In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics

Audain, Enrique; Uszkoreit, Julian; Sachsenberg, Timo; Pfeuffer, Julianus; Liang, Xiao; Hermjakob, Henning; Sanchez, Aniel; Eisenacher, Martin; Reinert, Knut; Tabb, David L.; Kohlbacher, Oliver; Perez-Riverol, Yasset

Published in:
Journal of Proteomics

DOI:
10.1016/j.jprot.2016.08.002

2017

Document Version:
Peer reviewed version (aka post-print)

In-depth Analysis of Protein Inference Algorithms using a Workflow Framework and Well-Defined Metrics

Enrique Audain[‡, §, £], Julian Uszkoreit [¶, £], Timo Sachsenberg [§], Julianus Pfeuffer [#], Xiao Liang [#], Henning Hermjakob [¥], Aniel Sanchez [□], Martin Eisenacher [¶], Knut Reinert [#], David L. Tabb [ⓘ], Oliver Kohlbacher [§, ‡], Yasset Perez-Riverol [¥, *]

‡ Department of Proteomics, Center of Molecular Immunology. Ciudad de la Habana. Cuba;

§ Center for Bioinformatics, Quantitative Biology Center and Department of Computer Science, University of Tübingen, Sand 14, 72076 Tübingen, Germany;

¶ Medizinisches Proteom-Center, Ruhr-Universität Bochum, Universitätsstr. 150, D-44801 Bochum, Germany;

¥ European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL- EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

□ Department of Translational Medicine, Faculty of Medicine, Lund University Malmo, Sweden

Algorithmic Bioinformatics, Institut für Informatik, Freie Universität Berlin, Takustraße 9, 14195 Berlin

ⓘ Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN, 37232.

‡ Biomolecular Interactions, Max Planck Institute for Developmental Biology, Spemannstr. 35, 72076 Tübingen, Germany

£ Both authors contributed equally to this work.

* Corresponding author: Dr. Yasset Perez-Riverol, E-mail: yperez@ebi.ac.uk, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL- EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. Tel.: +44 1223 492 610; Fax: +44 1223 494 484.

Abstract

In mass spectrometry-based shotgun proteomics, protein identifications are usually the desired result. However, most of the analytical methods are based in the identification of reliable peptides and not the direct identification of intact proteins. Thus, assembling peptides identified from tandem mass spectra into a list of proteins, referred as protein inference, is a critical step in proteomics research. Currently, different protein inference algorithms and tools are available for the proteomics community. Here, we evaluated five dominant software tools for protein inference (PIA, ProteinProphet, Fido, ProteinLP, MSBayesPro) using three popular database search engines: Mascot, X!Tandem, and MS-GF+. All the algorithms were evaluated using a highly customizable KNIME workflow using four different public datasets with varying complexities (different sample preparation, species and analytical instruments). We defined a set of quality control metrics to evaluate the performance of each combination of search engines, protein inference algorithm, and parameters on each dataset.

1. Introduction

Proteomics can be used for the study of the biological functions of proteins, cellular localization, post-translational modifications (PTMs), and interactions between proteins (1, 2). The field has seen great development in the last years due to advances in mass spectrometry (MS) instrumentation, the development of new analytical methods (3-5), and novel computational approaches (2, 6). Bottom-up proteomics is currently the standard analytical method to identify and quantify proteins based on the presence of peptides obtained by digestion of the protein mix during sample preparation. Current computational approaches can typically be broken down into three main steps: 1) peptide identification, 2) quality assessment of the peptide identifications, and 3) the assembly of the identified peptides into a final protein list using protein inference algorithms (7, 8). During peptide identification, peptide fragmentation spectra (MS/MS) are assigned to peptide sequences to generate a set of Peptide-Spectrum Matches (PSMs) using database search engines, such as Mascot (9), MS-GF+ (10), or X!Tandem (11). Then, it is necessary to assess the reliability of these identifications (12) by estimating collective false discovery rates or by assessing correctness probabilities for each PSM. Finally, the identified peptide sequences are assembled into a set of confident proteins, which enables protein quantitation or pathway analysis (13).

Ideally, protein inference produces a protein list from the identified peptides with all proteins of the original sample prior to digestion. Unfortunately, ambiguities arise when an identified peptide sequence can be explained by more than one entry in a protein database (14). Under certain assumptions, some of these ambiguities can be resolved when taking other peptide identifications, physicochemical properties, or quantities into account. Unfortunately, there are cases when it is not possible to resolve an ambiguity, e.g. if two protein entries map to exactly the same sets of identified peptides (15).

In 2003, PeptideProphet and ProteinProphet were published as some of the first algorithms and tools to address the challenges of protein inference, using a probabilistic model (16). ProteinProphet, a widely used algorithm integrated into the Trans-Proteomic Pipeline (TPP), employs an iterative heuristic probability model to estimate protein probabilities based on peptide probabilities. Other algorithms have been proposed using Bayesian methods (14) or linear programming (17), incorporating additional information like the isoelectric point, retention time, or detectability during protein inference (18, 19). As a result, several protein inference implementations are available to the proteomics community (20) including the implementations provided by search engines such as Mascot

(9) or Andromeda (21, 22). In addition, a number of commercial tools provide protein inference, such as ProteomeDiscoverer (Thermo Scientific, http://www.thermoscientific.com/en/products/mass-spectrometry.html) and Scaffold (23). Despite this wide range of tools and algorithms, only a few evaluations have been performed to benchmark their performance (20, 24, 25). In 2012, Claassen and co-workers benchmarked ProteinProphet with different "gene locus inference" approaches and opened the field to perform other studies including other inference approaches (24). A thorough comparison is hampered by the large number of possible combinations of tools, problems with interoperability of tools (e.g., the use of proprietary file formats, insufficient documentation or platform-dependence), and the lack of a clear set of metrics for unbiased evaluation of the performance.

Here, we evaluate and benchmark five leading tools for protein inference: ProteinProphet (16), MSBayesPro (26), ProteinLP (27), Fido (14) and PIA (28). To achieve this, three popular search engines including Mascot (9), X!Tandem (11), MS-GF+ (10) and their combinations were used with every protein inference tool. We implemented a workflow in the highly customizable KNIME (29) workflow environment using a series of OpenMS nodes and several new workflow nodes (https://github.com/KNIME-OMICS) to study all combinations of these search algorithms and inference algorithms. This approach is scalable to arbitrary numbers of algorithms. We provide different metrics to benchmark the algorithms under study. Amongst others, the numbers of reported proteins, peptides per protein, and uniquely reported proteins per inference method are used to evaluate the performance of each inference method. Four datasets of different complexities and from different species were employed to evaluate the performance of protein inference algorithms including one "gold standard" or "ground truth" dataset previously used to compare protein inference algorithms (28, 30). The final results for complex samples (the yeast "gold standard" dataset and the human lung cancer dataset - PXD000603 -) vary not only regarding the actual numbers of protein groups but also concerning the actually reported groups. The robustness of the numbers of reported proteins when using databases of differing complexities is depending on the applied inference algorithm. The final results also showed that merging the identifications of multiple search engines does not necessarily increase the number of reported proteins, but does increase the number of peptides per protein and thus can generally be recommended. At the same time, the present study shows that proper selection of search engine and inference algorithm is crucial to the yield of information from proteomic data sets.

2. Material and methods

2.1 The benchmark workflow

The presented protein inference comparison workflow is based on KNIME (29) and OpenMS (31). We made use of the existing OpenMS nodes, but we also implemented additional nodes for some of the tools. The developed workflow can be split into seven different steps (Figure 1). The first step **(A)** configures basic variables like the regular expression to identify decoys in the FASTA protein database and the allowed FDR q-value threshold. Also, if a gold-standard dataset is analyzed, the reference protein list is loaded with the set of proteins known to be in the data. Step **(B)** performs conversion to mzML, optional peak centroiding for spectra recorded in profile mode, and removal of MS1 spectra. The remaining tandem spectra are searched in step **(C)** using three different search engines (X!Tandem, Mascot, and MS-GF+) using the adapter nodes provided by OpenMS. Furthermore, the results are filtered for peptides with a minimum length of 7 amino acids and exported to idXML files, OpenMS's internal format, for further processing. In step **(D)** all possible combinations for the results of the three search engines are created. Peptide posterior error probabilities are calculated with the *IDPosteriorErrorProbability* tool, which is a standalone OpenMS node used for estimating the probability of peptide hits to be incorrectly assigned (32). For the assessment of the combined search engine results, results are combined using the *Consensus ID* (33) incorporated in OpenMS with the *PEPMatrix* algorithm. After calculating the PSM FDR using the target decoy information, all peptides with FDR q-value > 0.01 are filtered out and no longer considered in the analysis. To evaluate the FDR on the protein level later on, the target and decoy PSMs below the 0.01 FDR q-value threshold are passed together to all protein inferences.

Since MSBayesPro requires a peptide detectability additional to the probability during the inference process, we compute a detectability model of all results in step **(E)** using the OpenMS node *PTModel* (34). The *IDFilter* node was used to get the high scoring identifications (500 distinct peptides or at least one fourth of all available) to train the *PTModel* model. Additionally, we provide a subset of the PSMs for low-confidence peptides (i.e. those 500 peptides with lowest identification scores/probabilities or the

lowest scoring fourth of all available) as training input to the model (Supplemental File S1, section 6). In the next step **(F)** of the workflow the final list of peptides with the corresponding probabilities and detectability values are imported into the PIA (28, 35, 36), Fido (14), ProteinLP (17), MSBayesPro (26) and ProteinProphet (16) nodes (Supplemental File S1, section 7) to generate the protein lists.

PIA used the *SpectrumExtractor* algorithm with the recommended settings (using the best *PSM FDR Score* per peptide as basis for protein score with multiplicative protein scoring). This algorithm selects for each spectrum only the peptide that increases the total probability or score of the corresponding protein (28). The Fido node performs a fast Bayesian inference in order to solve the protein inference problem. The recommended parameters for gamma, alpha and beta (0.5, 0.1 and 0.01) were used for each run. ProteinProphet (PP) (16) takes a pepXML file as input that contains peptides with associated probability scores. Different peptide identifications corresponding to the same protein are combined together to estimate the probability that their corresponding protein is present in the sample. The pepXML files were refined using PP's xinteract to correct the decoy annotations and FASTA file connection. Afterwards, PP was executed without any parameters except MINPROB0.05 to include only peptides with probability of at least 5% into the inference. MSBayesPro (26) is a Bayesian protein inference algorithm. Besides peptide probabilities derived from the spectrum scoring it also incorporates the peptide detectability from the *PTModel* node in the probabilistic model. ProteinLP (17) introduces the marginal probability of each identified peptide being present is known. The algorithm tries to find a minimal set of proteins while peptide probabilities should be as close to its known value as possible. Also ProteinLP does not need any further parameters.

