Focus on the spectra that matter by clustering of quantification data in shotgun proteomics

Matthew The[1] and Lukas Käll[1]

[1]Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH – Royal Institute of Technology, Box 1031, 17121 Solna, Sweden

November 22, 2018

Abstract

In shotgun proteomics, the amount of information that can be extracted from label-free quantification experiments is typically limited by the identification rate and the noise level of the quantitative signals. This generally causes a low sensitivity in differential expression analysis on protein level. Here, we propose a quantification-first approach that reverses the classical identification-first workflow. Specifically, we introduce a method, Quandenser, that applies unsupervised clustering on both MS1 and MS2 level to summarize all analytes of interest without assigning identities. This prevents valuable information from being discarded prematurely in the identification process and allows us to spend more effort on the identification process due to the data reduction achieved by clustering. Applying this methodology to a dataset of partially known composition, we could now employ open modification and de novo searches to identify multiple analytes of interest that would have gone unnoticed in traditional pipelines. Furthermore, Quandenser reports error rates on feature level which we integrated into our probabilistic protein quantification method, Triqler, to propagate error probabilities from feature level all the way to protein level. Quandenser/Triqler outperformed the state-of-the-art method MaxQuant/Perseus, consistently reporting more differentially abundant proteins, even in a clinical dataset where none were discovered previously. Compellingly, in all three clinical datasets investigated, the differentially abundant proteins showed enrichment for functional annotation terms.

Introduction

In mass spectrometry-based proteomics, label-free quantification (LFQ) is one of the most comprehensive methods to analyze protein concentrations in complex mixtures. Its main advantage is that it allows for comparisons in large sample cohorts and can, hence, handle complex experimental designs [2]. Currently, LFQ and quantitative proteomics in general are struggling to obtain sufficient coverage of the proteome [3] and also suffer from low sensitivity for differentially expressed proteins at false discovery rate thresholds [24]. While this can partially be attributed to inherent limitations in the methodology of mass spectrometry, it is, to a high degree, caused by the inadequacy of our current data analysis pipelines. We also note that LFQ is sometimes seen as cumbersome, as contrary to, for instance, isobaric labeling, one is not guaranteed a readout for an identified peptide in each sample. Frequently, this is resolved by missing value imputation but this introduces a multitude of issues [33, 18]. Novel methods for LFQ data analysis are necessary to address these problems in sensitivity and specificity.

Two well-recognized issues regarding the sensitivity of LFQ pipelines are that many MS1 features remain unassigned to peptides [34] or even to MS2 spectra [22] and that a large number of fragment spectra remain unidentified [26]. Matches-between-runs (MBR) has proven to be an effective technique to propagate MS2 information to the MS1 features [5, 34, 1] and clustering of MS2 spectra from large repositories has allowed us to zoom in on frequently unidentified spectra for peptide identification [11]. Unsupervised clustering of MS2 spectra also significantly reduces the number of MS2 spectra that need to be searched [8, 10, 29]. This allows more computationally expensive searches to be conducted such as partial digestions, variable modification searches, open modification searches or even *de novo* searches.

Moreover, as a first step, conventional data analysis pipelines match the individual sample's fragmentation spectra to peptide sequences with search engines [5, 34], a procedure that already limits the number of peptides and proteins that can be quantified. More than the limitation in numbers, such an *identification-first* strategy has some undesirable properties. For instance, if one would want to search the data again with an open modification search engine, all quantification would have to be redone. Yet, in reality, the underlying experimental quantitative data does not change due to a new search engine result, just our interpretation thereof.

By using unsupervised clustering of MS1 features, we can capture the underlying quantitative information from the features. We can assign MS2 spectra to such clusters but postpone their interpretation to a later stage. Such a *quantification-first* approach allows us to focus our efforts on improving the coverage by investigating frequently unidentified features without having to go through the costly process of repeated iterations of identification and quantification.

On the other hand, specificity is often compromised by the multitude of thresholds in the various steps of a quantification pipeline, which cause a lack of error control and do not properly account for error propagation [28]. Also, missing value imputation is known to induce false quantification accuracy [13]. We have previously introduced a Bayesian method, Triqler [28], for protein quantification that propagates error rates of the different steps in protein quantification and compensates for missing readouts in a Bayesian manner. Two features that Triqler did not include yet were the error rates of the association of MS1 features with MS2 spectra, and support for MBR.

Here, we introduce a new method, *Quandenser* (QUANtification by Distillation for ENhanced Signals with Error Regulation), that we subsequently interface to Triqler to substantially increase the sensitivity of the LFQ analysis pipeline. Quandenser condenses the quantification data by applying unsupervised clustering on both MS1 and MS2 level and thereby paves the way for a quantification-first approach. The algorithm combines the unknowns from both levels into a condensed set of MS1 features and MS2 spectra that are likely to be interesting for further investigation. Specifically, Quandenser incorporates MBR for MS1 features to increase sensitivity and uses MS2 spectrum clustering to decrease the number of spectra to be searched. Importantly, it also provides *feature-feature match* error rates which can be used as input to Triqler to account for the errors as a result of the MBR step.

