In-Depth Benchmarking of DIA-type Proteomics Data Analysis Strategies Using a Large-Scale Benchmark Dataset Comprising Inter-Patient Heterogeneity

Klemens Fröhlich
Institute of Surgical Pathology
Eva Brombacher
University of Freiburg
Matthias Fahrner
University of Freiburg
Daniel Vogele
University of Freiburg
Lucas Kook
University of Zurich
Peter Bronsert
Faculty of Medicine, Institute for Surgical Pathology, Medical Center, University of Freiburg, Freiburg, https://orcid.org/0000-0001-8558-0347
Sylvia Timme
University of Freiburg
Alexander Schmidt
Biozentrum, University of Basel, Switzerland https://orcid.org/0000-0002-3149-2381
Katja Baerenfaller
Swiss Institute of Allergy and Asthma Research https://orcid.org/0000-0002-1904-9440
Clemens Kreutz
University of Freiburg https://orcid.org/0000-0002-8796-5766
Oliver Schilling (✉ oliver.schilling@uniklinik-freiburg.de)
Institute of Surgical Pathology

Article

Keywords: benchmarking, biomarkers, DIA-NN, heterogeneity

Posted Date: September 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-893982/v1

Version of Record: A version of this preprint was published at Nature Communications on May 12th, 2022. See the published version at https://doi.org/10.1038/s41467-022-30094-0.

Abstract

An overwhelming number of proteomics software tools and algorithms have been published for different steps of Data Independent Acquisition analysis of clinical samples. Nonetheless, there is still a lack of comprehensive benchmark studies evaluating which combinations of those isolated components perform best.

Here, we used 92 lymph nodes from distinct patients to create a unique benchmark dataset representing real-world inter-individual heterogeneity. The publicly available dataset comprises 118 LC-MS/MS runs with > 12 million MS2 spectra and allowed us to objectively evaluate how well different combinations of spectral libraries, DIA software, sparsity reduction, normalization and statistical tests can detect differentially abundant proteins, while also taking sample size into account.

Evaluation of 2 million data analysis workflows showed that a gas phase fractionation refined spectral library in combination with DIA-NN and Significance Analysis of Microarrays reliably detected differentially abundant proteins. Furthermore, DIA-NN and Spectronaut robustly avoided the false detection of truly absent proteins.

*KF and EB share first authorship. CK and OS share last authorship.

Introduction

Proteomics denotes the study of the entire set of proteins produced by an organism under defined conditions. While the genome of an organism is geared towards remaining static for almost every cell, the dynamics introduced by the proteome, including differential expression, altered activity, and modifications of proteins, allows cells, tissues and even the whole organism to undergo dramatic changes and to carry out a plethora of different functions. Often, the term ‘proteomics’ is specifically used to refer to large-scale studies of the proteome employing mass spectrometry (MS) and liquid chromatography (LC) coupled to mass spectrometry (LC-MS/MS).

Many studies, e.g. in the clinical context, focus on the detection of differentially abundant proteins; preferably on a proteome-wide scale. To identify such proteins, modern mass spectrometry-based proteomics techniques offer many ways to quantify and compare proteins between samples. Due to their simplicity and cost-effectiveness, label-free approaches have been used for decades. Historically, label-free samples were measured using data-dependent acquisition (DDA). In DDA, following a survey scan, masses of interest are selected for further fragmentation based on their intensity. This allows for narrow isolation windows and results in fragment spectra of low complexity. However, the fact that the masses of interest are selected during the measurement introduces stochastic sampling effects.

In contrast, parallel fragmentation of precursor ions implemented in data-independent acquisition (DIA) methods is independent of ion intensity and other properties, leading to constant data acquisition between samples. In DIA proteomics, assignment of fragment ions to a single analyte (i.e. peptide) happens post-measurement and is strongly dependent on properties of reference spectral libraries as well as on features of the data processing algorithms and tools. In DIA proteomics, quantification is typically performed on the fragment level and not on the precursor level as in the case of DDA [1].

While this increased complexity makes the handling and analysis of DIA data more laborious, it has been demonstrated that the quantification by DIA is more robust compared to DDA. Recently, DIA has reached a protein coverage that is comparable to, or even exceeds, the one of DDA [2,3].

To objectively compare data processing and quantitation methods, the proteomics community often employs so-called spike-in benchmark datasets, in which peptides with known properties (e.g. sequence and concentration) are added to 'background' peptides with likewise known properties. To mimic the complexity encountered in realistic settings often different organisms are added in combination to create benchmark datasets [4]. These benchmark datasets are valuable tools for controlling and optimizing different aspects of data acquisition and analysis, including LC-MS/MS parameters, library generation, analysis software parameters, data preprocessing and statistical analysis for detecting differentially abundant proteins. The critical importance of data processing in DIA proteomics renders benchmarking datasets particularly useful for this methodology.

Benchmark studies published to date have mainly focused on technical reproducibility and data acquisition [5], or on data analysis steps, such as data preprocessing in the form of data normalization or data imputation, and statistical methods [6]. Indeed, the downstream analysis of the data acquired by DIA software suites should be carefully reflected upon, going beyond peptide-spectrum-matching (PSM) and quantitative signal/feature integration. Furthermore, valid benchmarking datasets should represent inter-individual heterogeneity on a scale that is comparable to present-day, cohort-wide proteome studies, which is mostly not the case.

Hence, especially in biomarker discovery studies in which highly heterogeneous patient proteomes are investigated, current benchmark datasets provide little help for a user who has to decide whether and how to generate a library for DIA analysis, which tool to use for the DIA analysis itself, and, most importantly, how this will affect data preprocessing and statistical analysis for differentially abundant proteins.

We set out to create a large benchmark dataset reflecting real inter-individual heterogeneity to investigate the interplay between library generation, DIA software analysis, data preprocessing, and statistical analysis. To this end, we acquired sentinel lymph nodes from multiple patients as formalin-fixed paraffin-embedded tissue (FFPE) and used *Escherichia coli* (*E. coli)* as a spike-in peptide subpopulation of known concentration. For the DIA measurements a previously described, well established acquisition scheme was employed [7].

Three trends can be observed in current DIA analysis strategies: a) using spectral libraries generated by analysing pre-fractionated DDA runs, b) using spectral libraries generated by refining predicted libraries using gas-phase-fractionation (GPF) or c) using no additional experimental data to generate spectral libraries (e.g. using predicted libraries). All these prototypical approaches are integrated in this study.

For the main DIA analysis, we used four of the most widely applied DIA analysis software tools in the proteomics community: DIA-NN [8], OpenSwath [9], Skyline [10] and Spectronaut [11].

Combining spectral library analysis approaches with DIA data analysis software led to 17 different 'DIA workflows'. We used 'standard' parameters, opting for the recommended settings whenever possible, to reflect a realistic average user scenario and to prevent over-optimization on one dataset. The resulting different DIA workflow datasets were then combined with data analysis workflows combining bootstrapping with three sparsity reduction methods, four normalization methods and five statistical test options resulting in a total of over 2 million analyses.

This allowed us to investigate how library generation and choice of DIA software affect data properties, and how preprocessing and statistical analysis methods affect the identification of differentially expressed proteins by means of objective evaluation measures based on p-values and log2 fold-changes that resulted from each analysis workflow.

Results And Discussion

Significance of Benchmark Studies for the Field of Proteomics

Benchmark studies have become an invaluable tool to objectively assess the advantages and disadvantages of the choices made over the course of proteomics studies, including the choice of sample preparation, data acquisition, MS data analysis and statistical processing.

However, benchmark studies often suffer from small sample sizes and unrealistically low background variance [5,12,13].

Biomarker discovery studies often include hundreds of patients with heterogeneous proteomes and can contain high within- and between-person variance. In this setting, data characteristics and the resulting performance of data analysis workflows cannot be estimated using standard benchmark datasets.

Here, we set out to empirically investigate the performance of complete multi-step data analysis workflows for DIA-type proteomics, composed of library generation methods, DIA analysis software suites, data preprocessing, and statistical analyses.