The FDR q-values were calculated, based on the target-decoy approach, to control the false rates at the protein level (37). We employed the protein FDR q-value <= 0.01 threshold for all metrics except for the pseudo-ROC plots. Finally, in **(G)** the inference reports are generated, including both numbers and graphs (Supplemental File S1, section 8 and Supplemental Files S03-S14). For each search engine combination the number of FDR filtered PSMs is reported to give an overview of the identification step. Besides the target and decoy labels, all reference proteins in "gold standard datasets" were labeled to be true positives in the samples. A pseudo-ROC curve is generated with the number of true positives against the q-value of the FDR on protein level (28, 38). For all further metrics, the analyses were restricted to the high confidence proteins with a q-value below 0.01 or 1%.

2.2 Benchmark metrics

Benchmarking requires both a high-quality dataset/workflow and defined metrics to evaluate the improvements and potential pitfalls for these tools (39). We used a set of metrics based in previous studies to benchmark the inference algorithms and tools (24). The number of protein groups represents the first intuitive metric for a quick overview of the inference performance (20). We used the number of protein groups below the 1% FDR q-value on protein level for each inference algorithm (see Figure 5). A protein (ambiguity) group is an indistinguishable entity reported by an algorithms (40). In such groups, the sets of peptides overlap perfectly in the set of proteins from which they come. In addition, we studied the overlap of protein groups between all inference algorithms since the number of protein groups reported may be the same and yet the identities may be different. The proportions of mutually reported groups were calculated to gain deeper insight into the consensus of the reported protein groups (see Figure 4). It is furthermore possible for uniquely reported protein groups (i.e. groups reported by one inference algorithm alone) to distinguish whether the proteins in a group are reported in another combination of accessions as a group by any inference algorithm (light orange in plot) or whether they are truly uniquely reported (dark orange in plot). We additionally created Venn diagrams to visualize the overlap of the reported protein groups in a widely known way (see Figure 3a and Supplemental Files 3-14, sections 4 and 5).

Also, we used a heat-map to represent how many groups are shared by which reports (see supplemental Figures 4 in the supplemental reports). We studied the behavior of the number of reported protein groups along FDR q-values (0-5%) on protein level using pseudo-ROC curves (Figure 2 and 3) (28, 38). We also highlighted the true positive protein groups for the ground truth datasets (the yeast and the iPRG2008 dataset). Furthermore, we reported the number of identified peptides and peptide-spectrum matches by protein groups (see Figure 9 for peptides per protein). Finally, we plotted the number of reported proteins compared with the respective protein length to assess the performance of each algorithm retrieving proteins with short/long lengths (see section 13 in the Supplemental Files S3-S11). In Figure 8, we plotted comparisons of protein ranks by inference method for groups reported uniquely (i.e. groups which are not reported by any other method below the 1% protein FDR q-value threshold). All metrics for each search engine combination and dataset were plotted in the supplemental report files (Supplemental Files S3-S11). For all the current metrics we use the protein groups and protein sub-groups if

the inference algorithm reports them by default (Fido and ProteinLP). A protein sub-group is a protein group whose peptides are completely explained by another protein group (see an example in Section 10, Supplemental File S1).

2.3 Benchmark Datasets

In the present study we have tested four public datasets: the mouse lysate from the iPRG2008 study (http://www.abrf.org/research-group/proteome-informatics-research-group-iprg), a subset (the *070119-zl-mudpit07-1* files) of the "Gold Standard of Protein Expression in Yeast" also used by *Ramakrishnan et al.*, the lung cancer samples (LC1-LC12) of a more recent human dataset from the PRIDE repository (PXD000603) (41) and the HCD measurements of the histone enrichment study deposited in PRIDE under the identifier PXD001118 (42). For the iPRG2008 and the PXD00118 dataset the provided MGF files were used for identification, for the other two datasets the collections of spectra were converted to the mzML format using ProteoWizard (43). Tandem mass spectra were searched against appropriate protein sequence databases using the target/decoy approach (TDA) with three different search engines: X!Tandem (version Sledgehammer 2013.09.01.1), MS-GF+ (version beta v10089) and Mascot (version 2.5). The first two tools are not the most recent versions, but the ones shipped with the currently stable OpenMS version 2.0.

To analyze the influence of the database complexity in protein inference, each dataset was searched against three different databases: (i) UniProtKB/Swiss-Prot, (ii) Uniprot reference proteome, and (iii) Uniprot reference proteome containing known isoforms for each gene, in contrast to the first two, which contain only the longest isoform for each gene (Section 9, Supplemental Files S1). Only two databases were analyzed for the yeast dataset because the Swiss-Prot and the reference proteome sets are equal. The iPRG2008 dataset was additionally identified using the provided mouse database. The decoy databases were created with the DecoyDatabaseBuilder (27) by shuffling the protein sequences and appending them to the target database creating a concatenated target-decoy database. An exception was the provided *iPRG2008* database, where the provided target-decoy sequences with reversed decoys was used. We used the same search parameters wherever possible for each search engine, for the individual settings of each dataset (see Table 1). For the digestion of proteins to peptides a fully tryptic digestion was selected for the iPRG2008, yeast and PXD000603 datasets. For the histone dataset (PXD00118) the cleavage at lysine was masked by a fixed modification and therefore neglected. The

workflows, search engine results and all of the final results are available via ProteomeXchange and GitHub.

3. Results

3.1 General assessment of the protein inference algorithms

Running the aforementioned workflow, we analyzed 420 different protein lists due to the combination of the three different search engines, the five inference tools and the four datasets using ten different databases. We analyzed the number of FDR filtered PSMs for each single search engine and their combinations before performing any protein inference evaluation. The benefit of combining search engine results for spectrum identification has already been shown extensively in other publications (44, 45). It is generally accepted that search engines in combination yield more valid PSMs, especially in low-resolution fragment ion measurements (see section 1 in the Supplemental Files S3-S11). X!Tandem and MS-GF+ identified more PSMs than Mascot in almost all setups; the only exception was in the Swiss-Prot PXD000603 run, where X!Tandem was slightly outperformed by Mascot, and the UniProt proteome PXD001118 run. In the latter MS-GF+ alone was performing surprisingly suboptimal (X!Tandem reported 3.4, Mascot 2.7 times as many PSMs), due to relatively high-ranking decoy PSMs. The second biggest discrepancy between two single search engines was when the PXD000603 dataset was searched against the proteome set with isoforms, where MS-GF+ reported 1.55 times as many PSMs as Mascot. The average ratio between the lowest and highest single search engine was 1.62 (1.44 excluding the two prior mentioned outliers). Each combination of two search engines returned more than the respective single engines. The combination of all three engines yielded the most PSMs for each dataset, increasing the report of the best single result by 90% on average and ranging from 17% (iPRG2008 dataset with proteome database) to 173% (UniProt Proteome on the PXD001118 dataset). For all analyses in this work it must be considered that we inspected only the *Consensus ID* with the PEPMatrix algorithm provided by OpenMS for the combination of PSM results and the posterior error probabilities (PEPs) calculated by *IDPosteriorErrorProbability*.

Figure 2 shows pseudo-ROC curves of the number of reported target protein groups against the local protein FDR q-values for all datasets, the three search-engines combination and the respective Swiss-Prot database. The overall number of reported protein groups under a certain q-value varies slightly between the different inference

algorithms. Fido outperforms the other inference algorithms at 1% FDR q-value for the more complex datasets yeast and PXD000603 by 5.8% and 0.2% more protein groups respectively (see also further discussion of other metrics). However, all other approaches outperform Fido significantly on the iPRG2008 dataset. The main reason in this particular case is the highly unbalanced composition of target (34'127) versus decoy (332) PSMs. This resulted in much larger groups for target proteins reported by Fido, leading to reduced posterior probabilities of them, eventually boosting the ranks and therefore the q-values of decoys. The MSBayesPro results showed that the detectability algorithm implemented into *PTModel* did not perform as well as the algorithm from the original publication (Section 4, Supplemental Files S3 to S14). For this reason, the results from MSBayesPro were moved to Supplemental information; however, we will discuss in the manuscript some major drawbacks (but also advantages) of using detectability algorithms for protein inference.

3.2 Analysis of ground truth datasets: Yeast and iPRG2008

The reported protein groups for a given threshold is a basic metric to evaluate the performance of a given inference algorithm. However, it should be complemented with other metrics to label a protein inference superior to any other. In fact, it is more relevant to see whether the true protein groups are reported. There are several publicly available datasets containing ground truth data for peptides (46) and only relatively small protein datasets (47). Only the yeast dataset used can be considered as a complex mixture. We used three Venn diagrams for the reference set using Swiss-Prot database to examine the content of correctly identified proteins in the yeast dataset (Figure 3). We consider a protein group as true-positive if it contains at least one accession of the reference set of accessions, which are known to be in the sample. Figure 3a shows all the proteins identified in the yeast dataset by every inference algorithm without discrimination of true and false positives regarding the reference set. The yeast reference set contains 4,253 protein entries that are known to be in the sample (also validated by 2D-DIGE).

The number of reference proteins reported by Fido, PP, ProteinLP and PIA were 1193, 1152, 1149, 1095, respectively (Figure 3b). Fido identified 98, 44, and 41 more reference proteins than PIA, ProteinLP and PP, respectively. Most of the proteins uniquely identified by Fido (78%) are included in the reference set (Figure 3a and b). However, the inference algorithms are also reporting proteins that are not included in the reference set (Figure 3c). These proteins can be considered as potential false positives or new evidences that were not labeled properly. The number of proteins reported by Fido, PP, ProteinLP and PIA that

are not included in the reference set were 44, 29, 29, and 27, respectively. These proteins have more probability to be false positives if they are reported by only one inference algorithm. A close inspection shows that Fido uniquely reports the highest number of protein groups (12 groups) that are not in the reference set (Figure 3c). In contrast, ProteinProphet, ProteinLP and PIA uniquely reported 0, 1 and 2 proteins that are not in the reference set.

Also, we studied the protein clusters for the iPRG2008 dataset. A cluster is a set of proteins with partially shared peptides and in the iPRG2008 study a certain number of these protein groups should be reported as true positives (see the original study on http://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm). This allows the calculation of numbers for false positives (FP, i.e. too many reported protein groups in a cluster), false negatives (FN, too few reported groups in a cluster) and true positives (TP, exact number of reported group in a cluster). Furthermore, the total number of reported protein groups is of interest for this analysis. Using this information, the results can be ordered on that the highest number of TP, the fewest FN, the fewest FP and last the total number of reported proteins with 1% FDR (see Supplemental File S2 – iPRG2008 analysis). PIA reports the most TP with fewest FN for all search engine combinations, but the same numbers for the Mascot search alone.

ProteinProphet generally reports the TP more conservatively but is very good in controlling the FP. This result is correlated with the results in the yeast dataset where ProteinProphet does not uniquely report any protein that was not labeled in the reference set. On the other hand, it misses many proteins per cluster (higher FN rate than the before mentioned algorithms). Fido yields the highest number of FPs meaning that it reports too many separate protein groups per assumed cluster and relatively few TP.

