few topics we encountered in our work.

2. Results and Discussion - Challenges in Proteomics

2.1. Preanalytical Issues

Omics researchers are well aware of the need for standardization of the individual experiment as well as the entire project in order to limit, if not eliminate, variation introduced by sample handling (for exemplary studies, see [4–6]). That is easier said than done, especially when multiple laboratories are involved; then, this information is somehow often lost. Parameters such as temperature, sitting times, and centrifugation speed can have considerable influence, as well as personal ambition, skill and technical experience.

2.1.1. Example: Serum Quality

We noted varying serum quality in sample sets of several studies. In one case, only about two thirds of the samples sent to us were usable for our purpose. Using bradykinin as a reporter peptide [7], we detected a loss of protease activity in these samples compared to all others. We show a photograph of an exemplary 96-well plate of a sample cohort containing some hemolytic samples in Supplementary Figure S1. Clearly, the rupturing of red blood cells with release of their content into the blood fluid influences the results of subsequent analyses. According to the hemolysis palette published by the Centers for Disease Control and Prevention CDC [8] (Supplementary Figure S1), deep orange to red serum samples are not fit for certain analyses like our assay or omics-type of experiments. Although instructions to prevent hemolysis are abundantly available from different sources including the CDC [8] and taught to the medical staff, there still seem to be ample sources for mistakes or reasons for shortcuts.

In this particular sample set, we also detected significant differences in enzyme activity even for non-hemolytic serum as a batch effect resulting from the time and place of sampling. That was the more surprising as all collaborating parties were aware of the need for standardization and precautions had been taken. The reasons for the variation were never properly clarified. In another project, the lower protease activity in some non-hemolytic sera obtained from our clinic was traced back to the random use of different collection tubes by the medical staff.

2.1.2. Example: Tissue Origin

In typical omics projects, two or more groups of samples are compared, e.g. untreated versus treated cells. In cell culture experiments or even in research with laboratory animals, the respective proteomes of each group are well-defined and similar, which is a prerequisite for successful protein abundance analysis. The situation is much more complex in clinical studies, because of the large biological variation in patients, even when the obvious criteria such as gender and age are taken into account. There are simply too many parameters like nutrition, (self) medication, level of exercise, body weight, genetic predisposition and sanitary conditions to assemble perfect study cohorts. Thus, researchers try to work with large groups in order to alleviate some of the effects. However, for investigations of rare diseases only few individuals are available in the first place.

On top of these known limitations, we noted in several clinical studies that the proteomics data of biofluids or tissue samples are of such resolution and quality that individual subgroups can be identified. For instance, in a project working with stomach biopsies, we detected non-uniform protein profiles in the biopsies from the control group of healthy mucosa (Figure 1A) [9]. A double-check with the surgeon revealed that some samples were picked from corpus rather than antrum. It is known, however, since at least the work of Ni et al. [10] that mucosa proteomes differ across the stomach reflecting regions of different task and the presence of specific glands. Conclusively, in order not to distort the results of the investigation, we had to eliminate the corpus samples from the project. On a side note, we also had another problem in that project: in the operating theater, some biopsies had been stored in a different way from the others, promptly resulting in polymer contaminants in these samples, which showed up in the chromatogram; also these samples were lost to the study. Unfortunately, all of these observations were made after complete processing including MS analysis had been performed; in fact, the measurements themselves lead to questioning of the biopsy quality.

In the second example in Figure 1B, we show serum protein profiles of hospitalized COVID-19 patients [11]. There, it was not too surprising that they differed considerably, because the patients were very diverse to begin with; the only unifying parameters were hospitalization and proven COVID-19 infection. Sub-groups could be assembled from these data, which separated moderately from severely and critically ill patients. Again, it was not sensible to use the entire patient group for comparative analyses versus the control groups, but of course, the power of the analysis decreases with the number of subgroups.