Overview of LC-MS/MS Measurements and Initial Results

We obtained 92 FFPE-embedded tumor free lymph node tissue specimens derived from patients with primary acinary prostate cancer. Following protein digestion and sample clean-up, we split the samples into four groups and added *E. coli* peptides in human : *E. coli* peptide ratios of 6:1, 12:1, and 25:1, or did not add *E. coli* peptides at all. Those four groups are referred to as 'spike-in conditions' and have a size of n=23 each (Figure 1). The resulting set of DIA LC-MS/MS measurements (92 LC-MS/MS files) consists of 12.4 million MS2 spectra.

To generate experiment-specific spectral libraries, we performed GPF on a mastermix, which represents an average spike-in concentration of human to *E. coli* peptides of 15:1. Using DIA-NN to refine a combined human and *E. coli in silico* predicted DIA-NN spectral library, we generated a spectral library containing 84016 precursor entries mapping to 10459 proteins. Using an *in silico* predicted PROSIT spectral library refined by EncyclopeDIA, we generated a spectral library containing 45445 precursors mapping to 8472 proteins [14].

We also pre-fractionationated a mastermix to obtain samples for in-depth DDA library generation [15]. Using Fragpipe to generate a spectral library from these DDA files, we generated a spectral library containing 81409 precursors mapping to 7781 proteins. We also used MaxQuant to build a DDA-based spectral library containing 51260 precursors mapping to 7382 proteins.

Using DIA-NN in combination with the *in silico* predicted DIA-NN GPF-refined spectral library, on average 48,698 precursors were identified per measurement with an average chromatographic peak width of 8 seconds (full width at half height). No batch effects originating from sample preparation or order of measurement were apparent in this analysis (Supplementary Figure S1).

-

Assessment of data analysis workflows for DIA-type proteomics

We chose four commonly used DIA software analysis suites: DIA-NN, Skyline, OpenSwath, and Spectronaut. Whenever possible, we combined all generated libraries with all DIA analysis software solutions (especially the predicted spectral libraries pose a challenge to some software suites in combination with the high number of samples). We also included 'DirectDIA', a feature of Spectronaut, which does not require any additional experimental evidence for library generation. This resulted in a total of 17 different DIA analysis workflows. For all subsequent analysis steps, protein-level output from the DIA analysis workflows was used. For the sake of simplicity we focused our statistical analyses on the comparison between the two lowest *E. coli* spike-in conditions (Figure 1). This also represents the greatest challenge to any DIA analysis software as quantitations are usually less precise for lowly abundant proteins [16,17].

In this study, we used bootstrapping (see below) to investigate the effect of sample size, normalization, sparsity reduction and choice of a statistical test on the overall ability of the data analysis workflow to detect differentially abundant proteins.

In brief, we randomly drew samples from each of the two lowest *E. coli* spike-in conditions with group sizes of three to 23 samples. On each bootstrap dataset we applied different data analysis workflows composed of multiple options for the preprocessing steps in the form of sparsity reduction and normalization, followed by one of five statistical tests to identify differentially abundant proteins.

Taking into account the aforementioned 17 different types of LC-MS/MS data processing, we acquired prediction performance information for 1020 different analysis workflows, each of which was applied to 2100 bootstrap datasets resulting in over 2 million analyses.

This staggering number beautifully illustrates the amount of possible combinations of library generation methods, DIA software suites, and downstream data preprocessing and statistical analysis methods proteomics scientists are confronted with. As every study is different and there are no truly universally applicable methods available in proteomics, the level of experience and choices of the proteomics data analyst determine the reliability and reproducibility of a proteomics study, which was impressively demonstrated by Choi et al [13].

Analyses of the LC-MS/MS Data

We first assessed the number of identified and quantified proteins with a 1 % protein FDR cutoff being applied to all workflows (Figure 2 left panel). As the number of identified proteins depends on the number of proteins, which are physically present in a sample, the samples are grouped by spike-in condition. The DDA spectral libraries consistently led to smaller numbers of identified proteins. As the tissue used in this study was formalin-fixed, chemical modifications can reduce the number of identified peptides and proteins during spectral library generation. In our experience, GPF refined spectral libraries often lead to higher identification rates in DIA-type proteomics data of FFPE tissue. The total number of identifiable proteins increases with increasing amounts of spiked-in *E. coli* proteins. Using a GPF-refined *in silico* predicted PROSIT spectral library in combination with Skyline yielded the highest number of quantified proteins, ranging from a mean value of 7388 proteins for samples without spike-in to a median of 7480 proteins for the samples with a human: *E. coli* spike-in of 6:1 (w/w).

However, in quantitative proteomics, protein identifications only serve a useful purpose if they are accompanied with robust and reliable quantitation. When summarizing the protein abundances as calculated by the different DIA analysis workflows both the shape of the distribution of log-transformed protein abundances (Figure 2 center panel) as well as the correlation of log-transformed intensities between DIA analysis workflows mostly depend on the choice of DIA analysis software, and to a lesser extent on the spectral library (Supplementary Figure S2).

Further, we determined the variance of individual *E. coli* protein intensities per spike-in ratio. Since one single batch of *E. coli* was used for all spike-ins (thus reducing inter-sample variability between *E. coli* proteins), a minimal variance indicates a reproducible quantitation algorithm. While the absolute variance of protein intensities is similar across all DIA analysis workflows, the DIA workflows differ in how much this variance varies across each spike-in condition: for DIA-NN and Spectronaut DIA workflows, the

variability of variances decreases with higher *E. coli* spike-in concentrations (except for DIA-NN in combination with the refined PROSIT spectral library), while the opposite behaviour can be observed for OpenSwath and Skyline (with the exception of human to *E. coli* spike-in condition 6:1 (Figure 2 right panel). Therefore, we specifically investigated the reported quantitations for *E. coli* proteins (Supplementary Figure S3). While the number of identified and quantified *E. coli* proteins decrease in lower spike-in conditions, it remains constant for some DIA analysis software suites. This led us to more closely investigate how different DIA analysis suites report missing proteins.

Missing Values and False-Positive Quantitation

DIA-type proteomics promises to reduce missingness in multi-sample proteomic experiments [11]. In the present dataset, 25% of all samples are human-only and void of *E. coli* proteins. This experimental setting not only supports the illustration of missingness, but also the illustration of false-positive quantitation of proteins (here: *E. coli* proteins).

As can be appreciated from Figure 3A, the means of the human and *E. coli* protein intensities correlate negatively with the percentage of missing values per protein for all DIA software suites except for Skyline. This negative correlation has also been reported previously in a clinical proteomics study employing a tripleTOF instrument using OpenSwath for data analysis [18]. Furthermore, DIA-NN and Spectronaut correctly yield a level of 25 % missingness for most *E. coli* proteins, while Skyline and OpenSwath do not clearly reach this distinction. The nature of the spectral library only had a negligible impact in this regard in most cases. However, for DIA-NN the number of reported *E. coli* proteins for samples, which only contained human lymph node proteins increased when using the EncyclopeDIA-refined *in silico* predicted PROSIT spectral library.

Our observation suggests that OpenSwath and Skyline tend to report background ‘noise’ instead of missing values when no proteins can be confidently identified and quantified. In contrast, DIA-NN and Spectronaut tend in our analyses to report missing values when the proteins are physically absent in a sample.

Furthermore, we assessed how the missingness within each sample correlates with the sample mean of protein intensities. While for DIA-NN and Spectronaut this correlation is positive showing a separation of the spike-in conditions by sample mean of protein intensities, it is negative for Skyline and OpenSwath without showing such a separation (Figure 3B).

We hypothesize that the counter-intuitive positive correlation between protein intensity and missingness, as in the case of DIA-NN and Spectronaut, may be due to sample-dependent detection thresholds [19]. In other words, if the intensity of a protein lies below such a threshold, it is not included into the calculation of the sample mean of the protein intensities, thus, increasing the weight of proteins with higher intensities. This in turn increases the sample mean of protein intensities.

The implications of these findings are far-reaching and should be taken into consideration when planning studies, as in practically all proteomics experiments missing proteins are an issue that needs to be addressed, and in some studies such as knockout experiments or clinical biomarker discovery studies missing values are even of special interest. To our knowledge, detection of true missingness and false-positive quantitation is rarely investigated in benchmarking studies. Our dataset offers a well-suited platform to investigate (and possibly optimize) these aspects for future toolsets.