3.3 Evaluation of the overlap amount inference algorithm

The next inspected metric is the number of reported protein groups as well as the fraction of protein groups reported by other inference algorithms (Figure 4). The plots show the numbers of protein groups reported for the iPRG2008 (a-c), yeast (d, e), PXD000603 (f-h) and PXD001118 (i-k) datasets for the combination of the three search engines. First, we analyzed the impact of the database complexity used for the identification by comparing results with Swiss-Prot (Figure 4 a, d, f, i), UniProt proteome (Figure 4 b, d, g, j) and Uniprot proteome with isoforms (Figure 2 c, e, h, k). It is important to note that the yeast Swiss-Prot and UniProt proteomes are identical. The overlap of protein groups reported by

all inferences (teal) is bigger when using the least complex database (Swiss-Prot) for identification.

Figure 4 shows that Fido increases more than any other algorithm the number of uniquely reported proteins when more complex databases are used (UniProt proteomes). This is mainly because Fido reports sub-protein groups (groups whose peptides are contained in another group, section 10 in Supplemental File S1). In contrast, PIA, ProteinLP and PP seemed more robust against changes in the database complexity. PIA and ProteinLP tended to report the most groups on more complex databases (e.g. on PXD000603 PIA reports 16% more groups than PP for the UniProt proteome without isoforms and 15% more for the proteome with isoforms). On the iPRG2008 dataset, PIA and ProteinLP reported on average 42% and 40% more than PP, respectively. Here, Fido and PP reported similar numbers of protein groups for the Swiss-Prot dataset. However, on the other two databases (more complex ones) Fido reported 33% less than PP. In less complex databases Fido performs better than the other inference algorithms (e.g. an average of 5% more protein groups than PP on the yeast dataset). These results were also visualized using the more common Venn diagrams (Figure 3a and Supplemental Files 3-14). Both analyses also show that even if the actual numbers of reported protein groups may be similar between the inference algorithms (Figure 3a and Figure 4d), the actually reported groups and their overlaps differ between the algorithms.

The number of reported groups is increased when more search engines are combined compared with the results of single search engines (Figure 5). For each single dataset a pattern for the ratios between the inference algorithms and search engine combination is observed (e.g. Figure 5 showed that Fido reports most protein groups, followed by PIA and PP, then ProteinLP). However, the search engine and the inference algorithm should be selected carefully. For example in the PXD000603 dataset plotted in Figure 5, when X!Tandem alone is used with Fido the number of reported groups is decreased with respect to the other combinations.

Fido and the underlying generative (Bayesian) model relies on reasonable probabilities for the observed peptides, which are besides the three model parameters the only input to the algorithm. Although being relatively robust for multiple types and shapes of distributions of these input probabilities, even with parameter estimation, it cannot correct for heavily ill-shaped ones (this is similar to the saying “garbage in, garbage out”). This happens to be the case with unfiltered X!Tandem results as described in Figure 6 c, f. If we have a look at the different score distributions of the three used search engines: Mascot-EValue, MS-GF+-SpectralEValue, and X!Tandem-EValue in Figure 6 a-c, we can highlight different

points. Whereas Mascot and MS-GF+ yield distributions of an expected shape (a large peak in the lower region of the scores corresponding to false-positive hits and a flat right tail coming from an assumed second true-positive distribution), X!Tandem has a different distribution. It is not well-suited for the estimation of posterior probabilities with the fitting-based algorithm used by algorithms like IDPosteriorProbability in OpenMS. The distribution of the X!Tandem-EValues actually poses two problems when fitting a mixture of two distributions to it: Firstly, due to the bimodal nature of the distribution of the X!Tandem-EValues coming from an unexplainable valley at EValues around 1.0, the used expectation-maximization algorithm (EM) will try to distinguish between the two peaks and uses the second flat distribution of true-positives to explain parts of the second peak's density. In extreme cases where the algorithm fully picks up the second peak in the bad scoring region as a true-positive peak, it will yield wrongly interpreted posterior probabilities. Secondly, even if the impact of this second peak on the true-positive distribution is marginal, the density at bad values is very high compared to the remaining scores, resulting in a very strong and importantly narrow distribution for the false-positives. This results in the probability of a value to be generated by the false-positive distribution starting to be near zero at relatively low scores, which leads to many posterior peptide probabilities of 1.0 (Figures 6 d-f).

For Fido, one should be careful with inputs of extreme probabilities for the peptides, such as 0.0 or 1.0. In a Bayesian model this strictly excludes every combination not using this peptide although other information suggests differently (which is especially is a problem when assigning a probability of 1.0 to so many peptides as in the case for X!Tandem). A second problem is the lack of discriminative power between equal scores. Since the parameter estimation of Fido tries to create well-calibrated and well-discriminating results at the same time, this creates an issue. Extreme values for peptide probabilities as inputs are likely to generate extreme probabilities for the proteins. If more than a minor percentage of the proteins are assigned probabilities of 1.0 and these include decoy proteins, the first q-value-cutoff considered in a corresponding receiver operating curve (ROC) results in an uninformative straight line in the upper part of the curve covering all proteins of probability 1.0. This, in turn, makes it hard to compare it to other methods and limits the usefulness of the parameter estimation.

Interestingly, the inference of Fido in some runs that include results from Mascot alone is significantly superior regarding numbers of protein groups to all other inferences. Additionally, this can be seen on the combination of Mascot and MS-GF+ results in the iPRG2008 dataset with the provided database. This effect can be explained by the fact

that the local FDR and the FDR q-value on protein level differ under certain circumstances (Figure 7). Under specific conditions the local FDR (and therefore the q-value of all preceding elements in a sorted report) returns to a low value, after increasing steadily due to several reported decoys. If during this increase and decrease of the local FDR many targets are reported the respective pseudo-ROC shows a step or a peak (if this occurs at the end of the list). The effect can be seen in several of the created pseudo-ROC curves (section 5 in Supplemental Files S3-S11) for Fido and ProteinLP. Though except for only a few combinations this effect occurs on q-values exceeding the threshold of 0.01 (i.e. somewhere between 0.01 - 0.05, e.g. in Figure 7 at 0.022). For the analyses we used the q-value, as it is currently a widely accepted method. This behavior shows that a method controlling both FDR q-value and local FDR might be more applicable.

An evaluation of the ranks of the uniquely reported protein groups sorted by probability/score revealed some unintuitive distributions (Figure 8). If we assume that the top ranking protein groups are the most valid, an intuitive distribution should represent protein groups that are not reported in consensus (unique) at the end of the reported lists with low scores. For PIA almost all uniquely reported groups are at the end of the list. On the other hand, Fido and ProteinLP distributed the unique groups over the complete range of indices, with a tendency to the end of the list. The most extreme case is ProteinProphet that reports its unique groups at the very beginning. This reveals that the intuitive assumption that the majority of the uniquely reported groups are located at the end of the report is not correct.

3.4 Impact of multiple search engines

As a further quality assessment metric for reliable identifications, we analyzed the number of peptides per protein groups for each protein inference algorithm (46). The numbers of peptides per protein group were plotted in a heatmap-like way for the results of the PXD000603 dataset with the Swiss-Prot database (Figure 9). Independently of the inference algorithm, most protein groups are reported with few peptides and only a small fraction is represented by ten or more peptides. In the plotted dataset, the number of inferred protein groups with ten or more peptides from the single search engines' results with PIA, Fido and ProteinProphet are on average 5.7% (ranging from 5.1% - 6.7%) of all reported groups. Using the results from multiple search engines increases these groups in average to 6.5%, though for the X!Tandem-Fido combination the percentage is decreased by 0.2%. The actual numerical values are always increased by at least eight protein

groups with ten or more peptides. To assess the bias introduced by the reporting of sub-groups by Fido, we additionally analyzed all metrics after removing these sub-groups. For this, all protein groups, whose peptides were a subset of one or more other reported protein groups, were removed from the report, before calculating the FDR. This generally removed a big fraction of the groups, which were unique for Fido when not removing the sub-groups (Supplemental Files S3-14, sections 6 and 7). All analyses (except the spectrum identifications by Mascot) were performed on a laptop computer with an Intel(R) Core(TM) i7-4800MQ and 16 GB RAM.

4. Discussion

We have evaluated in detail the performance of difference inference algorithms using four different datasets and a set of well-define metrics. MSBayesPro needs detectability predictions for each peptide as an input of the inference algorithm. These values can only be calculated using the results of preceding experiments or estimated using algorithms like the *PTModel*. Both modeling approaches have drawbacks when experimenting with analytical methods (e.g., enrichment, different fractionation methods) for which there are no preceding reference results. In these cases, these inference algorithms will not perform well (see Supplemental Files S3-S14). Prediction of detectability increases the running time and the predicted model (MSBayesPro) is not available making difficult the integration into bioinformatics pipelines. However, different authors have demonstrated theoretically that the use of properties such as the isoelectric point, retention time or MS1 information can be used to improve the inference and identification process (15, 19, 47).

A uniqueness of the Fido implementation in OpenMS is that it requires a decoy database to find the best values of the parameters (α, β, and γ- the prior for the presence of proteins) by combining an ROC optimization (in a supervised manner) with FDR estimation. If the input data is biased as explained before (see Results section), this optimization step leads to suboptimal results. Fido is a very fast implementation with a small memory footprint. Meanwhile it is integrated into OpenMS and thus can easily be used in bigger workflows. Fido reports in most of the analysis more proteins that the others algorithms. However, its performance relies on multiple factors such as PSM score distribution, target/decoy database distributions, and redundancy of the database (isoforms). These factors make the results of Fido less constant than other algorithms and demand more benchmark and tuning of the pipeline (48).

ProteinLP and Fido have as a main concept not the parsimony of peptides or spectra but

the probability of proteins' occurrences given the PSM or peptide probabilities. By design, they report sub-proteins if the respective probabilities are sufficiently high. This difference with the parsimonious approaches such as PIA or ProteinProphet should be evaluated when choosing an inference algorithm. If many sub-protein groups were reported (e.g. in the data in Figure 4g and h, which shows many unique groups for Fido), the FDR q-value did increase due to reported decoy sub-groups as well, and the total number of reported protein groups decreased. In some combinations of databases, datasets and search engines the number of reported groups rises significantly above the reports of the other inference algorithms. This is due to an effect of the protein FDR q-value and the local protein FDR values (Figure 7). During this effect, the local FDR may exceed a given threshold significantly and drop below it after reporting many target proteins. This leads into steps in a corresponding pseudo-ROC curve and suggests more advanced methods than the q-value or local FDR alone, either combining these or algorithms like Mayu (49).

ProteinProphet has a very low memory imprint and thus scales to process big datasets. It is more conservative in reporting protein groups than other approaches reporting less false-positives in the reference datasets. It should be observed in the Venn diagrams and the iPRG2008 cluster analysis that it reports a low amount of unique proteins reducing the possibility of false positive identifications (Figure 3). One of its main strength is the integration into the Trans-Proteomic Pipeline, which incorporates multiple other search engines like SEQUEST (50), Comet (51) or Mascot. ProteinLP consumes the highest amount of memory and time. Although, it performs well for most of the datasets; in all the metrics it is out performed by other inference options.