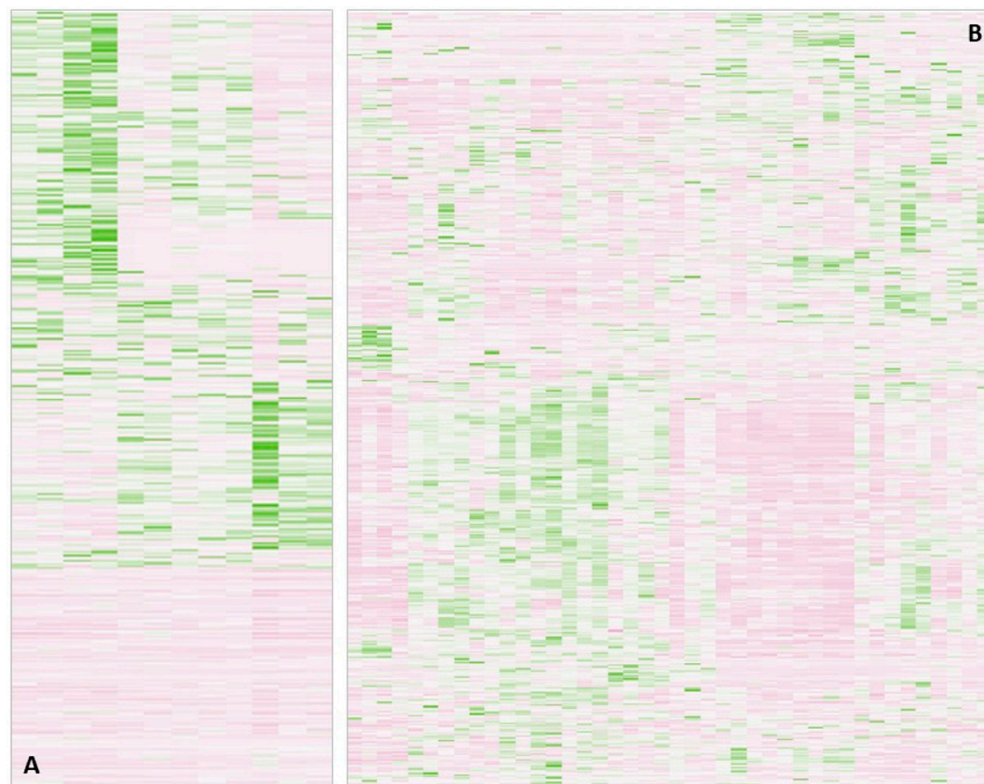


Figure 1. Heatmaps obtained in two different proteomics projects. (A) Proteins measured in healthy stomach mucosa used as control versus gastritis and cancer biopsies. B) Serum protein profiles from COVID-19 hospitalized patients. For more information, see references [9,11]. Both examples illustrate non-uniform sub-groups within the chosen sample cohort.

2.2. Peptide Sequencing

The majority of proteomics studies is based on the so-called bottom-up approach in which the proteins are enzymatically digested before analysis and the peptides are measured. This procedure has several advantages including the fact that peptides can be measured monoisotopically in the mass spectrometer. Moreover, they can be fragmented in the instrument and the resulting spectra of fragment ion species are the basis for subsequent protein identification (for tutorial, see [3]). Although it has been discussed before [2], we need to stress again that only high-quality spectra can provide confident peptide assignments. It is important to be aware that the proteomics experiment measures proteins across an unknown concentration range. Thus, protein matches are reported from abundant proteins as well as from species measured at the detection limit of the instrument. Conclusively, the spectral basis for some proteins hits may be excellent (typically for the abundant proteins), for others, however, it may be questionable. For illustration, we show two spectra of Glu-fibrinopeptide – one of great quality showing the entire fragment ion series, and another very noisy spectrum with only few peaks hinting at this peptide. Such low-level data are observed for peptides present only at low concentrations. Modern software is much more successful in finding useful information in these spectra than the human operator, increasing sensitivity. Still, assignments based on poor spectra remain problematic and prone to false-positives, especially when they are taken for proof of unique peptides (see below).

2.3. Protein Identification

Proteomics has to battle a more complex ome than genomics, because transcriptional and translational processes lead to many more protein forms than would be expected from the genetic

blueprint. Isoforms or variants originate from single genes as a result of processes like alternative splicings and variable promoter usage [12,13]. They are highly similar in sequence, but only some isoforms may have unique functions. Although most important, those proteins may be present at comparatively low concentrations in a total proteome digest. In order to illustrate this situation, we performed a Blast analysis of human tubulin α -1A chain (Uniprot.org) and show the sequences with more than 90% identity (Supplementary Figures S3/S4). Exemplarily, three tryptic peptides have been marked, which would need to be detected with good quality in order to verify a certain isoform.

As a result of the huge complexity, in measurements of total proteome digests it is not always possible to find all the unique peptides, which would distinguish one protein form from another. Therefore, search algorithms collect all accession numbers for protein sequences, which were detected. Consequently, with this global, unbiased experimental approach, isoform analysis is severely limited. As a result of the protein grouping phenomenon (as well as always, spectral quality), we encountered absurd situations such as not finding abundant proteins, which were clearly there, and, on the other hand, finding proteins, which were definitely not present. Therefore, the user has always to be aware that, although software is a fantastic tool, it has its limitations.

Protein grouping also influences subsequent gene ontology or pathway analyses. If the protein results table contains several accession numbers for a single protein hit, it is unclear, which protein to use for further considerations. Thus, typically, the first match or a preferred hit is chosen and all others are ignored. Obviously, such analysis delivers only superficial information, because it is not carried out based on validated protein forms. Nevertheless, it may be helpful for hypothesis generation.

Of course, the quality of database available for protein identification is very important. It is high for completely sequenced species, but questionable for species with only a few entries. This is often the case for plants, insects or, in general, species, which are not as widely researched as human, mouse or rat. Scientists try to circumvent the problem by using the database of a higher-order species with more entries, but this tends to confuse the protein output by the presence of multiple forms of the same protein from different sub-species. These forms may not be easily recognizable as identical or very similar proteins as a result of different protein names. Furthermore, the true protein sequence of the species under investigation will likely be similar but not identical to that of related species. Thus, quantitation efforts in such situations are very problematic.

2.4. Modifications

Adding to the complexity of the proteome are the more than 300 known post-translational modifications like glycosylation and phosphorylation, a fact which is summarized with the term "proteoform" [14,15]. Functionally relevant modifications are often sub-stoichiometric so that they may be hard to detect reliably in total proteome analyses. It is, thus, often sensible to enrich the modified proteoforms or, alternatively, deplete the matrix. Spectra of modified peptides should contain most expected fragment ions of the b- and y-fragment ion series for confident assignment; low level data as shown in Figure 2 are not acceptable as proof for the presence of a certain peptide; they can only serve as an indication that the peptide may be potentially there and that, with enrichment, spectral quality may improve.