Post-Processing and Measures of Performance Used in this Study

Although complex in its own realm, protein and peptide identification and quantitation from LC-MS/MS data are only the beginning of the complete analysis of a multi-sample, quantitative proteomics experiment. Subsequent steps typically include sparsity reduction, normalization and, ultimately, statistical assessment of differential protein abundance. For each of the aforementioned steps different algorithms exist, yielding a variety of possible combinations.

To investigate the performance of the analysis methods in different possible combinations, we jointly assessed commonly used approaches for sparsity reduction, normalization, and different statistical tests. For sparsity reduction we applied: a) no sparsity reduction (NoSR), b) requiring > 66% values per protein (SR66), and c) requiring > 90% values per protein (SR90). Four different methods were then applied to investigate the effect of normalization: a) unnormalized, b) quantile normalization (QN), c) tail-robust quantile normalization (TRQN), and d) median normalization. We then subjected each possible combination to five statistical tests to probe for differentially abundant proteins.

To systematically evaluate the performance of each of the above mentioned parameters, we focused on a sub-dataset, representing the two lowest *E. coli* spike-in conditions. We used bootstrapping to quantify the uncertainty of the observed assessment score and to investigate the effect of sample size on the overall ability of the data analysis workflow to detect differentially abundant proteins. To this end, we randomly drew (with replacement) from the set of samples of the two lowest *E. coli* spike-in conditions to receive group sizes of three to 23 samples. On each bootstrap dataset we applied all combinations of the aforementioned sparsity reductions, normalizations, and statistical testing options to determine differentially abundant proteins.

To objectively compare the performance of the different data analysis workflows, we introduced a unifying measure of performance for detecting differentially abundant proteins. The experimental design with the known *E. coli* spike-in conditions provides us with ground truth information based on which we can assess true positives (*E. coli* proteins, which are determined to be significantly differentially abundant between the two spike-in conditions), false positives (human proteins determined to be significantly differentially abundant between the two spike-in conditions), false negatives (*E. coli* proteins determined to be non-significantly differentially abundant between the two spike-in levels), and true negatives (human proteins determined to be non-significantly differentially abundant between the two spike-in

conditions). For each protein, we can then plot the true positive rate against the false positive rate to obtain a receiver operating characteristic (ROC) curve. To measure the ability of each workflow to detect differentially abundant proteins, the area under the ROC curve is then determined. We use the partial area under the curve (pAUC) for all analyses, as the pAUC reflects the performance in the relevant range of false positives (Figure 4A) (Walter 2005). While the AUC can be interpreted as the average sensitivity over the whole range of specificities, pAUCs correspond to the average sensitivity over a relevant (mostly high) specificity range only. In the literature, calculations of pAUCs for different specificities have been used [6,20]. Here, we focus on specificities larger than 80% (i.e. false positive rate (FPR) < 20%) for the pAUC calculations.

Although the fold-changes of the spiked-in *E. coli* proteins are known through our study design, it is unknown which human and *E. coli* proteins were actually present in the biological sample in the first place. Since the calculations of sensitivities and specificities strongly depend on the definition of the set of proteins present, we calculated them based on three different protein lists. This allows us to evaluate the robustness of the outcomes, while ensuring that no software or library is favoured.

The proteins, which are present in the DIA analysis workflow, i.e. the bootstrap dataset under investigation, are collectively referred to as 'DIA workflow' proteins (Supplementary Figure 4 & 5). The list of proteins, which were identified in at least one of the DIA analysis workflows is referred to as 'Combined' (11,516 Human proteins, 2,127 *E. coli* proteins). The list of proteins, which were identified in more than 80% (at least 14 out of 17) of the DIA analysis workflows is referred to as 'Intersection' (4,499 Human proteins, 745 *E. coli* proteins), and represents a core protein set.

Evaluation of Post-Processing Approaches and Statistics of Differential Abundance

The highest pAUC values are achieved when no sparsity reduction is performed, while more strict criteria for missing values lead to a decrease in pAUC (Figure 4B). However, depending on the measure of performance being used, different aspects of an analysis workflow are rewarded or punished, representing an underlying challenge in benchmark studies. In our study, the reference protein list based on which the ROC is generated represents an additional that impacts the outcome of our comparisons.

The pAUC values shown in Figure 4B have all been generated using the 'DIA workflow' protein list. When we removed protein entries based on sparsity reduction in the 'DIA workflow', the pAUC values decreased. While the removal of proteins via sparsity reduction can lead to a situation in which the maximum sensitivity cannot be reached, the same can happen if the reference protein list, based on which sensitivity and specificity are calculated, is larger than the list of proteins, for which statistical results are available. This indicates that the choice of an appropriate reference protein list is crucial.

Furthermore, we observed a steep initial increase in the ROC curve for SR90, after which a plateau is reached. Differences in the pAUC are then solely based on the height of this plateau, which itself depends

on the number of quantified *E. coli* proteins in a given dataset. This steepness decreases from SR90 over SR66 to NoSR (data not shown).

If testing for differential abundance of a protein returned a missing value, the p-value for this comparison was set to one. As human proteins are overrepresented in our benchmark dataset, this might lead to a bias when performing sparsity reduction, limiting inter-comparability of sparsity reduction levels.

We next investigated the relative performance of each DIA analysis workflow separated by reference protein list (Figure 4C). Note the tri-model distribution of pAUC, which is particularly prominent for DIA-NN and Spectronaut, and can be explained by the three included sparsity reduction options. The performance of some DIA data analysis workflows differs drastically between the reference protein lists.

"Within workflow" performance

Using the 'DIA workflow' protein lists to measure the prediction performance of differentially abundant proteins, we find that Spectronaut's 'DirectDIA' performs best. DIA-NN, Skyline and Spectronaut all perform well using the more classical DDA spectral libraries generated by MaxQuant and MSFragger. Combining OpenSwath with the MSFragger-based spectral library leads to a better prediction performance than combining it with the MaxQuant spectral library. Overall, the GPF-refined libraries show an inferior performance, except for the refined DIA-NN spectral library in combination with OpenSwath.

"Overall sensitivity" performance of each workflow compared to all other workflows

When using the Combined reference protein list, the GPF-refined libraries, but not the *in silico* predicted DIA-NN unrefined library, perform well for DIA-NN and OpenSwath workflows. These libraries do not, however, perform as well for Spectronaut. Skyline clearly performs better with the refined PROSIT spectral library as compared to the refined DIA-NN spectral library for this specific reference protein list. Also, the DDA-based spectral libraries perform worse than in the case of the DIA workflow protein list.

Performance against "core protein dataset"

When using the 'Intersection' reference protein list, on average, DIA-NN performs slightly better compared to the other software solutions. The refined DIA-NN library (also in combination with OpenSwath, but not with Skyline and Spectronaut) leads to a good prediction performance against the protein core dataset.

The performance of each DIA analysis software suite strongly depends on the spectral library with which it is combined, and on the protein list against which it was benchmarked. For instance, while the relative performance against the DIA workflow protein list was below average for the combination of refined DIA-NN spectral library and DIA-NN analysis, the performance against the Combined protein list and against the Intersection protein list was among the best. In contrast, Skyline in combination with the refined PROSIT spectral library does not perform well when measured against the respective protein list, but performs well when measured against the Combined protein list.

These data highlight the strengths, but also the limitations of any spectral library-DIA software combination. While some combinations lead to a high number of reported proteins, others will give more accurate results. This can also be observed when comparing the detected log2 fold-changes of *E. coli* proteins with the actual fold-changes of the spiked-in *E. coli* lysates (Supplementary Figure S6).

Normalization

We found that virtually all DIA software-spectral library combinations do not benefit from normalization and perform best with non-normalized data (Supplementary Figure S7). All normalization methods included in this study normalize by distribution and, thus, act under the assumption of a symmetric differential expression, i.e. that the number of up- and down-regulated proteins is equal [21]. In this benchmark dataset, however, the differentially expressed proteins solely change in one direction. Thus, we hypothesize that the observed decline of performance when normalization is performed could, at least partly, be an artifact of the study design.

Furthermore, due to the higher number of human than *E. coli* proteins in the samples, the impact of human proteins on the normalised outcome is higher. As a result, the distribution of the human proteins is comparable across the normalised samples, while this is not the case for the *E. coli* proteins. This might lead to a bias in the identification of differentially abundant *E. coli* proteins.