Among parsimonious approaches, PIA mostly reports more target protein groups than ProteinProphet in the studied datasets. PIA consumes a relatively large amount of memory analyzing a not-FDR-filtered or very big dataset. It reports high numbers of confident protein groups and like other parsimonious approaches it is relatively fast. However, ProteinProphet yields in less false positive identifications when it was used to analyzed the ground truth datasets (section 3.2). For both of these datasets Fido reports more proteins than other algorithms but also more possible false positives (proteins that are not labeled in the reference set).

A feature that should be considered when choosing an inference algorithm is the robustness when using complex databases for spectrum identification. While PIA, ProteinLP and ProteinProphet were only slightly affected by this, Fido and MSBayesPro reported significantly fewer valid protein groups when using more complex databases at 1% FDR q-value (Supplemental Files S5, S8, S11 and S14). Figure 9 presented a

drawback for parsimonious approaches that reported more single peptide proteins compared with probabilistic models such as Fido and ProteinLP. The “two peptides”-rule, applied quite often in proteomics to control protein false discoveries (46), can affect and change the results of the experiment depending of the inference algorithm used.

Also, the interoperability and ease of use of an inference algorithm will influence its application by a user. All analyzed algorithms except PIA need special non-standard input formats. It would be very beneficial for users, if standard formats like mzIdentML or even the search engines’ default result files could be used as input. PIA also has the advantage that it works with spectrum identifications coming from various file formats, search engines and bioinformatics workflows (28, 35). It is the only implementation that works natively with standard file formats such as mzIdentML and mztab and it is integrated in PRIDE Inspector Toolsuite for the analysis of public datasets (36). It also can be fully integrated into OpenMS pipelines, when using KNIME as workflow environment. ProteinProphet uses the pepXML format that does have converters for many search engine results, but is well known in the proteomics community (52). PIA and Fido are the only algorithms at the moment which can be fully integrated into an OpenMS workflow and thus inside KNIME.

5. Conclusion

We introduced a workflow that uses three search engines and five open-source and generally applicable protein inference algorithms for the fair and in-depth comparison of protein inference results. The workflow and inference methods were tested on four datasets with different complexities of protein databases. While there is no explicit best inference algorithm, different considerations for choosing a tool can be given.

The analysis of identifications using protein databases with varying complexity shows some algorithm specific results. Due to the occurrence of more decoys, all inference algorithms report fewer groups when more complex databases are used. The numbers of reported groups by PIA and ProteinProphet are much less dependent on the database complexity than Fido. If the detection of specific isoforms is important in the scientific context, this could compensate for slightly less protein groups.

Depending on the underlying report, the FDR q-value may be not sufficient to filter for good identifications. This is especially the case if the local FDR exceeds a given threshold for a big part of the report, but finally drops below the threshold again. To improve on this problem, other strategies should be developed. Another very interesting comparison of

protein inference algorithms and the fundamental search engines would be how the reported isoforms or splice variants matched on a gold standard dataset, containing the knowledge of isoforms and splice variants. This could not be tested thoroughly, due to the lack of current publicly available datasets at the time of writing. We also expect that more “gold standard” datasets in larger scale will lead to a fairer comparison of protein inference algorithms, than the usage of target decoy strategies alone.

Using Fido and PIA results in more proteins than the other approaches. However, ProteinProphet has a more conservative approach and reports less false positives in all the analysis. The parsimonious approaches are less dependent on the search engine scores distribution than Fido. Although being relatively robust for multiple types and shapes of distributions of the input probabilities, even with parameter estimation, Fido cannot correct for heavily ill-shaped distributions like some results from X!Tandem in the discussed analyses.

Furthermore, the created workflow could easily be adjusted to benchmark different protein inference algorithms in the future and thus gives a fair framework for testing of protein inference algorithms in general. Overall, one possibility for future improvements to the inference methods could be the use of additional information during the inference process, especially, since data from the MS1 level (e.g. deviation from predicted retention time or intensities) is readily available in almost all experiments (15, 19, 47). Another source of information could come from technical replicates, where agreeing identifications may boost the confidence of its correctness. It might additionally be helpful to view protein inference as a (binary) subtask of quantification to leverage knowledge from this related problem.

Availability

All of the analyzed protein inference algorithms are available as KNIME nodes and can be used together with OpenMS workflows to yield protein identification lists. The designed workflow also allows the exchange of the tested inference algorithms and thus comprehensive benchmarking of new implementations. The plugins for the newly developed nodes and the complete workflow are available as open source on https://github.com/KNIME-OMICS. The workflows, search engine results and all of the final results are available via ProteomeXchange with the identifiers PXD003066, PXD003067, PXD003068, PXD003072, PXD003953, PXD003954, PXD003955, PXD003956, PXD003957, PXD003958, PXD003959, PXD003960.

Acknowledgements

E.A. was supported by a grant from the Boehringer Ingelheim Fonds, J. U. and T. S. are funded by the BMBF grant de.NBI - German Network for Bioinformatics Infrastructure (FKZ 031 A 534A resp. FKZ 031 A 535A); funding of ME is related to PURE and Valibio, Projects of North Rhine-Westphalia; Y.P-R. is supported by the BBSRC ‘PROCESS’ grant [BB/K01997X/1].

References

1. Altelaar, A. F., Munoz, J., and Heck, A. J. (2013) Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature reviews. Genetics* 14, 35-48
2. Perez-Riverol, Y., Wang, R., Hermjakob, H., Muller, M., Vesada, V., and Vizcaino, J. A. (2014) Open source libraries and frameworks for mass spectrometry based proteomics: a developer's perspective. *Biochimica et biophysica acta* 1844, 63-76
3. Betancourt, L. H., De Bock, P. J., Staes, A., Timmerman, E., Perez-Riverol, Y., Sanchez, A., Besada, V., Gonzalez, L. J., Vandekerckhove, J., and Gevaert, K. (2013) SCX charge state selective separation of tryptic peptides combined with 2D-RP-HPLC allows for detailed proteome mapping. *Journal of proteomics* 91, 164-171
4. Ramos, Y., Gutierrez, E., Machado, Y., Sanchez, A., Castellanos-Serra, L., Gonzalez, L. J., Fernandez-de-Cossio, J., Perez-Riverol, Y., Betancourt, L., Gil, J., Padron, G., and Besada, V. (2008) Proteomics based on peptide fractionation by SDS-free PAGE. *Journal of proteome research* 7, 2427-2434
5. Ramos, Y., Garcia, Y., Perez-Riverol, Y., Leyva, A., Padron, G., Sanchez, A., Castellanos-Serra, L., Gonzalez, L. J., and Besada, V. (2011) Peptide fractionation by acid pH SDS-free electrophoresis. *Electrophoresis* 32, 1323-1326
6. Chick, J. M., Kolippakkam, D., Nusinow, D. P., Zhai, B., Rad, R., Huttlin, E. L., and Gygi, S. P. (2015) A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature biotechnology* 33, 743-749
7. Martens, L. (2011) Bioinformatics challenges in mass spectrometry-driven proteomics. *Methods in molecular biology* 753, 359-371
8. Perez-Riverol, Y., Hermjakob, H., Kohlbacher, O., Martens, L., Creasy, D., Cox, J., Leprevost, F., Shan, B. P., Perez-Nueno, V. I., Blazejczyk, M., Punta, M., Vierlinger, K., Valiente, P. A., Leon, K., Chinea, G., Guirola, O., Bringas, R., Cabrera, G., Guillen, G., Padron, G., Gonzalez, L. J., and Besada, V. (2013) Computational proteomics pitfalls and challenges: HavanaBioinfo 2012 workshop report. *Journal of proteomics* 87, 134-138
9. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551-3567
10. Kim, S., and Pevzner, P. A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature communications* 5, 5277
11. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20, 1466-1467
12. Nesvizhskii, A. I. (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of proteomics* 73, 2092-2123
13. Villavicencio-Diaz, T. N., Rodriguez-Ulloa, A., Guirola-Cruz, O., and Perez-Riverol, Y. (2014) Bioinformatics tools for the functional interpretation of quantitative proteomics results. *Current topics in medicinal chemistry* 14, 435-449
14. Serang, O., MacCoss, M. J., and Noble, W. S. (2010) Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *Journal of proteome research* 9, 5346-5357
15. Perez-Riverol, Y., Sanchez, A., Ramos, Y., Schmidt, A., Muller, M., Betancourt, L., Gonzalez, L. J., Vera, R., Padron, G., and Besada, V. (2011) In silico analysis of accurate proteomics, complemented by selective isolation of peptides. *Journal of proteomics* 74, 2071-2082
16. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Analytical chemistry* 75, 4646-4658
17. Huang, T., and He, Z. (2012) A linear programming model for protein inference problem in shotgun proteomics. *Bioinformatics* 28, 2956-2962
18. Audain, E., Sanchez, A., Vizcaino, J. A., and Perez-Riverol, Y. (2014) A survey of molecular descriptors used in mass spectrometry based proteomics. *Current topics in medicinal chemistry* 14, 388-397
19. Perez-Riverol, Y., Audain, E., Millan, A., Ramos, Y., Sanchez, A., Vizcaino, J. A., Wang, R.,