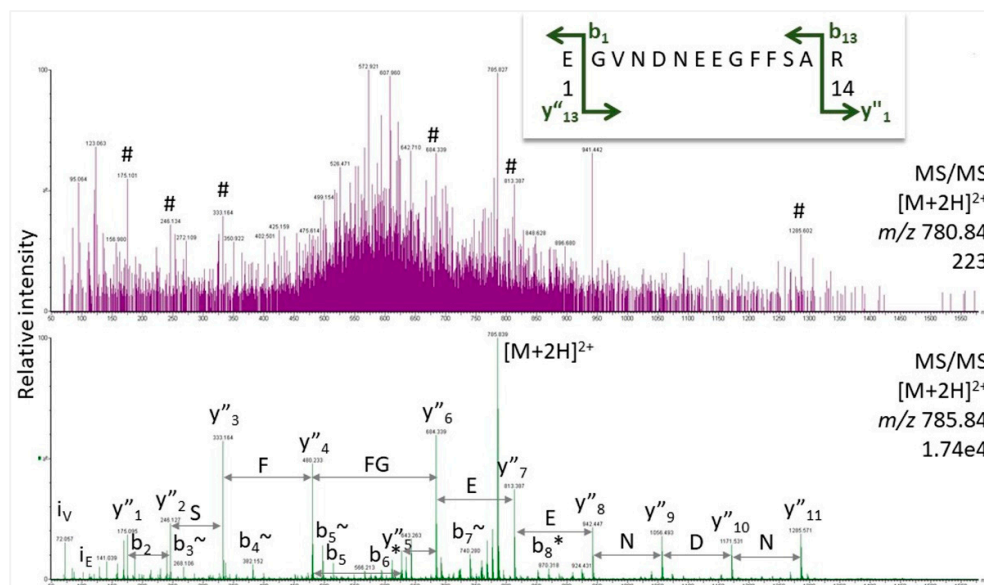


Figure 2. Gas phase fragmentation spectra for Glu-fibrinopeptide obtained with Q-TOF MS (Synapt G2 Si, Waters Corp.). In the bottom trace, the instrument was set at the proper isolation mass for the doubly-charged peptide (m/z 785.84). For the top trace, the isolation window was shifted by five mass units to simulate a weak spectrum for illustration purposes. For masses of the expected fragment ions and labels, see Supplementary Figure S2. # indicates fragment ions of Glu-fibrinopeptide at low intensity.

Attention also has to be paid to the fact that the electrospray ionization process may introduce oxidative changes on the analyte [16,17] (for an example from phosphorylation analysis, see ref. [3]); the search algorithm will not be able to distinguish those artificial modifications from endogenous ones requiring further validation.

2.5. Bioinformatics / Data Analysis Issues

Proteomics based on high-resolution MS is a highly digitalized discipline. It is mandatory to analyze the hundreds of thousands of spectra of a study in-silico. Bioinformatics steps should, however, not be misunderstood as post-experiment salvation of suboptimal study design or sample quality; they are rather inherent steps of the workflow that can deal properly only with data obtained in established standardized experiments of a guaranteed quality level.

2.5.1. Quality Assessment / Experiment Reproducibility

Proteomics research requires complex sample preparation and instrumental set-ups. Reporting of all steps of the workflow as well as quality assessment became more and more important since the mid-2000s [18,19]. During the following decade, diverse metrics and tools for quality assessment have been proposed [20–22]. In silico quality assessment of experimental data is first and foremost a relative evaluation of the parameters of the dataset (e.g., comparison of the detected number of peptide ions, charge states or peptide intensities per run). In this way, outliers within a dataset can be identified. Moreover, if a known reference sample is routinely injected every n runs into the instrument, the comparison of results with those obtained with a new or cleaned instrument may deliver information about increasing contamination of the mass spectrometer source and its internal parts or decreasing quality of the liquid chromatography (LC) column with its lifetime.

There was hope among proteomics researchers that with progress in mass spectrometer accuracy and resolution the technical reproducibility of measurements would improve. In order to test this hypothesis, we here assessed measurements obtained with the in-house prepared standard peptide sample ISA, which is routinely used in our laboratory, and which were stored with the

sample raw files of each individual study. To that end, we used a quality control (QC) tool that we are developing in order to validate study coherence. It is based on the evaluation of many characteristics of the runs. The QC tool calculates dozens of summarizing values per experiment that characterize the run (e.g., the number of MS2 spectra, or the fraction of doubly charged pre-cursors). Some of those values are connected to certain workflow steps (e.g., fractions of identified peptides with missed cleavage sites or quantiles of the total ion chromatogram (TIC) to describe TIC shapes). We applied the tool to 10 consecutive ISA runs each recorded with three different orbitrap mass spectrometers from ThermoFisher (data sets labeled as PROETD (2015), QExHF (2019) and EXII (2023)). Principal component analysis (PCA) of all the available quality features has been performed (Figure 3). The technical variance within the ISA runs of each machine is similar (see diameter of dot clouds in Figure 3). The results obtained with the three instruments differ, however, from each other, which is not surprising as their technical parameters vary. There is one outlier in the EXII (2023) runs (indicated in Figure 3 by an arrow). It originated from one run showing heavy deviation at late retention times most likely resulting from column contamination (e.g., samples containing CHAPS detergent would generate similar ion chromatograms; see Supplementary Figure S5). Moreover, the dot clouds in the PCA show a linear trend suggesting a gradual change in measurement quality due to column aging or source contamination. Furthermore, technical reproducibility within a data set does not seem to have improved over the last eight years. The progress in mass spectrometer parameters has not considerably impacted on data variability, which is not surprising, because a lot of the experimental variation originates from pre-analytical factors (see above), or, as in the case shown here, from contamination.