Additionally, all employed normalization steps assume that the relative abundance of proteins within one sample can be used to normalize all proteins. This, however, cannot be assumed for this dataset as *E. coli* and human proteins were pipetted separately, which leads to changes in the protein abundance ranks between samples (which are assumed to be stable by the normalization methods). This highlights the need to employ special strategies to evaluate normalization strategies in future benchmark studies. A dilution series of the same samples being measured with different injection amounts may be more suitable to investigate normalization methods.

Statistical tests

Finally, we evaluated the prediction performance of statistical tests for two-group comparisons, which have previously been used in proteomics data analysis (Figure 4D), again for all three reference protein lists. In general, the non-parametric tests ROTS and SAM consistently perform best for all DIA analysis workflows. Interestingly, SAM performed better than ROTS with the exception of data analysed by Skyline, where ROTS consistently performed best (Supplementary Figure S8). However, the superiority of SAM over ROTS may also be due to the set hyperparameters, as the SAM statistic can be derived from the more general ROTS statistic. The good performance of non-parametric methods has been described previously [20,22]. Interestingly, SAM did not perform well in the study of Pursiheimo et al, who caution when using SAM but highlight the good performance of ROTS [23].

Connection between data characteristics and statistical prediction performance

We investigated the connection between data properties of the bootstrap datasets and statistical prediction performance. As we identified DIA-NN in combination with the *in silico* predicted GPF-refined spectral library as an overall well performing DIA analysis workflow, we further investigated, which data properties (Supplementary Figure S9) correlate with benchmarking performance measures (Supplementary Figure S10).

In general, sample variance, kurtosis, skewness and the ratio of variances between two spike-in conditions only show little influence on the performance of statistical tests to detect differentially abundant proteins. The correlation behavior of the remaining data characteristics differ between the DIA analysis workflows (not shown).

As we included different sample sizes during bootstrapping to mimic limited replicate availability, we were also able to investigate the performance of the different DIA analysis workflow for different sample sizes (Supplementary Figure S11). We observed a moderate positive correlation between pAUC values and sample size for all DIA analysis workflows (exemplarily shown in Supplementary Figure S12 for DIA-NN in combination with the *in silico* predicted GPF-refined spectral library).

Influence of sample size on statistical testing and performance

Overall, SAM performed best for all sample sizes over all workflows, except for Skyline, which showed the best performance in combination with ROTS. However, irrespective of the software suite, for small sample sizes ($n < 5$) limma achieved a similar performance to SAM and ROTS. Van Ooijen et al., who compared different statistical tools to detect differentially abundant proteome features, also found limma to perform well for small sample sizes [24].

Conclusions

With this comprehensive benchmark study in which we assessed multiple processing options simultaneously, we strive to support the proteomics community by providing novel insights into the interplay between spectral libraries, DIA software suites, data preprocessing, and statistical testing for differentially abundant proteins.

We found that the most extensive proteome coverage was achieved using Skyline in combination with an *in silico* predicted PROSIT spectral library, which was refined using EncyclopeDIA. DIA-NN and Spectronaut robustly avoided the false detection of *E. coli* proteins, which are truly absent in human-only samples. This is highly relevant for studies inferring biological relevance from missing values, especially in a clinical context.

Naturally, the amount of missing values also influences the effect of sparsity reduction. This effect is smaller for OpenSwath and Skyline as compared to DIA-NN and Spectronaut, potentially due to the differing nature of missing values.

In our study we found a very limited effect of data normalization on the prediction performance of differentially abundant proteins. This highlights the quality of internal protein inference and summarization algorithms for all tools, especially for DIA-NN and Spectronaut.

As for statistical testing, the non-parametric statistical tests SAM and ROTS consistently performed well, with SAM outperforming ROTS when DIA-NN, OpenSwath, or Spectronaut were used, while ROTS performed best in case of Skyline. Limma performed well compared to other parametric statistical tests for very low sample sizes. The performance to detect differentially abundant proteins changed for the spectral library - DIA analysis software combinations depending on the employed performance measure. When the number of identified proteins was considered irrespective of the DIA software-library combination they were derived from, OpenSwath in combination with the *in silico* predicted DIA-NN spectral library (GPF-refined by DIA-NN) performed best. When considering the core protein dataset (identified in 80 % of all DIA analysis workflows), DIA-NN in combination with the *in silico* predicted DIA-NN spectral library (GPF-refined by DIA-NN) performed best. When only the proteins are taken into account, which were found in the respective DIA software-library combination, Spectronaut's "DirectDIA" excelled.

This highlights the importance of spectral library generation and the quality of the resulting spectral libraries. *In silico* prediction of spectral libraries (and their refinement on LC-MS/MS measurements) is gaining momentum. In this setting, the choice of library prediction algorithm, possible LC-MS/MS refinement, and the actual DIA-analysis software lead to even more complex combinatorial workflows.

In summary, we found that the reliability and reproducibility of proteomics data analyses heavily depend on properly choosing and combining the options provided for each proteomics workflow step, as

downstream analyses may rely on certain assumptions about data characteristics, which are themselves heavily influenced by the choice of DIA software and spectral libraries.

We invite others to test their approaches on our dataset as it provides a unique opportunity to test DIA analysis workflows as well as data analysis workflows in a heterogeneous background setting. We furthermore recommend that workflows used in clinical settings should be tested against our dataset to control their performance and expected data structure.

Materials And Methods

Sample Preparation

The study has been approved by the Ethics Board of the University Medical Center Freiburg (approval 280/18). Histologically non-infiltrated lymph nodes from patients with acinary prostate cancer were collected as sentinel samples and preserved as FFPE tissue. Consecutive slices of 10 µm thickness were deparaffinized, stained, and macrodissected to acquire 0.5 - 1 mm^3 of lymph node tissue per patient. Subsequently, antigen retrieval was performed in 4% (v/v) SDS, 100 mM HEPES pH 8.0, with samples being sonicated using a Bioruptor device for 10 cycles (40 sec / 20 sec, high intensity), heated to 95°C for one hour, and sonicated again.

E. coli K12 bacteria were provided by Christoph Schell (University Medical Center, Freiburg) as cell pellets. *E. coli* samples were lysed in 4 % SDS in 100 mM HEPES pH 8 and heated to 95 °C for 10 min and subsequently sonicated using a Bioruptor for 15 cycles (40 sec / 20 sec, high intensity).

All samples were centrifuged at 15000 rcf for 10 min at room temperature. Only the supernatant was used for MS sample preparation.

FFPE tissue samples and *E. coli* samples were reduced at 95°C for 10 min using 5 mM TCEP. Samples were alkylated for 20 min at room temperature in the dark using 10 mM iodoacetamide. Samples were prepared for MS analysis using micro S-TRAP columns (PROTIFY) according to manufacturer's instructions. For digestion, a mix of trypsin and Lys-C (1:20 w/w to sample protein amount) was used. Following purification, peptide content was measured using BCA assay (Thermo) and aliquots containing 5 µg peptide per sample were dried and frozen at -80°C until measurement.

For DDA library generation, a mastermix of 10 different lymph node peptide preparations and *E. coli* peptides with a human / *E. coli* ratio of 15:1 was used. The sample was pre-fractionated using offline high-pH prefractionation as described previously resulting in 10 fractions [25].

LC-MS/MS Measurements

All LC-MS/MS runs were acquired using an Orbitrap Eclipse Mass Spectrometer (Thermo) coupled to an Easy nLC 1200 (Thermo). Precolumns with 100 µm ID were self-packed with 3 µm C18 AQ (Dr. Maisch) beads to a length of 2 cm. A 75 µm Picofrit column (New Objective) was self-packed with 1.9 µm C18 AQ

(Dr. Maisch) beads to a length of 20 cm as previously described [26]. For every injection, 500 ng of peptides were used. iRT peptides (Biognosys) were added to a final quantity of 50 fmol / injection. Buffer A consisted of 0.1% formic acid, buffer B consisted of 80 % acetonitrile in 0.1 % formic acid. All samples were separated using a 70 min linear gradient from 5 % to 31 % B followed by a 5 min linear gradient from 31 % to 44 % buffer B. For the data acquisition of the dilution series the mass spectrometer was operated in DIA mode and the standard parameters from the staggered DIA method editor node were used. Briefly, a survey scan (60k resolution) from 390 to 1010 m/z was followed by MS2 scans (15k resolution) with 8 m/z isolation width covering 400 m/z to 1000 m/z. A second survey scan was followed by MS2 scans with an offset of 4 m/z as compared to the first cycle. For MS2 scans, peptides were fragmented using HCD and stepped collision energy 30 (5%), and maximum injection time was set to 22 ms. The data were recorded in centroid mode.
For spectral library generation, a masterpool sample with a lymph node to *E. coli* peptide ratio of 15:1 was generated by combining peptides from 12 randomly chosen samples (3 from each spike-in condition).