Muller, M., Machado, Y. J., Betancourt, L. H., Gonzalez, L. J., Padron, G., and Besada, V. (2012) Isoelectric point optimization using peptide descriptors and support vector machines. *Journal of proteomics* 75, 2269-2274
20. Huang, T., Wang, J., Yu, W., and He, Z. (2012) Protein inference: a review. *Briefings in bioinformatics* 13, 586-614
21. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research* 10, 1794-1805
22. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* 26, 1367-1372
23. Searle, B. C. (2010) Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics* 10, 1265-1269
24. Claassen, M., Reiter, L., Hengartner, M. O., Buhmann, J. M., and Aebersold, R. (2012) Generic comparison of protein inference engines. *Molecular & cellular proteomics : MCP* 11, O110 007088
25. Li, Y. F., and Radivojac, P. (2012) Computational approaches to protein inference in shotgun proteomics. *BMC bioinformatics* 13 Suppl 16, S4
26. Li, Y. F., Arnold, R. J., Tang, H., and Radivojac, P. (2010) The importance of peptide detectability for protein identification, quantification, and experiment design in MS/MS proteomics. *Journal of proteome research* 9, 6288-6297
27. Reidegeld, K. A., Eisenacher, M., Kohl, M., Chamrad, D., Korting, G., Bluggel, M., Meyer, H. E., and Stephan, C. (2008) 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* 8, 1129-1137
28. Uszkoreit, J., Maerkens, A., Perez-Riverol, Y., Meyer, H. E., Marcus, K., Stephan, C., Kohlbacher, O., and Eisenacher, M. (2015) PIA: An Intuitive Protein Inference Engine with a Web-Based User Interface. *Journal of proteome research* 14, 2988-2997
29. Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., and Wiswedel, B. (2008) *KNIME: The Konstanz information miner*, Springer
30. Meyer-Arendt, K., Old, W. M., Houel, S., Renganathan, K., Eichelberger, B., Resing, K. A., and Ahn, N. G. (2011) IsoformResolver: A peptide-centric algorithm for protein inference. *Journal of proteome research* 10, 3060-3075
31. Sturm, M., Bertsch, A., Gropl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Zerck, A., Reinert, K., and Kohlbacher, O. (2008) OpenMS - an open-source software framework for mass spectrometry. *BMC bioinformatics* 9, 163
32. Bertsch, A., Gropl, C., Reinert, K., and Kohlbacher, O. (2011) OpenMS and TOPP: open source software for LC-MS data analysis. *Methods in molecular biology* 696, 353-367
33. Nahnsen, S., Bertsch, A., Rahnenfuhrer, J., Nordheim, A., and Kohlbacher, O. (2011) Probabilistic consensus scoring improves tandem mass spectrometry peptide identification. *Journal of proteome research* 10, 3332-3343
34. Pfeifer, N., Leinenbach, A., Huber, C. G., and Kohlbacher, O. (2007) Statistical learning of peptide retention behavior in chromatographic separations: a new kernel-based approach for computational proteomics. *BMC bioinformatics* 8, 468
35. Perez-Riverol, Y., Uszkoreit, J., Sanchez, A., Ternent, T., Del Toro, N., Hermjakob, H., Vizcaino, J. A., and Wang, R. (2015) ms-data-core-api: an open-source, metadata-oriented library for computational proteomics. *Bioinformatics* 31, 2903-2905
36. Perez-Riverol, Y., Xu, Q. W., Wang, R., Uszkoreit, J., Griss, J., Sanchez, A., Reisinger, F., Csordas, A., Ternent, T., Del-Toro, N., Dianes, J. A., Eisenacher, M., Hermjakob, H., and Vizcaino, J. A. (2016) PRIDE Inspector Toolsuite: Moving Toward a Universal Visualization Tool for Proteomics Data Standard Formats and Quality Assessment of ProteomeXchange Datasets. *Molecular & cellular proteomics : MCP* 15, 305-317
37. Serang, O., and Kall, L. (2015) Solution to Statistical Challenges in Proteomics Is More Statistics, Not Less. *Journal of proteome research* 14, 4099-4103
38. Perez-Riverol, Y., Sanchez, A., Noda, J., Borges, D., Carvalho, P. C., Wang, R., Vizcaino, J. A., Betancourt, L., Ramos, Y., Duarte, G., Nogueira, F. C., Gonzalez, L. J., Padron, G., Tabb, D. L., Hermjakob, H., Domont, G. B., and Besada, V. (2013) HI-bone: a scoring system for identifying

phenylisothiocyanate-derivatized peptides based on precursor mass and high intensity fragment ions. *Analytical chemistry* 85, 3515-3520
39. (2015) The difficulty of a fair comparison. *Nature methods* 12, 273
40. Seymour, S. L., Farrah, T., Binz, P. A., Chalkley, R. J., Cottrell, J. S., Searle, B. C., Tabb, D. L., Vizcaino, J. A., Prieto, G., Uszkoreit, J., Eisenacher, M., Martinez-Bartolome, S., Ghali, F., and Jones, A. R. (2014) A standardized framing for reporting protein identifications in mzIdentML 1.2. *Proteomics* 14, 2389-2399
41. Ahn, J. M., Kim, M. S., Kim, Y. I., Jeong, S. K., Lee, H. J., Lee, S. H., Paik, Y. K., Pandey, A., and Cho, J. Y. (2014) Proteogenomic analysis of human chromosome 9-encoded genes from human samples and lung cancer tissues. *Journal of proteome research* 13, 137-146
42. Yuan, Z. F., Lin, S., Molden, R. C., and Garcia, B. A. (2014) Evaluation of proteomic search engines for the analysis of histone modifications. *Journal of proteome research* 13, 4470-4478
43. Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* 24, 2534-2536
44. Shteynberg, D., Nesvizhskii, A. I., Moritz, R. L., and Deutsch, E. W. (2013) Combining results of multiple search engines in proteomics. *Molecular & cellular proteomics : MCP* 12, 2383-2393
45. Jones, A. R., Siepen, J. A., Hubbard, S. J., and Paton, N. W. (2009) Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics* 9, 1220-1229
46. Omenn, G. S., Lane, L., Lundberg, E. K., Beavis, R. C., Nesvizhskii, A. I., and Deutsch, E. W. (2015) Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification. *Journal of proteome research* 14, 3452-3460
47. Moruz, L., Hoopmann, M. R., Rosenlund, M., Granholm, V., Moritz, R. L., and Kall, L. (2013) Mass fingerprinting of complex mixtures: protein inference from high-resolution peptide masses and predicted retention times. *Journal of proteome research* 12, 5730-5741
48. Serang, O. (2013) Concerning the accuracy of Fido and parameter choice. *Bioinformatics* 29, 412
49. Reiter, L., Claassen, M., Schrimpf, S. P., Jovanovic, M., Schmidt, A., Buhmann, J. M., Hengartner, M. O., and Aebersold, R. (2009) Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Molecular & cellular proteomics : MCP* 8, 2405-2417
50. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* 5, 976-989
51. Eng, J. K., Jahan, T. A., and Hoopmann, M. R. (2013) Comet: an open-source MS/MS sequence database search tool. *Proteomics* 13, 22-24
52. Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nature biotechnology* 22, 1459-1466

Table 1. The datasets and search engine settings used in this work.

Datasets	URL	Instrument	Fragmentation method	Peptide/fragment tolerance	Modifications
iPRG2008	http://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm	3200 QTRAP	CID	0.45 Da 0.45 Da	**Fixed**: iTRAQ 4-plex (K, N), Methylation (C) **Variable**: Oxidation (M) **Cleavage:** [KR]|{P}
Yeast Gold Dataset	http://www.marcottelab.org/MSdata/Data_02	LTQ Orbitrap	CID	25 ppm 0.5 Da	**Variable**: Oxidation (M) **Cleavage:** [KR]|{P}
PXD000603	http://www.ebi.ac.uk/pride/archive/projects/PXD000603	LTQ Orbitrap XL	CID	10 ppm 0.8 Da	**Fixed**: Carbamidomethyl (C) **Variable**: Oxidation (M) **Cleavage:** [KR]|{P}
PXD001118	http://www.ebi.ac.uk/pride/archive/projects/PXD001118	LTQ Orbitrap Velos	HCD	10 ppm 0.02 Da	**Fixed**: Propionyl (N-Term and K) **Enzyme:** [R]|{P} (K is blocked by modification)

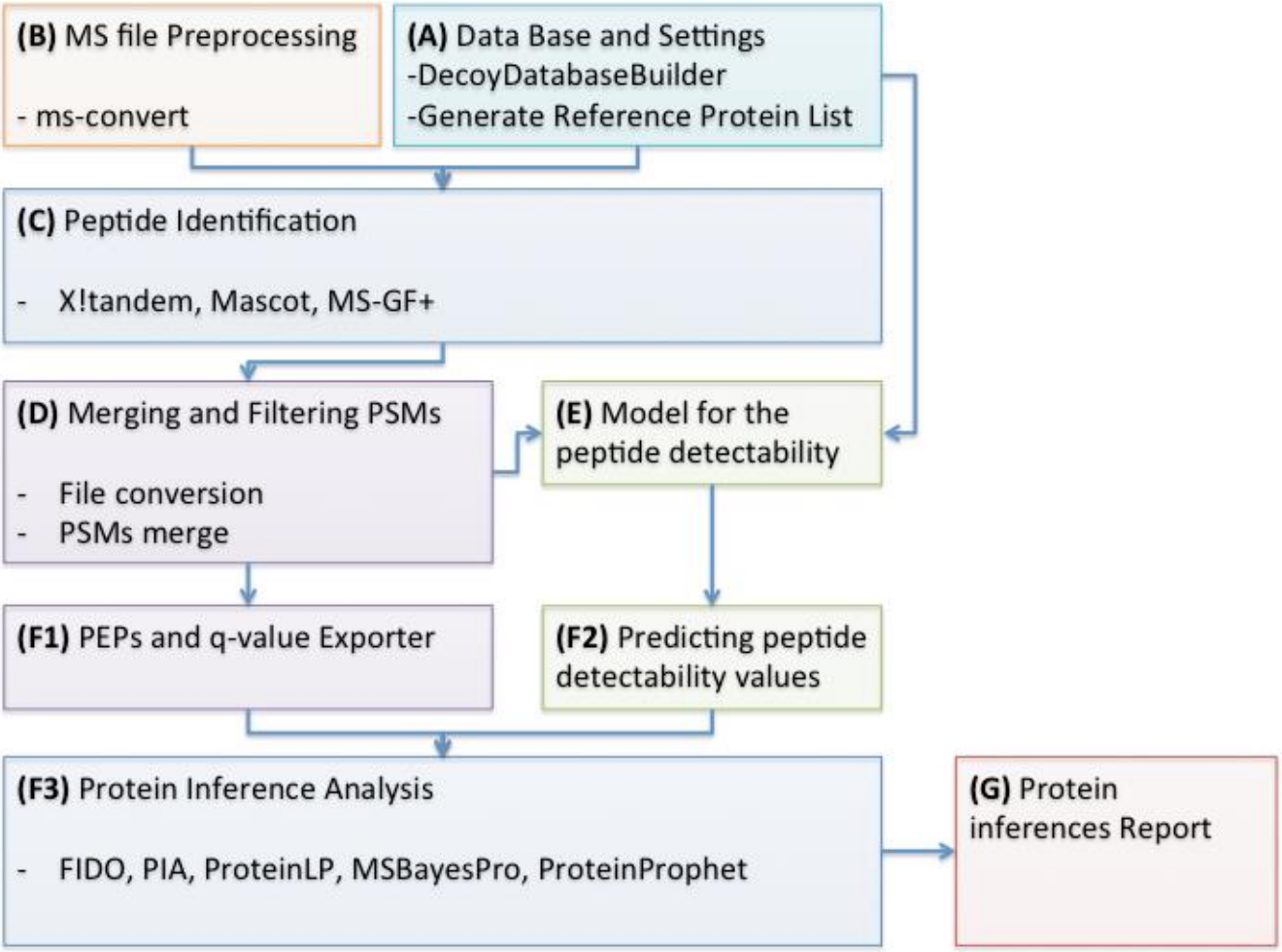

Figure 1: Simplified representation of the workflow used for the peptide identification and protein inference in KNIME. As input of the workflow, the raw MS data in mzML format is used; the output consists of graphs and numbers, as well as a complete report of the analyzed protein inferences. This workflow can be split into seven different stages A-G. (A) Settings and database, import of protein knowledge of gold standard datasets, (B) spectrum pre-processing, (C) peptide identification, (D) merging of PSMs (E) creating a model for peptide detectability, (F) protein inferences, (G) calculating numbers and graphs of the inferences.