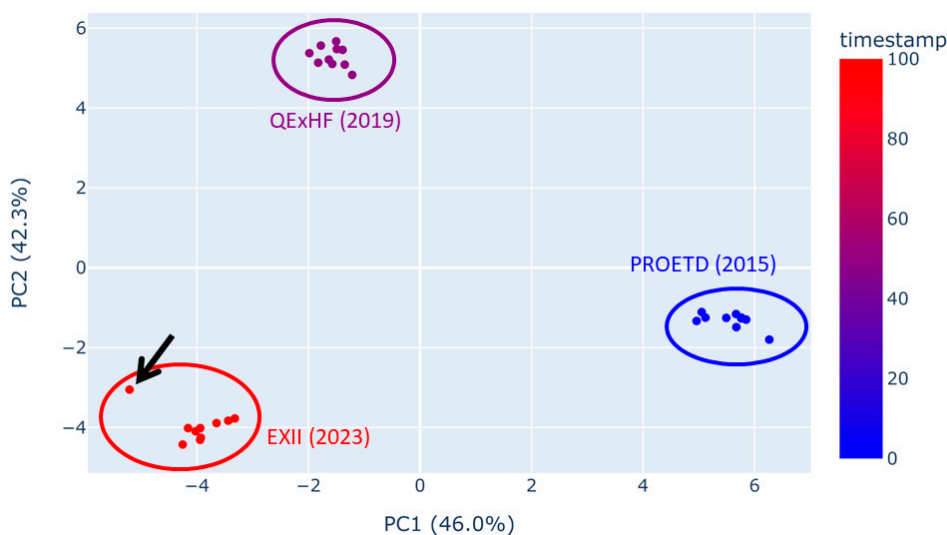


Figure 3. Scatter plot of the first two principal components (PCA-plot) calculated on the QC metrics of the raw data (prior to peptide identification). Each data point represents one raw file. The colour gradient is determined by the timestamp, here also indicating the year and machine. The arrow points to an outlier in the EXII raw files.

2.5.2. False-Discovery Estimation With Decoy Sequences

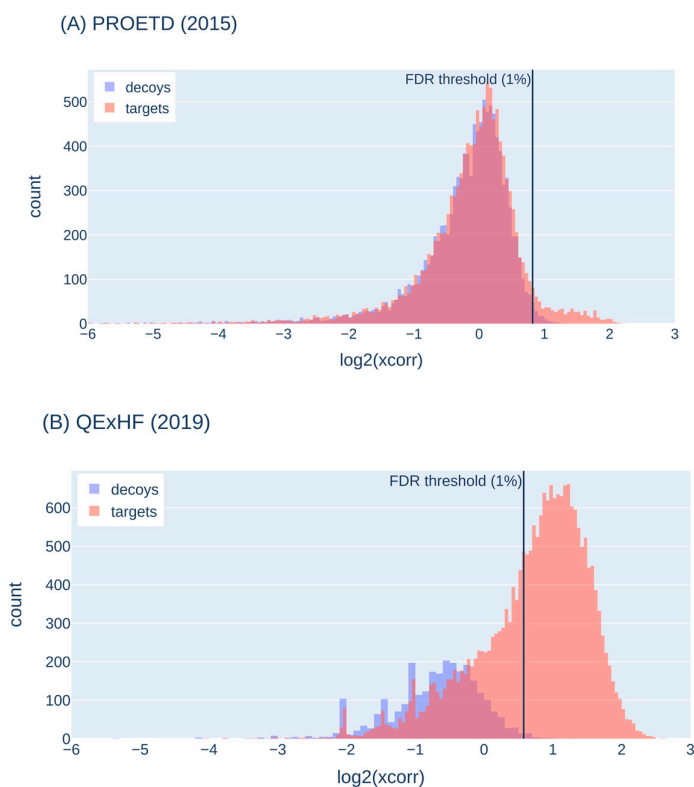
When after the Human Genome Project manual spectrum identification was widely replaced by sequence database-driven spectrum identification, it was instantly clear that a race for the number of identified peptide-spectrum-matches (PSMs) alone was not sufficient or appropriate. For each pair of measured spectrum and theoretical spectrum of a tryptic database peptide, a score was calculated. Thresholding the good from the bad scores was mandatory; it can still be done score system-agnostic with a target-decoy approach [23,24] and a local false discovery rate (FDR) estimation at each position of the score-sorted peptide-spectrum results list. This method is useful and part of state-of-the-art

software suites, albeit not aggressively promoted. Moreover, target-decoy implementations face challenges, when they fail to generate a sufficient number of decoys to account for the highly accurate precursor masses [25], or when they do not meet other crucial assumptions such as similar properties of decoy PSMs and incorrect target PSMs [26].