Methods

Data sets

We downloaded RAW files for 3 datasets. The first dataset was a set of partially known composition with the UPS1 protein mixture spiked in at different concentrations in a yeast background (PRIDE project: PXD002370, 9 RAW files) [9]. We also downloaded RAW files for three clinical datasets. The first was a dataset studying bladder cancer [17] (PRIDE project: PXD002170, 8 RAW files), which will henceforth be referred to as the *Latosinska* dataset. The second was a dataset studying Hepatitis C Virus-associated hepatic fibrosis [4] (PRIDE project: PXD001474, 27 RAW files), which will be referred to as the *Bracht* dataset. The third dataset concerned a recent advancement in nanoscale proteomics applied to type 1 diabetes [35] (PRIDE project: PXD006847, 18 RAW files), which will be referred to as the *Zhu* dataset.

For the UPS-Yeast mixture, a UPS1 protein mixture was spiked into a 1 μg yeast background at respectively 25, 10 and 5 fmol concentration, with triplicates for each concentration. The Latosinska dataset featured 8 samples of tumor tissues of non-muscle invasive (stage pTa, $n = 4$) and muscle-invasive bladder cancer cases (stage pT2+, $n = 4$), without technical replicates. The Bracht dataset featured 27 samples of biopsies from patients with HCV-associated hepatic fibrosis, classified in a low fibrosis group ($n = 13$) and a high fibrosis group ($n = 14$), without technical replicates. The Zhu dataset featured 18 samples with $10 - -100$ cells each of human pancreatic islet section with nine samples from a type 1 diabetes donor and nine from a non-diabetic control donor, without technical replicates.

Prior to processing the runs with Quandenser, all RAW files were converted to mzML format with ProteoWizard [15], where we applied peak picking both on MS1- and MS2-level.

Quandenser

An overview of the Quandenser process is given in Figure 1. First, MS1 features were detected with Dinosaur v1.1.3 [27] and assigned to the MS2 spectra that were obtained inside the retention time and precursor isolation window. We will refer to a combination of an MS1 feature and an MS2 spectrum as a *mass-charge state*. Next, clustering of MS2 spectra was applied with MaRaCluster v0.05 [29]. One advantage of applying MS2 clustering first was that we could align retention times between two runs through pairs of spectra that end up in the same cluster. This alignment was done by fitting a spline function using iteratively reweighted least squares regression (IRLS). The IRLS algorithm provided protection against outliers that might have resulted from incorrect clusterings. We then estimated the standard deviation of the aligned retention times, which gave us a way to select a reasonable window to search for matching precursors (by default, 5 standard deviations), instead of having this window user-specified, as is usually needed.

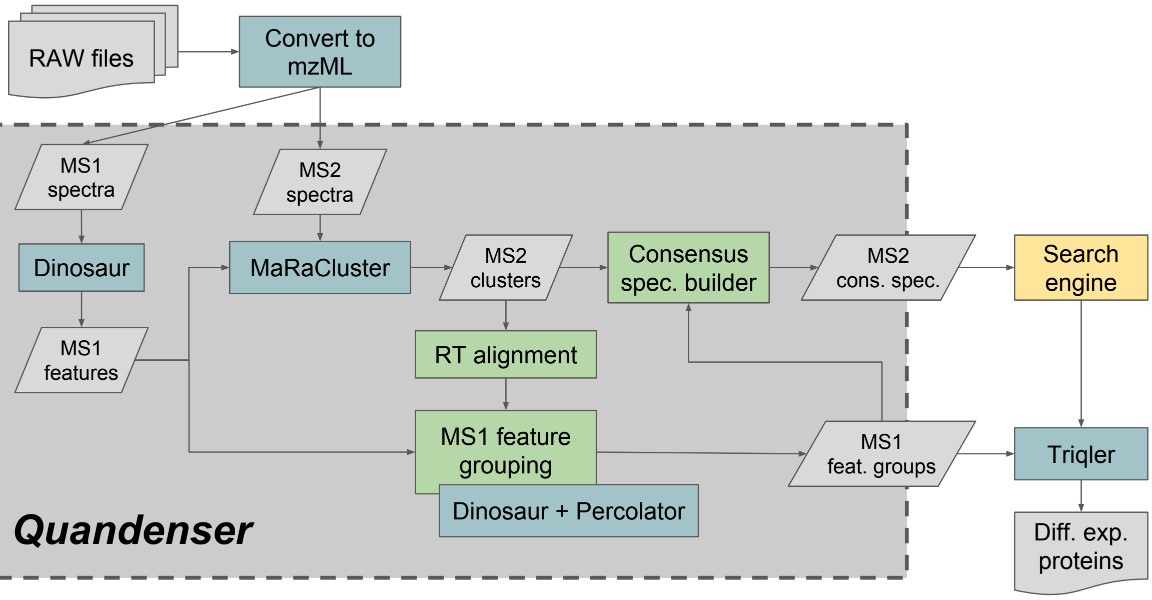

Figure 1: **Workflow within Quandenser and its employment in the protein quantification pipeline.** Blue boxes represent existing software packages, some of which are directly integrated in the Quandenser package, green boxes represent in-house packages, gray boxes represent intermediate results or files and the yellow box represents any search engine(s) of choice.