For GPF measurements, the masterpool sample was repeatedly measured. A tSIM scan with an isolation width of 110 m/z was followed by MS2 scans with 4 m/z isolation width over 100 m/z. A second tSIM scan with 110 m/z was followed by MS2 scans with an offset of 2 m/z as compared to the first cycle. A total of 6 measurements were performed to cover a scan range from 400 to 1000 m/z.

For data dependent acquisition measurements, the masterpool sample was pre-fractionated offline prior to LC-MS/MS measurement as described previously [25]. A survey scan of 120k ranging from 390 m/z to 1010 m/z was recorded. Following the survey scan, a Top 15 method was employed. MS2 scans were recorded at 15k resolution with the isolation window set to 1.6 m/z and maximum injection time set to 60 ms. DDA data integrity was validated using PTXQC [27].

Peptide-Spectrum Matching (PSM) and Signal Quantitation for LC-MS/MS data

Spectral Library Generation

For all spectral libraries, a reviewed human and *E. coli* K12 FASTA (one entry per gene) were downloaded from Uniprot on Nov 22nd 2020 [28]. The GPF-refined PROSIT [29] spectral library was generated as described previously [14]. In brief, EncyclopeDIA [17] (0.9.5) was used to generate PROSIT input csv files. PROSIT (2019 iRT prediction model) was used to predict spectra and retention times, which were reimported into EncyclopeDIA. Destaggered GPF mzml files were then used to generate a GPF-refined library, which was exported in tabular format.
The GPF-refined DIA-NN spectral library was predicted and refined using DIA-NN. In brief, DIA-NN was provided with a combined FASTA protein database (human + *E. coli*) as input and neural networks were used to generate spectra and retention times for the appropriate mass range of 390 m/z to 1010 m/z.

The GPF mzml files were then used to generate a GPF refined library, which was exported in tabular format. For the DIA-NN *in silico* predicted library, the refinement step was skipped and the unrefined *in silico* predicted library was directly used for DIA analysis.

The MaxQuant DDA library was generated using MaxQuant (1.6.14.0) searching the DDA files resulting from prefractionation directly as raw files. The MaxQuant output was imported into Spectronaut, converted to library format and exported in tabular format.

The MSFragger DDA library was generated using MSFragger (3.2) in the Fragpipe GUI (14.0) in conjunction with SpectraST, following conversion of DDA raw files to mzXML format. MSFragger output was converted to tabular format using DIA-NN.

Raw files were destaggered and converted to mzml or mzxml format using MSConvert in conjunction with ProteoWizard (3.0.20315) [30] or demultiplexed and converted to htrms format using Spectronaut (14.0).

DIA Data Analysis

DIA-NN (1.7.12) was used with recommended settings. Mass ranges were set appropriately for the search space and RT profiling was activated. For the *in silico* predicted library search, the reduced memory option was additionally activated. Protein FDR was set to 1.0 %. All DIA-NN computations were performed on an Intel(R) Xeon(R) Gold 6246 CPU.

Skyline [31] (64 Bit) (20.2.0.343) analyses were performed as described in the Skyline tutorials 'Analysis of DIA/SWATH data' and 'Advanced Peak Picking Models'. In brief, the 'Import Peptide Search' daemon was used to import spectral libraries and implement the iRT retention time predictor [32]. Mass accuracy was set to 10 ppm. A mProphet model was trained not including MS1 information and results were filtered based on the q-value given by the mProphet model (1 % peptide FDR).

The OpenSwathWorkflow (2.6) was used in Galaxy with the default settings except for minor adjustments [9]. Briefly, the mass accuracy on MS1 and MS2 level were set to 10 ppm. For iRT peptide extraction, 20 ppm was used, and a minimum of seven iRT peptides was requested. Target-decoy scoring was performed using PyProphet (2.1.4.2) in Galaxy with the 'XGBoost' classifier for semi-supervised learning including MS1 as well as MS2 information [33]. Identification results were filtered based on a peptide and protein FDR of 1% using PyProphet.

Spectronaut's (14.0) 'import' function was used for converting tabular libraries into Spectronaut format. Before import, the retention times of the PROSIT-EncyclopeDIA and the DIA-NN libraries were converted to minutes and a linear model was used to convert retention times to iRT values.

All raw data, libraries, analysis log files, and analysis output files are available at the European Genome-phenome Archive (https://ega-archive.org EGAS00001005589).

Data Post-Processing and Visualization of DIA Analysis Software Output

Prior to performing data analysis on the protein intensity datasets derived from the 17 DIA workflow analyses, data were transformed to a common format, in which all proteins were annotated with their respective UniProtKB/Swiss-Prot entry names. For some DIA analysis workflows, multiple protein identifiers were composed of multiple protein names. Proteins were excluded if they were labelled both as a human and as derived from *E. coli.* Proteins without reported quantitations were removed. For further analysis and visualization, the resulting protein intensities were log2-transformed.

Bootstrapping

To evaluate statistical tools to identify differentially abundant proteins in omics data, the two lowest human : *E. coli* spike-in ratios 25:1 and 12:1 were used (Figure 1). Bootstrap datasets were generated by randomly drawing (with replacement) a defined number of samples from each of the two spike-in conditions. We varied the group sizes for each spike-in condition from three to 23 samples and generated 100 bootstrap datasets for each of those group sizes, resulting in 2100 bootstrap datasets in total. To each of those bootstrap datasets all data analysis workflows consisting of different combinations of sparsity reduction, normalization, and statistical testing options were applied.

Sparsity Reduction and Normalization

We included the following three sparsity reduction options: including all protein entries (NoSR), or only those protein entries present in at least 66 % (SR66), or 90 % (SR90) of all samples, respectively. We included the following four normalization options: no normalization (unnormalized), tail-robust quantile normalization (TRQN) [19], quantile normalization (QN) [34,35], and median normalization (median).

Statistical Testing for Differentially Abundant Proteins

The following five statistical tests were included in our analyses: Student's t-test (with equal variances), linear models for microarray data (limma) [36], generalized linear model (GLM), significance analysis of microarrays (SAM) (250 permutations used to estimate false discovery rates) [37], and reproducibility-optimized test statistic (ROTS) [38,39] (with 100 bootstrap and permutation resamplings and the largest top list size considered being 500).

In total, we acquired performance information for 17 DIA analysis workflows x 2100 bootstrap datasets x 3 sparsity reduction options x 4 normalization options x 5 statistical test options = 2’142’000 cases. For each of those cases we received for each protein of a bootstrap dataset a p-value as a result of a

statistical test and the estimated log2-fold change (log2FC) between the two different human to *E. coli* spike-in ratios 12:1 and 25:1.

Measure of Performance

To evaluate which analysis performed best in predicting the differentially abundant proteins, we used the partial area under the curve (pAUC), specifically the area under the receiver operating characteristic (ROC) curve. If statistical tests returned a missing value for a given protein the p-value of this protein was set to 1 in the respective analysis. We also calculated the sensitivity at a significance level of 0.05.

To quantify precision, we calculated the root-mean-square error (RMSE) based on the estimated log2FC and the true log2FC, which is 0 for human proteins and 1.11 for *E. coli* proteins.

We calculated the evaluation measures for three reference protein lists in parallel: the 'core'/'intersect' protein set with proteins appearing in more than 80% (at least 14 of 17) DIA analysis workflow datasets, the 'combined' protein set with proteins appearing in at least one DIA analysis workflow dataset, and the 'DIAWorkflow' protein set, which is specific for each DIA analysis workflow.