We visualized the target-decoy behavior on representative runs of the three ISA samples to characterize the target-decoy-behavior in our experiments (Figure 4). The decoy PSMs (with their score) represent false-positive target matches around that score, so that we can distinguish true from false-positive target scores. Usually, a FDR threshold of 1% is applied, i.e. all target scores higher than this threshold are accepted (shown as a black vertical line in Figure 4).

In the ISA data set PROETD (2015), the false-positive peak is rather high (targets and decoys nearly perfectly overlaid) and vastly exceeds the peak of true positives (between 1 and 2 on the x axis). In 2023, the peak for true positives (target scores) is higher than in 2019 and higher than the peak for the false-positive targets (peak to the left of the threshold). The peak for the false-positive targets is accompanied by a peak for the false-positive decoys, but they do not perfectly overlay, which hints on problems with the decoy assumption (the number of decoy sequences offered to low-quality spectra may not be high enough). The target-decoy histograms of QExHF (2018) indicate even larger problems with the decoy assumption.

In summary, the decoy approach is helpful in separating good from poor matches. With the most recent machines, more true-positives than with older instruments and more true-positives than false-positives are identified, so that the disadvantage of having not a perfect separation between good and poor matches is partially compensated. The observed behavior also allows to use a more stringent FDR (0.01, 0.001 or even lower) without losing to many true-positive hits. Nevertheless, bioinformatics groups are already working on addressing the false-positive detection for state-of-the-art instruments, e.g., by spectrum-specific decoy generation [26].



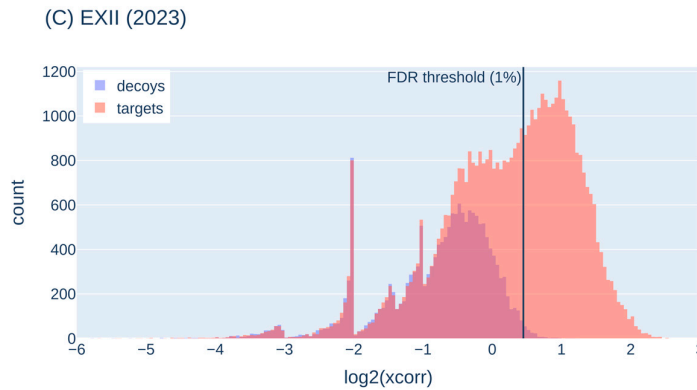


Figure 4. Score histograms for decoy and target PSMs; representative runs selected from PROETD (2015), QExHF (2019) and EXII (2023). The x-axis was cut at -6. The vertical line shows the score threshold for 1% FDR calculated from the respective decoy score distributions.

3. Materials and Methods

3.1. Description of ISA Files

At the Medizinisches Proteom-Center (MPC), ISA is used for regular QC and the monitoring of the performance of the LC-MS set-up. The generation of ISA samples is described in Supplementary Methods 1 in detail [27–29]. The sample consists of A549 cell culture with six spiked-in peptides. After 15–20 analyses, one ISA is measured for QC; between all runs + ISA, blanks are run to reduce take-over.

Since 2015, the MPC archives all measured raw files. We used files from the most popular mass spectrometers in the years 2015, 2019 and 2023 (Table 1), and compared the results of ten ISA runs, each of which followed at least ten other analyses. We avoided selecting files from times of instrument repair or calibration, when ISAs were measured more frequently. We chose consecutive raw files, meaning that no other ISA samples were measured in between the chosen raw files. Details on the MS analysis of the ISA samples are given in Supplementary Methods 2 [30].

Table 1. MPC mass spectrometers used for this study.

Year	2015	2019	2023
Instrument code	PROETD	QExHF	EXII
Nano electrospray mass spectrometer (ThermoFisher)	Velos Plus	Q Exactive HF	Orbitrap Exploris 480
Analyzer	radial ejection linear ion trap	quadrupole	quadrupole
Detector	electron multiplier	inductive detector	inductive detector
Number of raw files	3172	1930	3287

3.2. QC Tool

For calculating and visualization of the quality measures of the different raw files, we used an updated version of the MaCProQC tool programmed by us [22]. The workflow was re-written in Python 3.8 [31] and is executed via Nextflow [32]. The tool calculates different QC metrics of which many are based on the paper by Bittremieux et al. [33]. In contrast to the earlier version of the tool [22], the new workflow uses Comet version v2022.01.0 [34] as a search engine instead of Mascot. For feature finding, the FeatureFinderCentroided from OpenMS version 2.8.0 [35] is used. In brief, the workflow first extracts quality metrics directly from the raw files (e.g., number of MS1 and MS2 spectra). Different quartiles are calculated based on the TIC or number of MS events [33]. Then, a

peptide identification with Comet is performed as well as protein inference by PIA version 1.4.8 [36,37]. This results in metrics like the number of PSMs, peptides and protein groups, as well as the distribution of charge states and missed cleavages for the PSMs. The feature detection then reports the number of features and identified features. These metrics are subsequently used for PCA and visualized using the plotly python package (Plotly Technologies Inc.) [38]. In Comet, two missed cleavages were allowed and the five best hits were reported. As variable modification, oxidation of methionine, and as fixed modification, carbamidomethylation of cysteine, were set. The tolerance values were chosen depending on the instrument as given in Table 2. For feature finding, a mass trace m/z -tolerance of 0.02 and a isotopic pattern m/z -tolerance of 0.04 was used for PROETD (low resolution settings) and mass trace m/z -tolerance of 0.004 and a isotopic pattern m/z -tolerance of 0.005 for EXII and QExHF (high resolution settings).