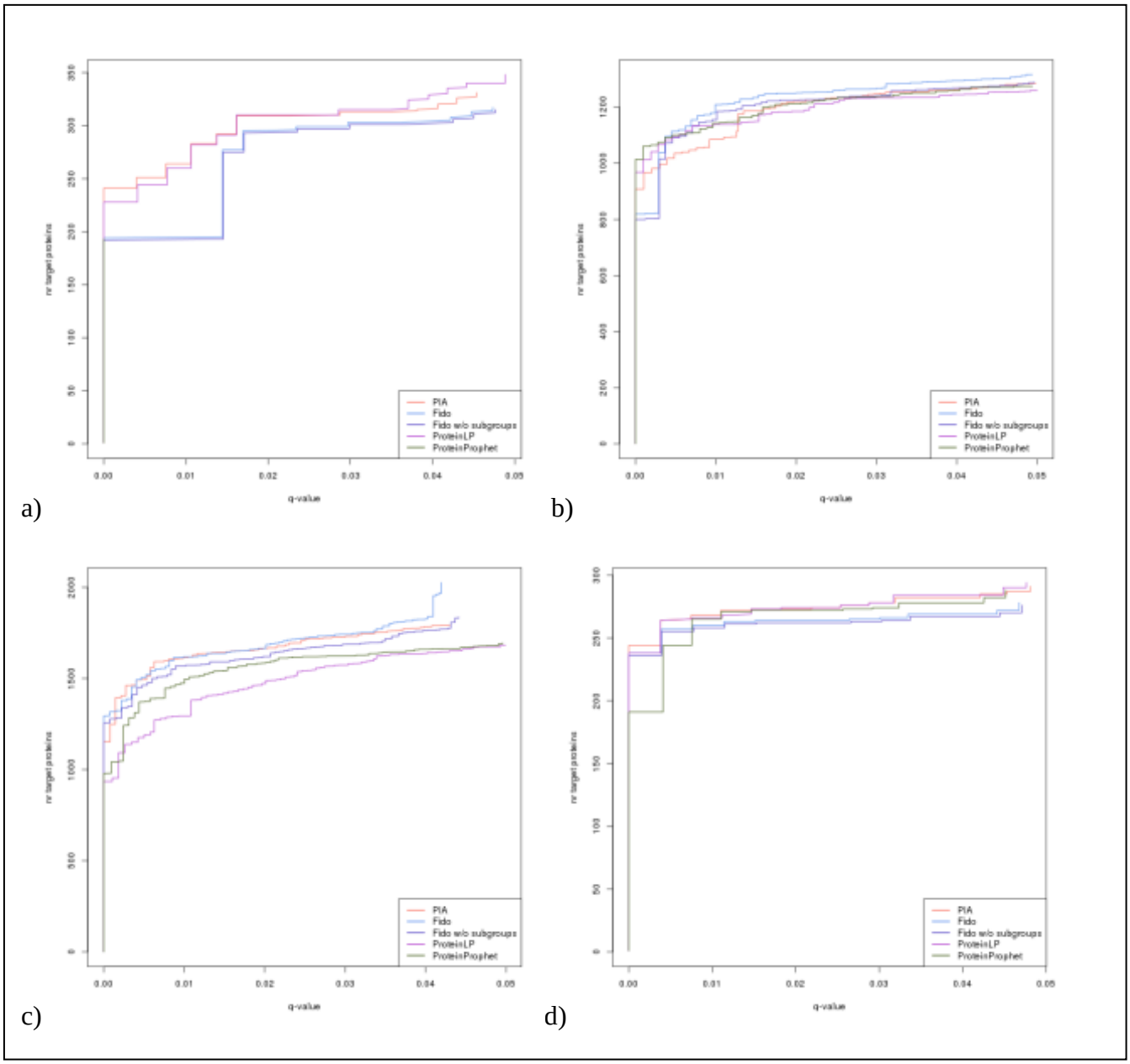

Figure 2: Pseudo-ROC curves show the number of reported protein groups against the FDR q-value for the four datasets using the Swiss-Prot database and the combination of the three search engines: a) iPRG2008, b) yeast-, c) PXD000603, and d) PXD001118 dataset. The plots indicate that the main trend is similar for all inference algorithms. Depending on the dataset, different algorithms perform worse than others for certain q-value ranges, like Fido in the iPRG2008 dataset, PIA in the yeast dataset and ProteinProphet in the PXD000603 dataset.

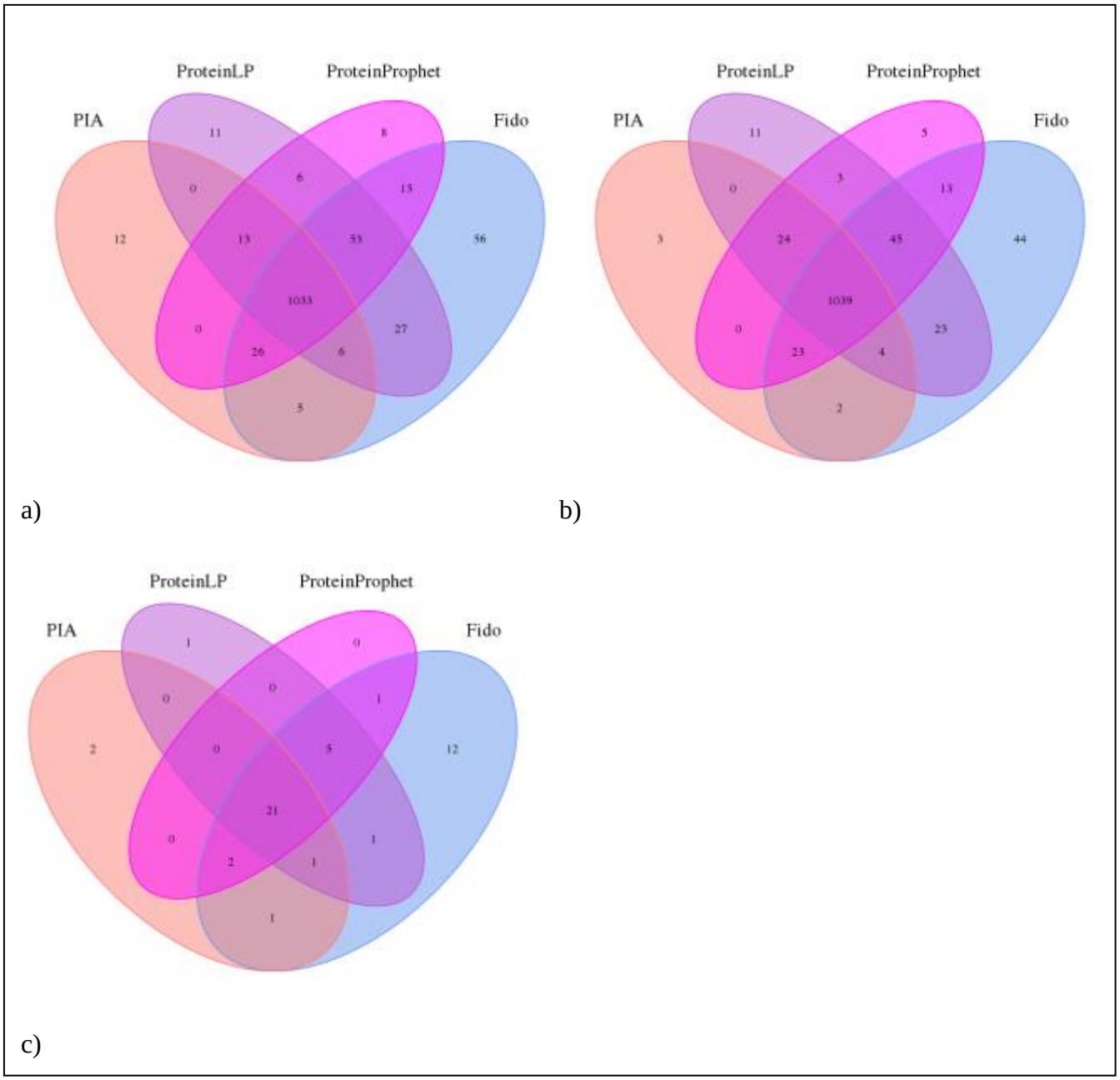

Figure 3: Venn diagrams plotting the number of reported protein for each inference algorithm at 1% FDR for the ground truth yeast dataset using the Swiss-Prot database. **(a)** Number of reported proteins for the yeast dataset (not only the labeled proteins). **(b)** Number of reported proteins included in the reference set. **(c)** Number of reported proteins that are not in the reference set.