The R code used for the statistical analyses is available at github.com/kreutz-lab/dia-benchmarking.

Declarations

Acknowledgements

The authors acknowledge support by the state of Baden-Württemberg through bwHPC

and the German Research Foundation (DFG) through grant INST 35/1134-1 FUGG. OS acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, SCHI 871/17- 1, NY 90/6-1; SCHI 871/15-1, GR 4553/5-1, PA 2807/3-1, project-ID 431984000 – SFB 1453 (project P12); project-ID 441891347 - SFB 1479 (project S1); project-ID 423813989/GRK2606 (RTG "ProtPath"), INST 39/766-3 (Z1)), the ERA PerMed programs (BMBF, 01KU1916, 01KU1915A), the German-Israel Foundation (grant no. 1444), and the German Consortium for Translational Cancer Research (project Impro-Rec).

Conflict of Interest

The authors declare no conflict of interest

Author Contributions

Conceptualization: KF, EB, LK, PB, AS, KB, CK, OS

Methodology: KF, EB, MF, DV, LK, PB, STB

Validation: KF, EB, MF

Formal Analysis: KF, EB, MF, DV, PB, STB, CK, OS

Resources: PB, STB, AS, KB, CK, OS

Data Curation: KF, EB, MF, PB, CK, OS

Writing (Original Draft): KF, EB, LK, KB, CK, OS

Writing (Review): KF, EB, KB, CK, OS

Visualizations: KF, EB

Project Administration: CK, OS

Funding acquisition: CK, OS

References

1. Gillet, L. C. *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics* **11**, O111.016717 (2012).

2. Muntel, J. *et al.* Surpassing 10 000 identified and quantified proteins in a single run by optimizing current LC-MS instrumentation and data analysis strategy. *Mol Omics* **15**, 348–360 (2019).

3. Lou, R. *et al.* Hybrid Spectral Library Combining DIA-MS Data and a Targeted Virtual Library Substantially Deepens the Proteome Coverage. *iScience* **23**, 100903 (2020).

4. Gotti, C. *et al.* Extensive and accurate benchmarking of DIA acquisition methods and software tools using a complex proteomic standard. *Cold Spring Harbor Laboratory* 2020.11.03.365585 (2020) doi:10.1101/2020.11.03.365585.

5. Navarro, P. *et al.* A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* **34**, 1130–1136 (2016).

6. Suomi, T. & Elo, L. L. Enhanced differential expression statistics for data-independent acquisition proteomics. *Sci. Rep.* **7**, 5869 (2017).

7. Amodei, D. *et al.* Improving Precursor Selectivity in Data-Independent Acquisition Using Overlapping Windows. *J. Am. Soc. Mass Spectrom.* **30**, 669–684 (2019).

8. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17**, 41–44 (2020).

9. Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **32**, 219–223 (2014).

10. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).

11. Bruderer, R. *et al.* Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteomics* **14**, 1400–1410 (2015).

12. Dowell, J. A., Wright, L. J., Armstrong, E. A. & Denu, J. M. Benchmarking Quantitative Performance in Label-Free Proteomics. *ACS Omega* **6**, 2494–2504 (2021).

13. Choi, M. *et al.* ABRF Proteome Informatics Research Group (iPRG) 2015 Study: Detection of Differentially Abundant Proteins in Label-Free Quantitative LC-MS/MS Experiments. *J. Proteome Res.* **16**, 945–957 (2017).

14. Searle, B. C. *et al.* Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nat. Commun.* **11**, 1548 (2020).

15. Yang, F., Shen, Y., Camp, D. G., 2nd & Smith, R. D. High-pH reversed-phase chromatography with fraction concatenation for 2D proteomic analysis. *Expert Rev. Proteomics* **9**, 129–134 (2012).

16. Demichev, V. *et al.* High sensitivity dia-PASEF proteomics with DIA-NN and FragPipe. *bioRxiv* (2021) doi:10.1101/2021.03.08.434385.

17. Searle, B. C. *et al.* Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nat. Commun.* **9**, 1–12 (2018).

18. McGurk, K. A. *et al.* The use of missing values in proteomic data-independent acquisition mass spectrometry to enable disease activity discrimination. *Bioinformatics* **36**, 2217–2223 (2020).

19. Brombacher, E., Schad, A. & Kreutz, C. Tail-Robust Quantile Normalization. *Proteomics* **20**, e2000068 (2020).

20. Wang, J. *et al.* In-depth method assessments of differentially expressed protein detection for shotgun proteomics data with missing values. *Sci. Rep.* **7**, 3367 (2017).

21. Evans, C., Hardin, J. & Stoebel, D. M. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.* **19**, 776–792 (2018).

22. Klammer, M., Dybowski, J. N., Hoffmann, D. & Schaab, C. Identification of significant features by the Global Mean Rank test. *PLoS One* **9**, e104504 (2014).

23. Pursiheimo, A. *et al.* Optimization of Statistical Methods Impact on Quantitative Proteomics Data. *J. Proteome Res.* **14**, 4118–4126 (2015).

24. van Ooijen, M. P. *et al.* Identification of differentially expressed peptides in high-throughput proteomics data. *Brief. Bioinform.* **19**, 971–981 (2018).

25. Baumert, H. M. *et al.* Depletion of histone methyltransferase KMT9 inhibits lung cancer cell proliferation by inducing non-apoptotic cell death. *Cancer Cell Int.* **20**, 52 (2020).

26. Kovalchuk, S. I., Jensen, O. N. & Rogowska-Wrzesinska, A. FlashPack: Fast and Simple Preparation of Ultrahigh-performance Capillary Columns for LC-MS. *Mol. Cell. Proteomics* **18**, 383–390 (2019).

27. Bielow, C., Mastrobuoni, G. & Kempa, S. Proteomics Quality Control: Quality Control Software for MaxQuant Results. *J. Proteome Res.* **15**, 777–787 (2016).

28. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).

29. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).

30. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).

31. Pino, L. K. *et al.* The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. *Mass Spectrom. Rev.* **39**, 229–244 (2020).

32. Escher, C. *et al.* Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* **12**, 1111–1121 (2012).

33. Teleman, J. *et al.* DIANA—algorithmic improvements for analysis of data-independent acquisition MS data. *Bioinformatics* vol. 31 555–562 (2015).

34. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).

35. Amaratunga, D. & Cabrera, J. Analysis of data from viral DNA microchips. *J. Am. Stat. Assoc.* **96**, 1161–1170 (2001).

36. Smyth, G. K. limma: Linear Models for Microarray Data. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* 397–420 (Springer-Verlag, 2005).

37. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 5116–5121 (2001).

38. Suomi, T., Seyednasrollah, F., Jaakkola, M. K., Faux, T. & Elo, L. L. ROTS: An R package for reproducibility-optimized statistical testing. *PLoS Comput. Biol.* **13**, e1005562 (2017).

39. Elo, L. L., Filen, S., Lahesmaa, R. & Aittokallio, T. Reproducibility-optimized test statistic for ranking genes in microarray studies. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **5**, 423–431 (2008).