Table 2. Search parameters used in the Comet search engine for the different mass spectrometers.

Project code	Tolerance		
	Peptide mass	Fragment bin	Fragment bin offset
PROETD (2015)	10 ppm	1.0005	0.4
QExHF (2019)	10 ppm	0.02	0
EXII (2023)	5 ppm	0.02	0

3.3. Description of Decoy Workflow

The 30 raw files were searched with the Comet search engine, version v2022.01.0 [34]. The FASTA database was 20230223_human_proteome_with_spikeins.fasta (in total 81,843 entries), which contained the human reference proteome version 2022_05 from UniProt as well as the six spike-in peptides (GEPAAAAPEAGASPVEK, NLVVGDETTSSLR, LQPGDIGIYR, VVVLPSGALQISR, YPGAYYIFQIK, NIPTVNENLENYYLEVNQLEK). Decoys were generated by reversing the original protein sequences within Comet. Only the best hit was reported by Comet, all other settings were kept the same. The text output file from Comet was further processed with Python scripts. A PSM was counted as a target, if it matched with at least one target sequence. If all matched sequences were decoys, it was counted as a decoy. The $xcorr$ scores from comet were \log_2 -transformed, plotted in histograms and coloured by type (targets or decoys) to compare the score distributions between targets and decoys using plotly [38].

4. Conclusions

Many challenges in the field of bioinformatics for proteomics are due to the high-throughput nature of the experiments. Unambiguous protein assignment is complicated by shared peptides [39,40]. Peptide measurement and subsequently protein quantification is influenced by suppression effects from the analytical matrix as well as the physico-chemical properties of the individual peptides, namely their ionization efficiency. For an accurate assessment of identification and quantification algorithms, there is a strong need for a standard peptide sample [41] with thousands of potentially shared tryptic peptides in known quantities. However, it is a challenge to experimentally express proteins in correctly known amounts. Even with the use of peptide synthesizers, there are several practical hurdles to generate properly characterized reference samples. Once a highly complex peptide references standard becomes available, sample concentrations could be directly calculated from detector signals by measurement-specific "correction coefficients". It would help enormously to assess not only the decoy, but also the quantitative algorithms. Despite continuous advances in mass spectrometer technology as well as analysis software, there is still need for data set-wise QC. With a continuously measured peptide mixture, QC can go beyond one data set. Work has already begun on the stability of proteotypic peptides [42] used for quantification.

False-discovery detection and avoidance is also necessary. The traditional methods are working, but bioinformatics groups are developing better suited approaches such as spectra-wise decoys.

The best bioinformatician cannot rescue poor experiments. Our runs with reference samples demonstrate that sample preparation remains the major factor in measurement variability and cannot be compensated by advanced high technology. Standardisation of pre-analytical procedures and proper experimental design remain the most important factors in omics studies.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org. Supplementary Figures and experimental details of reference measurements: Supplement.doc.

Author Contributions: Conceptualization, S.K. and M.E.; methodology, S.K., K.S. and M.E.; software, K.S. and M.E.; writing—original draft preparation, S.K., K.S. and M.E.; writing—review and editing, S.K. and M.E.; visualization, S.K. and K.S. All authors have read and agreed to the published version of the manuscript.

Funding: Please add: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from martin.eisenacher@rub.de. They will also be made available in PRIDE.