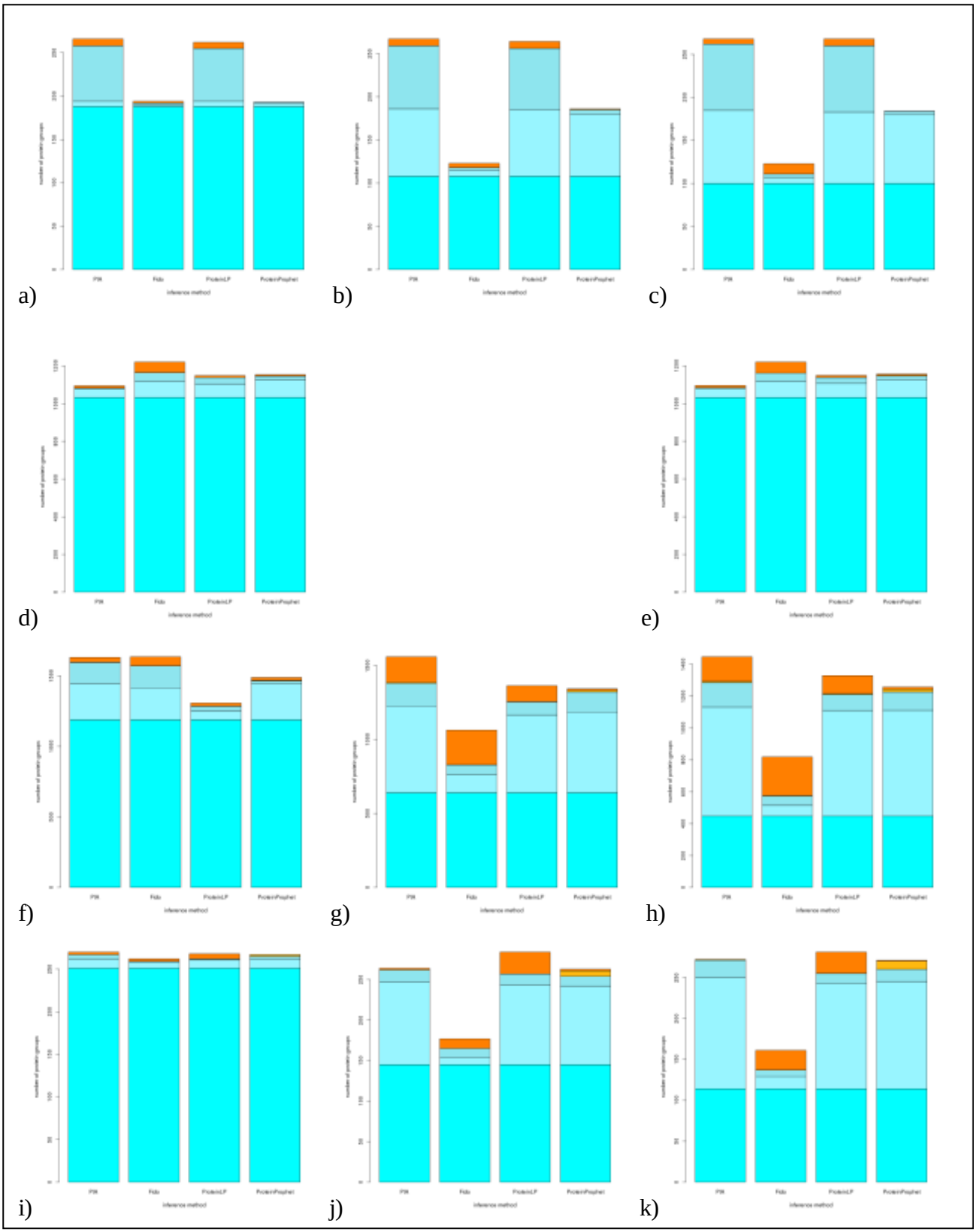

Figure 4: Number of protein groups reported for the datasets using different inference algorithms and databases. Number of protein groups under a 1% FDR q-value for the iPRG2008 (a-c), yeast (d, e), PXD000603 (f-h) and PXD001118 (i-k) dataset with the corresponding Swiss-Prot (a, d, f, i), Uniprot proteome (b, d, g, j, Swiss-Prot and proteome set being equal for the yeast dataset) and Uniprot proteome

with isoforms (c, e, h, k). The bars color-code represent the overlap: protein groups reported by all inferences are in teal (bottom), groups reported by 2, 3 and 4 groups in light blue with increasing darkness. Unique groups are orange, where light orange codes for groups whose accessions are reported in different combinations by other inferences and dark orange stands for groups with members, which are not reported by other inference algorithms. It can be seen, that with increasing complexity of the database, the reports' consensus decrease. Fido's results are also decreasing but the number of uniquely reported groups increases. This can be explained due to reported sub-proteins. PIA, ProteinLP and PP seem to be relatively robust against the change in database complexity

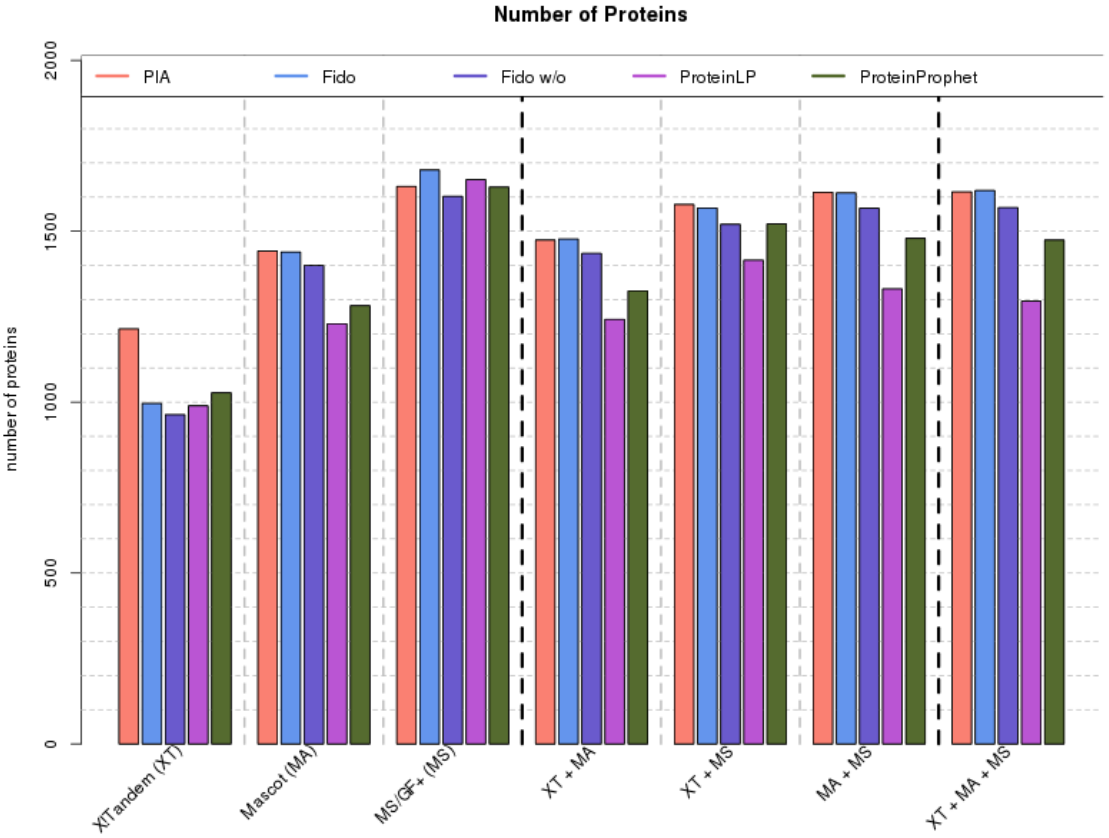

Figure 5: Number of protein groups reported using different inference algorithms and the Swiss-Prot database for the PXD000603 dataset. The bars show the number of FDR 1% valid protein groups reported for all analyzed inference algorithms and combinations of search engine identifications. For most combinations the same pattern for a ratio between the inference algorithms can be seen, as well as an increase in the number of reported protein groups when combining the search engines.

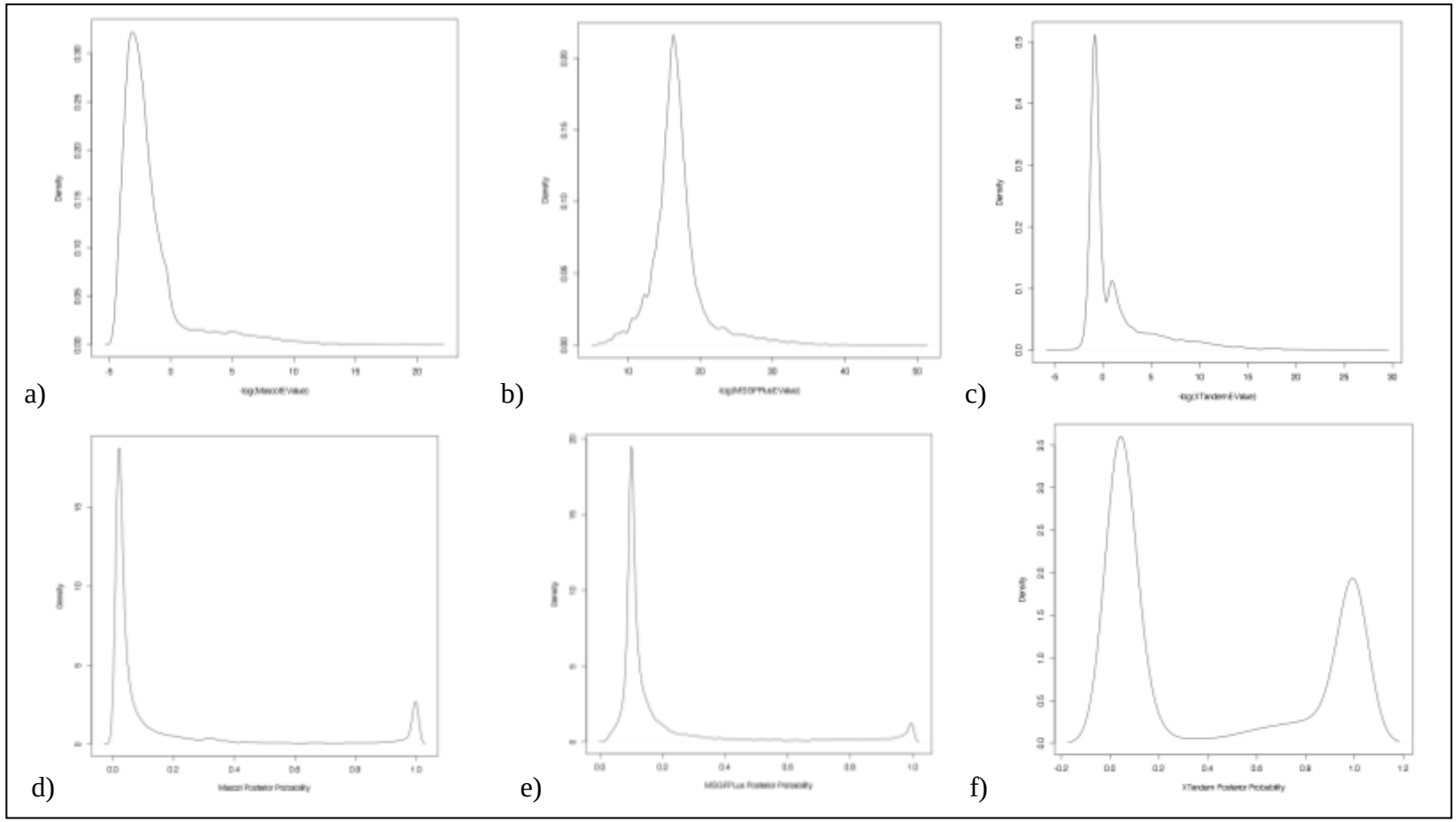

Figure 6: Distribution of scores for reported PSMs for the three search engines, the Swiss-Prot database for the PXD000603 dataset. **a)** Density of EValues of the Mascot search, **b)** Density of EValues of the MS-GF+ search, **c)** Density of EValues of the X!Tandem search, **d)** Posterior probabilities with OpenMS' IDPosteriorErrorProbability tool on the Mascot EValues, **e)** Posterior probabilities with OpenMS' IDPosteriorErrorProbability tool on the MS-GF+-SpectralEValues. **f)** Posterior probabilities with OpenMS' IDPosteriorErrorProbability tool on the X!Tandem-EValues.