Figures

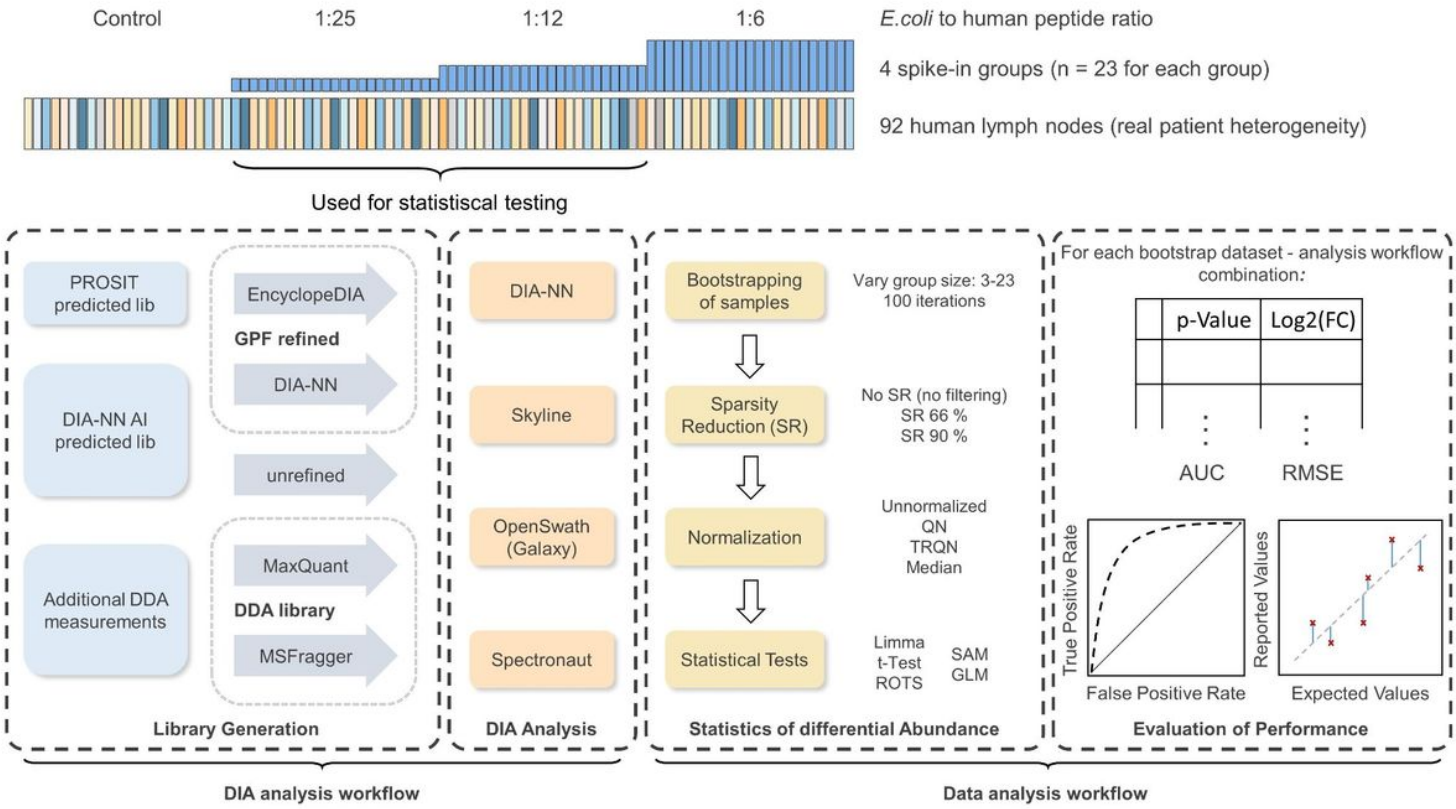

Figure 1

Benchmarking Workflow A data-independent acquisition (DIA) benchmark dataset was created by adding E. coli peptides in known ratios to peptide preparations of lymph nodes of 92 individuals. We analyzed the raw data with different spectral libraries and DIA software suites. From samples to which E. coli peptides were added in the two E. coli : human peptide ratios 25:1 and 12:1, bootstrap datasets with group sizes of 3 to 23 were generated. For each of those 21 different group sizes, 100 bootstrap datasets were generated. On each bootstrap dataset different data analysis workflows, composed of different sparsity reductions, normalization options, and different statistical testing methods for differentially

abundant proteins, were applied. The results were returned in a table containing p-values and log2-fold changes (log2FCs). As the ground truth about the changed proteins (E. coli) is known, the prediction performance of each workflow can be assessed. This can be done based on the p-values from the statistical tests by calculating the receiver operating characteristic (ROC) curve, based on which the area under curve (AUC) is calculated. To quantify the accuracy of quantification the root-mean-square error (RMSE) is calculated based on the detected log2FC.

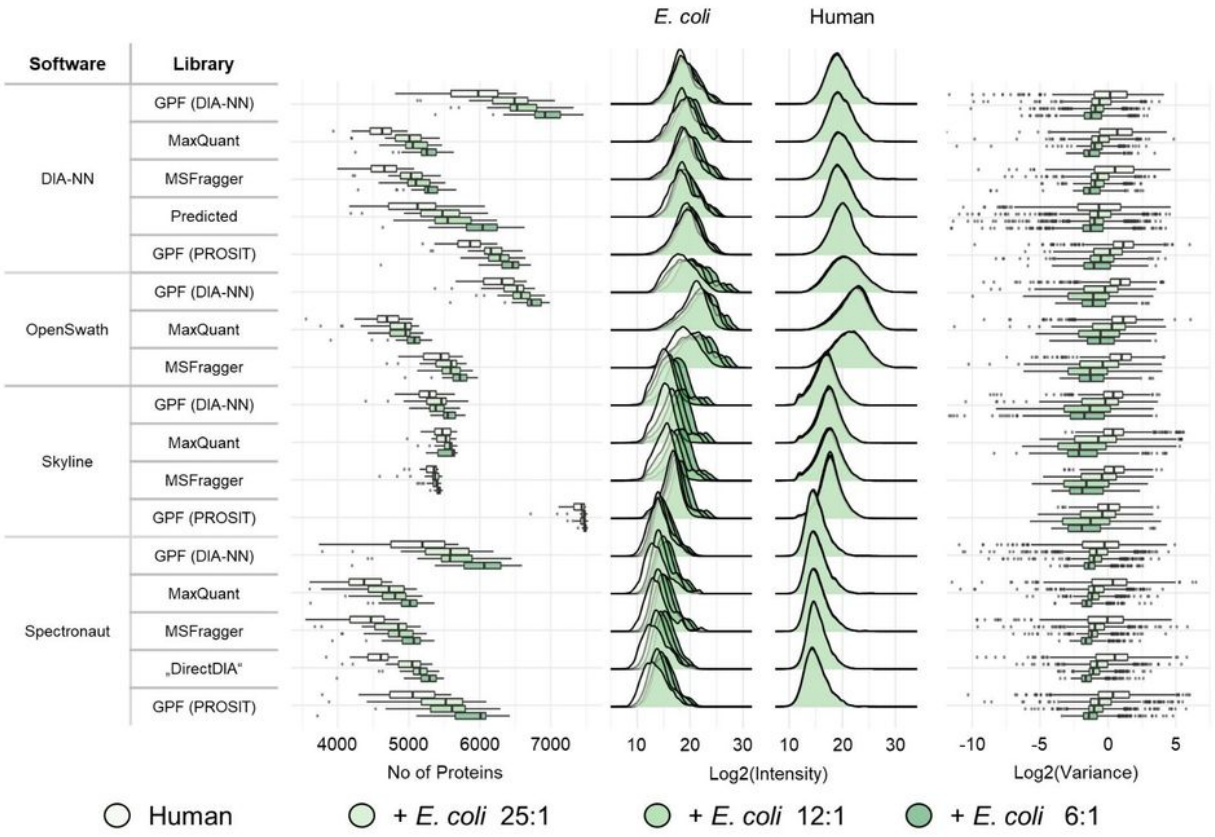

Figure 2

Choice of Spectral Library and DIA Analysis Software Influences Number of Identified Proteins. Left: Number of all identified and quantified proteins (human and E. coli) in all 92 samples. Center: Log2 intensity distributions of proteins. Right: Log2 variance of E. coli proteins. Log2 Variance values smaller than -12 were excluded from this plot. Color-coded by spike-in condition