Acknowledgments: The authors thank the collaborators in the projects used as examples (all co-authors and M. Bayer, formerly CUP). The authors thank K. Marcus-Alic for providing the data and K. Fuchs, N. Keil and C. Bunse for measurement of the ISA samples. The authors are also grateful to S. Rozanova and B. Eggers for providing experimental information regarding the ISA standard and MS measurements and D. Lux for assisting with the Comet searches.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Duncan, M. Good mass spectrometry and its place in good science. *J. Mass Spectrom.* **2012**, *47*, 795-809.
2. Coorsen, J.R.; Yergey, A.L. Proteomics is analytical chemistry: Fitness-for-purpose in the application of top-down and bottom-up analyses. *Proteomes* **2015**, *3*, 440-453.
3. König, S. Spectral quality overrides software score—A brief tutorial on the analysis of peptide fragmentation data for mass spectrometry laymen. *J. Mass Spectrom.* **2021**, *56*, e4616.
4. Gegner, H.M.; Naake, T.; Dugourd, A.; Müller, T.; Czernilofsky, F.; Kliewer, G.; Jäger, E.; Helm, B.; Kunze-Rohrbach, N.; Klingmüller, U.; Hopf, C.; Müller-Tidow, C.; Dietrich, S.; Saez-Rodriguez, J.; Huber, W.; Hell, R.; Poschet, G.; Krijgsveld, J. Pre-analytical processing of plasma and serum samples for combined proteome and metabolome analysis. *Front. Mol. Biosci.* **2022**, *9*, 961448.
5. West, J.; Atherton, J.; Costelloe, S.J.; Pourmahram, G.; Stretton, A.; Cornes, M. Preanalytical errors in medical laboratories: a review of the available methodologies of data collection and analysis. *Ann. Clin. Biochem.* **2017**, *54*(1): 14-19.
6. Hassis, M.E.; Niles, R.K.; Braten, M.N.; Albertolle, M.E.; Witkowska, E.H.; Hubel, C.A.; Fisher, S.J.; Williams, K.E. Evaluating the effects of preanalytical variables on the stability of the human plasma proteome. *Anal. Biochem.* **2015**, *478*: 14-22.
7. Bayer, M.; König, S. A vote for robustness: Monitoring serum enzyme activity by thin-layer chromatography of dabsylated bradykinin products. *J. Pharmaceut. Biomed. Anal.* **2017**, *143*, 199-203.
8. CDC hemolysis palette, <https://www.cdc.gov/nceid/dvbd/specimensub/hemolysis-palette.html>, accessed Sept. 9, 2023.
9. Aziz, S.; Rasheed, F.; Zahra, R.; König, S. Gastric cancer pre-stage detection and early diagnosis of gastritis using serum protein signatures. *Molecules* **2022**, *27*: 2857.
10. Ni, X.; Tan, Z.; Ding, C.; Zhang, C.; Song, L.; Yang, S.; Liu, M.; Jia, R.; Zhao, C.; Song, L.; et al. A region-resolved mucosa proteome of the human stomach. *Nat. Commun.* **2019**, *10*, 39.
11. Tepasse, P.-R.; Vollenberg, R.; Steinebrey, N.; König, S. The dysregulation of the renin-angiotensin-system in COVID-19 studied by serum proteomics: Angiotensinogen increases with disease severity. *Molecules* **2022**, *27*, 2495.
12. Brett, D.; Pospisil, H.; Valcárcel, J.; Reich, J.; Bork, P. Alternative splicing and genome complexity. *Nature Genetics* **2002**, *30*(1): 29-30.
13. Liu, Q.; Fang, L.; Wu, C. Alternative splicing and isoforms: From mechanisms to diseases. *Genes* **2022**, *13*(3): 401.
14. Smith, L.M.; Kelleher, N.L. Proteoforms as the next proteomics currency. *Science* **2018**, *359*(6380): 1106-1107.