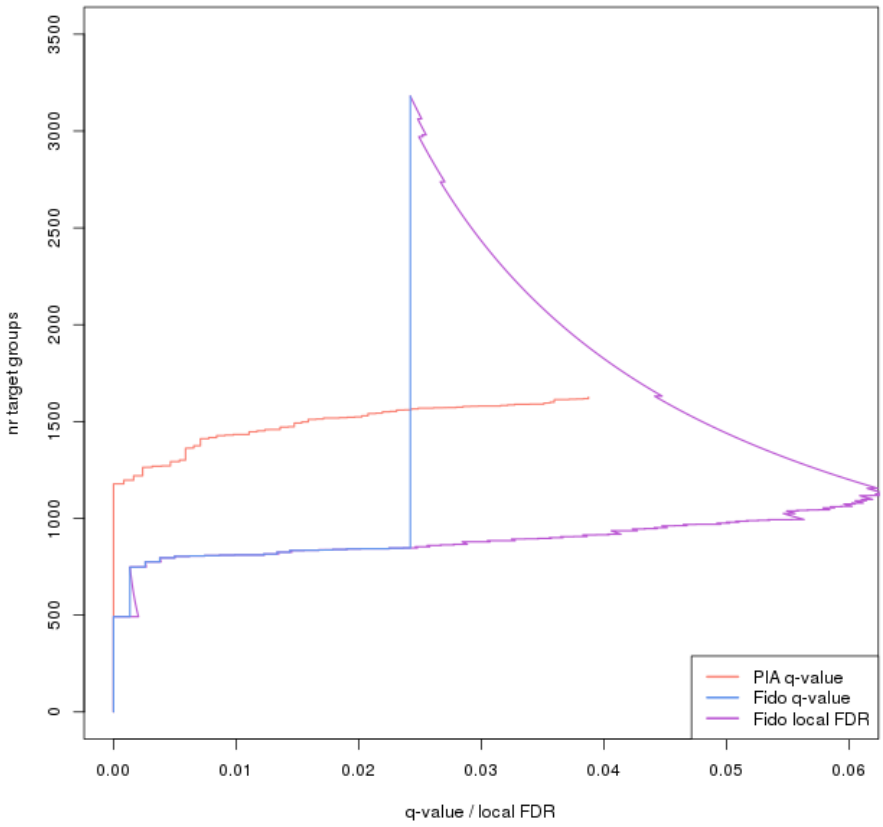

Figure 7: Pseudo-ROC plots of the protein groups reported for the PXD000603 datasets using the merged results of all search engines on the proteome database with isoforms and either the FDR q-values or the local FDR. This plot shows, plotted for the Fido results, that under certain circumstances the q-value can differ significantly from the local FDR. If this effect emerges under the given q-value threshold (usually 1%), the affected method generates more reports than expected. Larger differences between the local FDR and q-values can be seen at two ranges: one at q-values of 0.001 and one for q-values of 0.022. The respective plot for the PIA q-values is given for reference, here no larger discrepancies could be detected, and therefore the PIA local FDR values were not plotted.

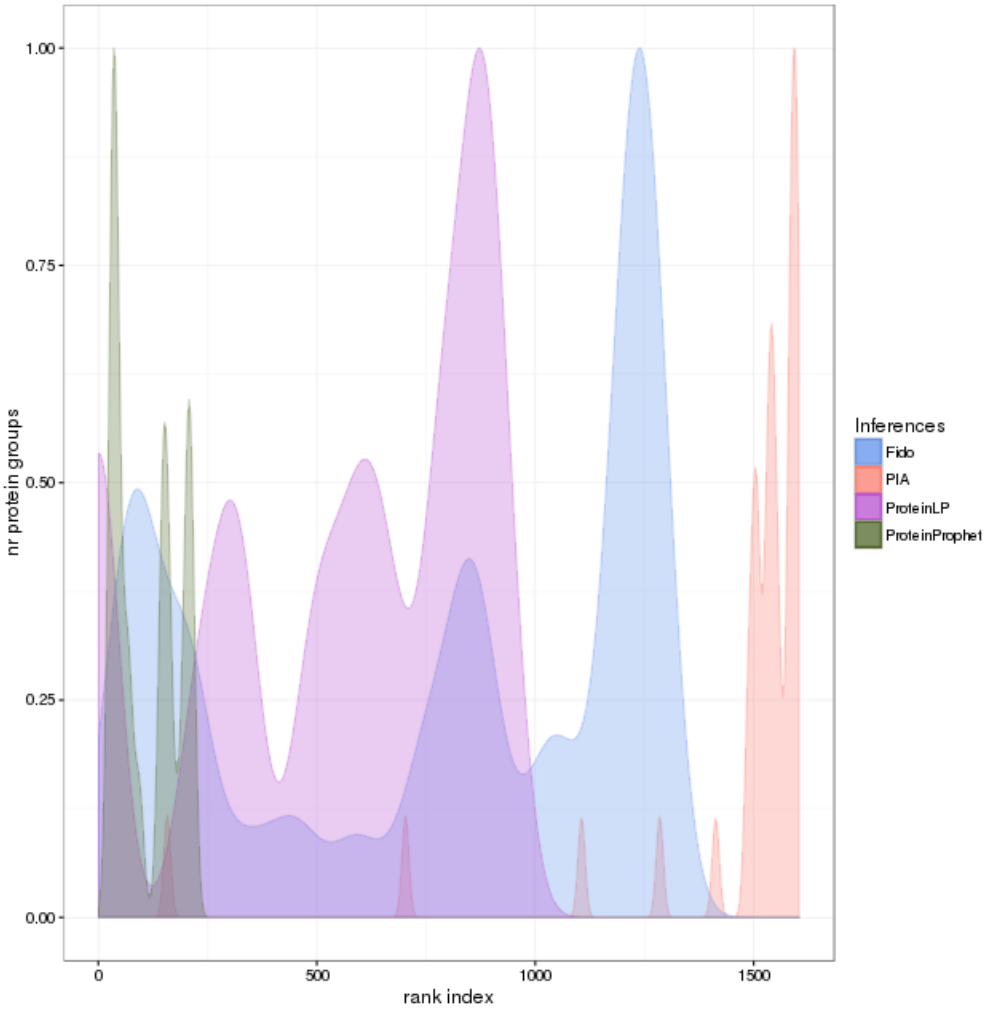

Figure 8: Distribution of the ranks of uniquely reported protein groups. This plot shows for the analyzed inference methods, on which ranks in the reported list of protein groups uniquely reported groups occur. Depicted is the data from the merge of PSMs from Mascot, X!Tandem and MS-GF+ for the PXD000603 dataset using the Swiss-Prot database. For PIA it can be seen, that almost all uniquely reported groups are at the end of the list. Fido and ProteinLP, on the other hand, distribute the unique groups over the complete range of indices, though with a tendency to the list's end. The most extreme case is ProteinProphet which reports its unique groups at the beginning. This reveals that an intuitive assumption, that the relatively high consensus of reported groups is found in the top of the report, is not correct for all algorithms.

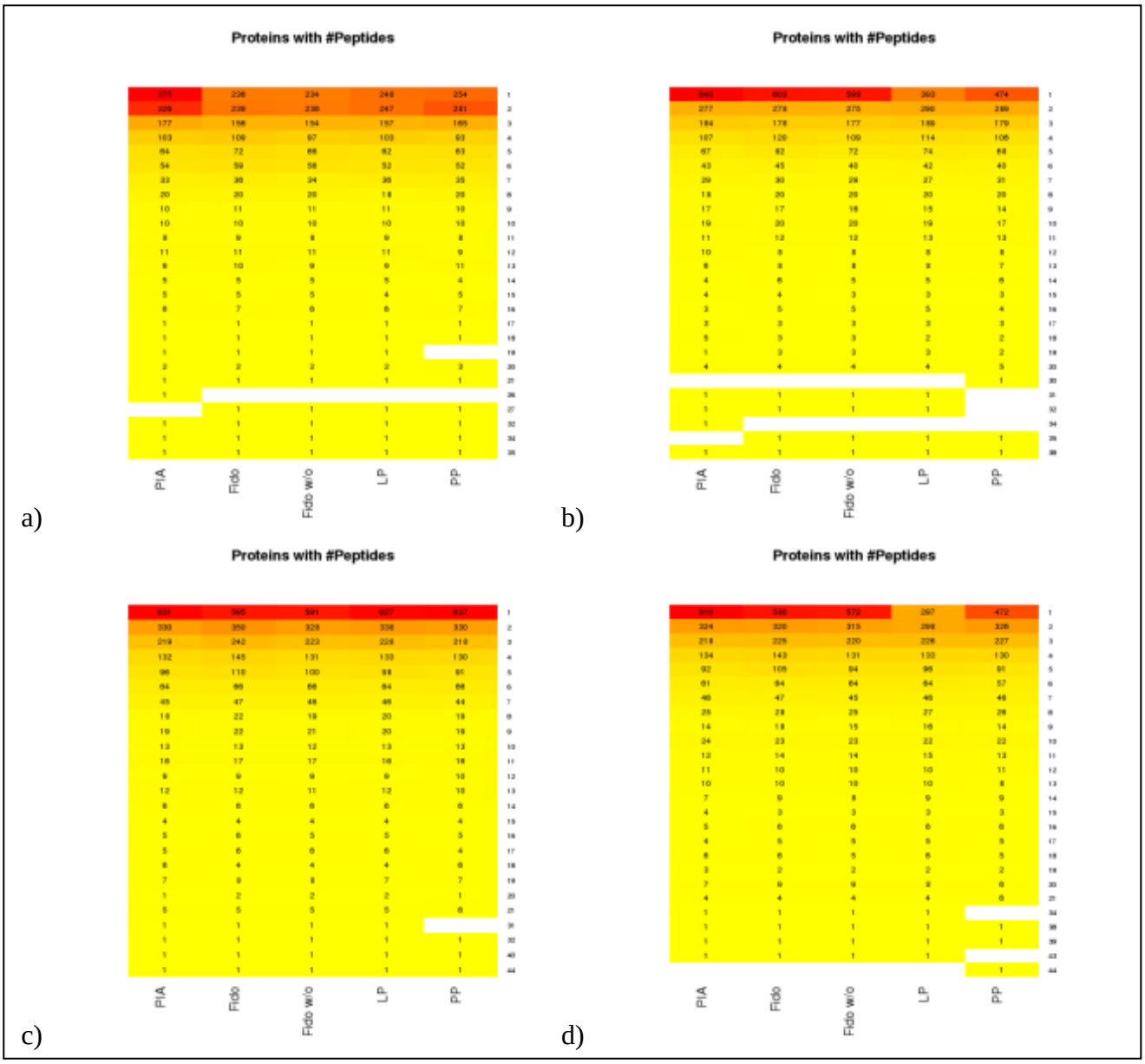

Figure 9: Numbers of identified peptides per protein. The graphics show the numbers of peptides identified per protein in a heatmap-like plot for identifications from a) X!Tandem, b) Mascot, c) MS-GF+ and d) combination of all for the PXD000603 dataset and the Swiss-Prot database. It can be seen, that the most proteins are identified with relatively few peptides, while only few proteins have ten or more peptides in this dataset. With the single search engines, PIA, Fido and ProteinProphet report on average 5.7% of proteins with ten or more peptides, while with the merge of the PSM results they report 6.5% with at least ten peptides, and also numerically at least eight proteins more with these many peptides. This shows that also on protein level a qualitative improvement is yielded by merging search results.