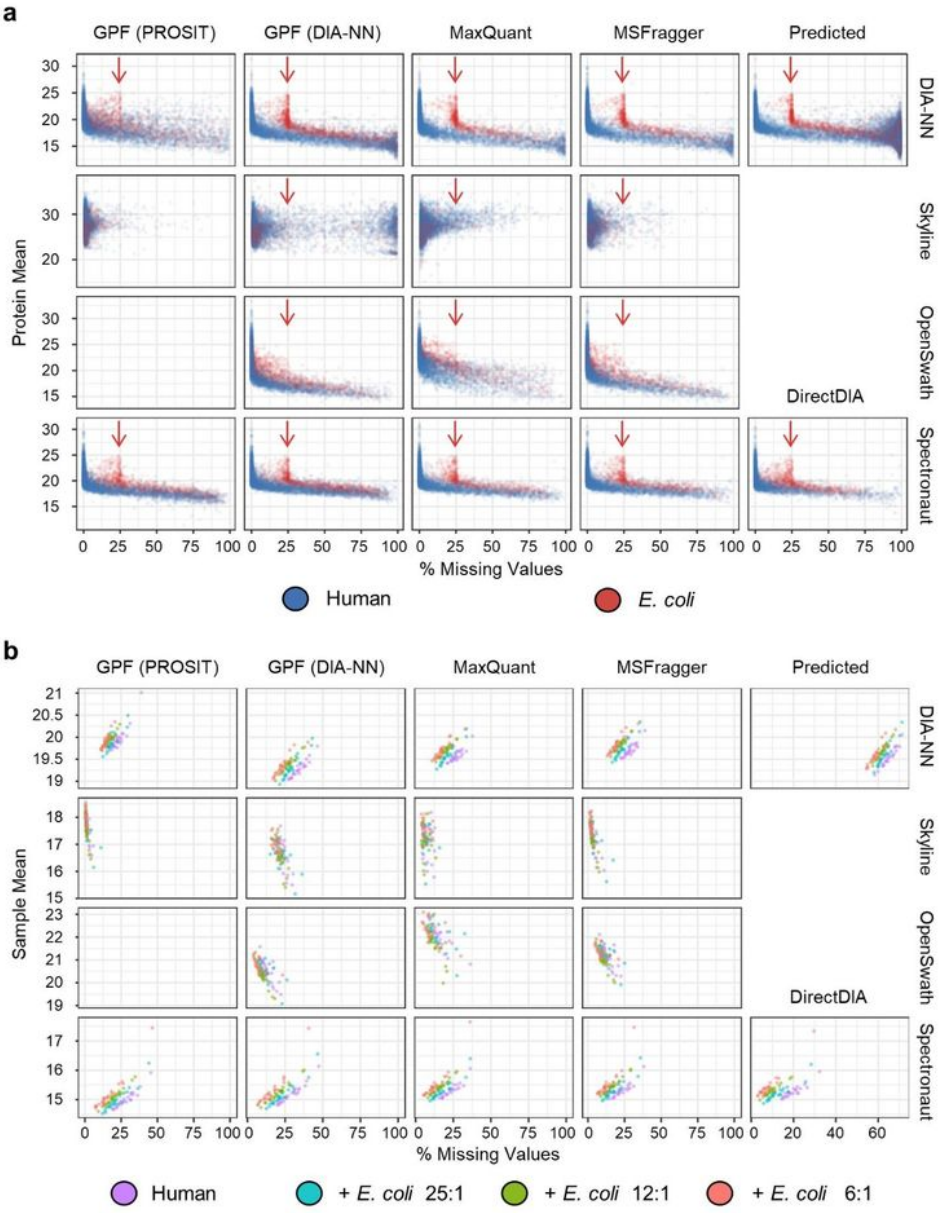

Figure 3

Missing Value Characteristics. a) Missingness of proteins is reported differently and mainly varies with the employed DIA software. Means of Log2 intensities of identified human (blue) and E. coli (red) proteins plotted against the percentage of missing values in the respective protein. E. coli proteins are not physically present in 25% of samples (indicated by the red arrow). b) The correlation between the missingness within samples and the sample means of the protein intensities varies with the employed DIA software. Sample means of protein intensities are plotted against the percentage of missing values in the respective sample.

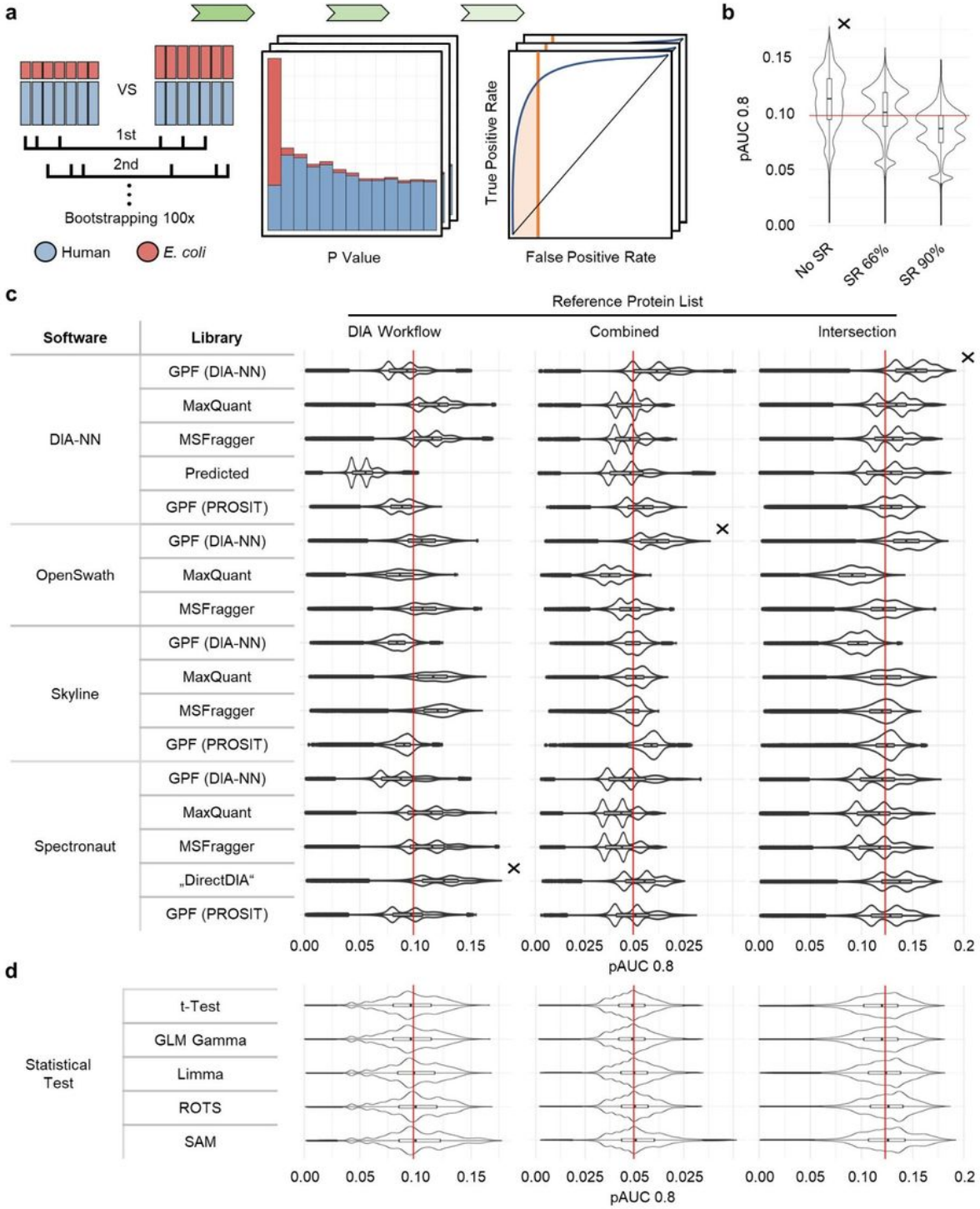

Figure 4

Statistical Analysis of Benchmark Dataset a) Workflow schematic: For the generation of bootstrap datasets random samples were drawn with replacement from samples of the spike-in conditions 25:1 and 12:1 mimicking two groups containing differentially abundant proteins, here represented by all E. coli proteins. The p-values acquired after data preprocessing and statistical analysis were used to build receiver operating characteristic (ROC) curves. The partial area under the curve (pAUC) for specificities larger than 0.8 was used as a measure of prediction performance. b) pAUC distribution for the different sparsity reduction options (as measured against 'DIA workflow' protein list) c) pAUC for the different DIA analysis workflows as measured against the three different reference protein lists d) pAUC distributions for the statistical tests. 'DIA workflow' describes the performance against the proteins present in the given DIA workflow only, 'Combined' describes the performance against proteins identified at least by one of all DIA analysis workflows. 'Intersection' describes the performance against proteins which were found in more than 80% (in at least 14 of 17) of the DIA analysis workflows. For each reference protein list the respective median of pAUC performance is indicated by a red line, and the best performing option with a cross.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BenchmarkPaperNatCommSubmissionSuppl.docx
- BenchmarkPaperNatCommSubmissionSuppl.docx