15. Aebersold, R.; Agar, J.N.; Amster, I.J.; Baker, M.S.; Bertozzi, C.R.; Boja, E.S., et al. How many human proteoforms are there? *Nature Chemical Biology* **2018**, *14*(3): 206–214.
16. Chen, M.; Cook, K.D. Oxidation artifacts in the electrospray mass spectrometry of Abeta peptide. *Anal. Chem.* **2007**, *79*, 2031-2036.
17. Morand, K.; Talbo, G.; Mann, M. Oxidation of peptides during electrospray ionization. *Rapid Commun. Mass Spectrom.* **1993**, *7*, 738-743.
18. Taylor, C.F.; Paton, N.W.; Lilley, K.S.; Binz, P.A.; Julian, R.K. Jr; Jones, A.R.; Zhu, W.; Apweiler, R.; Aebersold, R.; Deutsch, E.W.; Dunn, M.J.; Heck, A.J.; Leitner, A.; Macht, M.; Mann, M.; Martens, L.; Neubert, T.A.; Patterson, S.D.; Ping, P.; Seymour, S.L.; Souda, P.; Tsugita, A.; Vandekerckhove, J.; Vondriska, T.M.; Whitelegge, J.P.; Wilkins, M.R.; Xenarios, I.; Yates, J.R. 3rd; Hermjakob, H. The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol.* **2007**, *25*(8), 887-893.
19. Walzer, M.; Pernas, L.E.; Nasso, S.; Bittremieux, W.; Nahnsen, S.; Kelchtermans, P.; Pichler, P.; van den Toorn, H.W.; Staes, A.; Vandenbussche, J.; Mazanek, M.; Taus, T.; Scheltema, R.A.; Kelstrup, C.D.; Gatto, L.; van Breukelen, B.; Aiche, S.; Valkenborg, D.; Laukens, K.; Lilley, K.S.; Olsen, J.V.; Heck, A.J.; Mechtler, K.; Aebersold, R.; Gevaert, K.; Vizcaino, J.A.; Hermjakob, H.; Kohlbacher, O.; Martens, L. qcML: an exchange format for quality control metrics from mass spectrometry experiments. *Mol Cell Proteomics.* **2014**, *13*(8), 1905-1913.
20. Pichler P.; Mazanek M.; Dusberger F.; Weillböck L.; Huber C.G.; Stingl C.; Luider T.M.; Straube W.L.; Köcher T.; Mechtler K. SIMPATIQCO: a server-based software suite which facilitates monitoring the time course of LC-MS performance metrics on Orbitrap instruments. *J Proteome Res.* **2012** *11*(11), 5540-5547.
21. Rudnick, P.A.; Clauser, K.R.; Kilpatrick, L.E.; Tchekhovskoi, D.V.; Neta, P.; Blonder, N.; Billheimer, D.D.; Blackman, R.K.; Bunk, D.M.; Cardasis, H.L.; Ham, A.J.; Jaffe, J.D.; Kinsinger, C.R.; Mesri, M.; Neubert, T.A.; Schilling, B.; Tabb, D.L.; Tegeler, T.J.; Vega-Montoto, L.; Variyath, A.M.; Wang, M.; Wang, P.; Whiteaker, J.R.; Zimmerman, L.J.; Carr, S.A.; Fisher, S.J.; Gibson, B.W.; Paulovich, A.G.; Regnier, F.E.; Rodriguez, H.; Spiegelman, C.; Tempst, P.; Liebler, D.C.; Stein, S.E. Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Mol Cell Proteomics.* **2010**, *9*(2), 225-241.
22. Rozanova, S.; Uszkoreit, J.; Schork, K.; Serschnitzki, B.; Eisenacher, M.; Tönges, L.; Barkovits-Boeddinghaus, K.; Marcus, K. Quality control - A stepchild in quantitative proteomics: A case study for the human CSF proteome. *Biomolecules.* **2023**, *13*(3), 491.
23. Elias, J.E.; Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **2007**, *4*(3), 207-214.
24. Reidegeld, K.A.; Eisenacher, M.; Kohl, M.; Chamrad, D.; Körting, G.; Blüggel, M.; Meyer, H.E.; Stephan, C. An easy-to-use decoy database builder software tool, implementing different decoy strategies for false discovery rate calculation in automated MS/MS protein identifications. *Proteomics* **2008**, *8*(6), 1129-1137.
25. Keich, U., Noble, W.S. On the importance of well-calibrated scores for identifying shotgun proteomics spectra. *J Proteome Res.* **2015**, *14*(2), 1147-1160.
26. Debrie, E.; Malfait, M.; Gabriels, R.; Declercq, A.; Sticker, A.; Martens, L.; Clement, L. Quality control for the target decoy approach for peptide identification. *J Proteome Res.* **2023**, *22*(2), 350-358.
27. Bradford, M.M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem.* **1976**, *72*, 248-254.
28. Li, Y., Wang, Y.; Mao, J.; Yao, Y.; Wang, K.; Qiao, Q.; Fang, Z.; Ye, M. Sensitive profiling of cell surface proteome by using an optimized biotinylation method. *J Proteomics.* **2019**, *196*, 33-41.
29. May, C.; Serschnitzki, B.; Marcus, K. Good old-fashioned protein concentration determination by amino acid analysis. *Methods Mol Biol.* **2021**, *2228*, 21-28.
30. Kley, R.A.; Maerkens, A.; Leber, Y.; Theis, V.; Schreiner, A.; van der Ven, P.F.; Uszkoreit, J.; Stephan, C.; Eulitz, S.; Euler, N.; Kirschner, J.; Müller, K.; Meyer, H.E.; Tegenthoff, M.; Fürst, D.O.; Vorgerd, M.; Müller, T., Marcus, K. A combined laser microdissection and mass spectrometry approach reveals new disease relevant proteins accumulating in aggregates of filaminopathy patients. *Mol Cell Proteomics.* **2013**, *12*(1), 215-227.
31. Van Rossum, G., Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.
32. Di Tommaso, P.; Chatzou, M.; Floden, E.W.; et al. Nextflow enables reproducible computational workflows. *Nature Biotechnology.* **2017**, *35*(4), 316-319.
33. Bittremieux, W.; Meysman, P.; Martens, L.; Valkenborg, D.; Laukens, K. Unsupervised quality assessment of mass spectrometry proteomics experiments by multivariate quality control metrics. *J Proteome Research.* **2016**, *15*(4), 1300-1307.
34. Eng, J.K.; Jahan, T.A.; Hoopmann, M.R. Comet: an open-source MS/MS sequence database search tool. *Proteomics.* **2013** *13*(1), 22-24.
35. Weisser, H.; Nahnsen, S.; Grossmann, J.; et al. An automated pipeline for high-throughput label-free quantitative proteomics. *J Proteome Research.* **2013**, *12*(4), 1628-1644.

36. Uszkoreit, J.; Maerkens, A.; Perez-Riverol, Y.; Meyer, H.E.; Marcus, K.; Stephan, C.; Kohlbacher, O.; Eisenacher M. PIA: An intuitive protein inference engine with a web-based user interface. *J Proteome Res.* **2015**, *14*(7), 2988-2997.
37. Uszkoreit, J.; Perez-Riverol, Y.; Eggers, B.; Marcus, K.; Eisenacher, M. Protein inference using PIA workflows and PSI standard file formats. *J Proteome Res.* **2019**, *18*(2), 741-747.
38. Plotly Technologies Inc. Collaborative data science. Montréal, QC, 2015. <https://plot.ly>.
39. Dost, B.; Bandeira, N.; Li, X.; Shen, Z.; Briggs, S.P.; Bafna, V. Accurate mass spectrometry based protein quantification via shared peptides. *J Comput Biol.* **2012**, *19*(4), 337-348.
40. Schork, K.; Turewicz, M.; Uszkoreit, J.; Rahnenführer, J.; Eisenacher M. Characterization of peptide-protein relationships in protein ambiguity groups via bipartite graphs. *PLoS One* **2022**, *17*(10), e0276401.
41. The, M.; Edfors F.; Perez-Riverol, Y.; Payne, S.H.; Hoopmann, M.R.; Palmblad, M.; Forsström, B.; Käll, L. A protein standard that emulates homology for the characterization of protein inference algorithms. *J Proteome Res.* **2018**, *17*(5), 1879-1886.
42. Chiva, C.; Elhamraoui, Z.; Solé, A.; Serret, M.; Wilhelm, M.; Sabidó, E. Assessment and prediction of human proteotypic peptide stability for proteomics quantification. *Anal Chem.* **2023**, *95*(37), 13746-13749.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.