Retention times were aligned pair-wise between runs using a minimum spanning tree based on the similarity of chromatography runs [25]. For each pair-wise retention time alignment, the MS1 features discovered by Dinosaur from the two corresponding runs were matched based on a set of features, such as the difference between the observed and aligned retention time and precursor mass difference.

Decoy features were generated by shifting the precursor m/z by 5 × 1.000508 Th. We noted that the probability of observing a certain precursor mass is constrained by its composition of amino acids and therefore the probability of a match quickly tapers off when non-integer offsets are employed. Adding the random mass perturbation would, therefore, give an underestimation of the true FDR. Using these decoy features, we estimated error probabilities of these feature-feature matches through Percolator v3.02 [30]. We filtered out feature-feature matches with a posterior error probability (PEP) above 0.25, and at the same time retained the information of the error probabilities to be used as input to Triqler. For features for which no corresponding feature could be found in the opposite run, we searched for previously missed features using the targeted mode of Dinosaur, again scored by Percolator according to the same scheme as above. If a feature still had no match, a placeholder feature, representing a missing value, was added at the corresponding precursor mass and aligned retention time.

The MS1 features were grouped by feature-feature matches from the pair-wise alignments using single-linkage clustering, resulting in MS1 *feature groups*. The feature groups that had a missing value in more than M runs were filtered out. Subsequently, for each feature group, we selected all MS2 spectra that were previously assigned to a feature in the group and assigned the MS2 spectrum's corresponding cluster from MaRaCluster to the feature group. We then scored each match between a feature group and a spectrum cluster according to the following score:

$$s = \sum_{r=1}^{R} \mathbb{I}(c_r \geq 1) \cdot \log_2(I_r),$$

where $r = 1, \ldots, R$ indicates the run index, $\mathbb{I}$ is the indicator function that is 1 if the condition holds and 0 otherwise, c_r are the number of spectra in the spectrum cluster that can be linked to the feature for run r, and I_r is the intensity of the feature in run r. For each spectrum cluster, we only retained the feature groups with $s \geq \frac{\max\{s_g \mid g \in G\}}{2}$, with G the set of all feature groups associated with the spectrum cluster. The idea behind this filter is that the MS1 feature intensity is a good predictor of the dominant peptide species in the MS2 spectrum. This filter typically more than halved the number of mass-charge states that had to be searched, while maintaining or even increasing the number of peptide identifications due to the reduction of tested hypotheses.

Finally, Quandenser produces (1) a list of MS1 feature groups with their corresponding MS2 spectrum clusters, and (2) a spectrum file with, for each spectrum cluster, its consensus spectrum together with the mass-charge states of the assigned feature groups. The spectrum file can be any of the file formats that are supported by Proteowizard, such as mzML and mgf, and can then be processed by the user's search engine of choice.

Peptide and protein identification and quantification

For the UPS-yeast mixture, we created a concatenated FASTA file containing both the UPS1 proteins (`https://www.sigmaaldrich.com/`, accessed: 2018 Jan 17) and the Swiss-Prot database for yeast (`http://www.uniprot.org/`, accessed: 2016 Mar 15). The Latosinska, Bracht and Zhu sets were searched against the Swiss-Prot database for human (accessed: 2015 Nov 12). We created concatenated target-decoy databases as input to the search engines.

We used several search engines both as standalone packages as well as part of a cascade search approach [14]. We employed Tide [6], through the interface of the Crux 2.1 [21] package, and used Percolator v3.02 [30] to post-process the resulting PSMs. All parameters in Tide and Percolator were left to their default values, except for allowing up to 2 oxidations for all datasets. Furthermore, we used MODa v1.51 [23] and MSFragger (build 20170103.0) [16] for open modification searches. We extracted several relevant features of the respective search results with in-house python scripts and subsequently processed the PSMs with Percolator v3.02. Finally, we also used Novor v1.05.0573 [20] for de novo searches as a discovery tool and searched the resulting sequences with BLAST through the UniProtKB website. We did no statistical analysis on the Novor results.

After the search engine search, the output file from Quandenser and the search engine results can be combined into an input file to Triqler v0.1.4 with a python script available from the Quandenser repository (`bin/prepare_input.py`). This step includes the option for retention time-dependent normalization [34] which is the default option and was also applied to all datasets presented here. For the UPS-Yeast dataset, we allowed $M = 3$ missing values, for the Latosinska set we used $M = 4$, for the Bracht set $M = 7$ and for the Zhu set $M = 11$.

Some adaptations to Triqler were necessary to deal with the changes and additions to the pipeline. Firstly, the feature-feature match PEPs are used explicitly as an extra input to the feature node in Triqler's probabilistic graphical model (Supplementary Figure S1). Triqler then proceeds in normal fashion to calculate relative protein expression levels and finally outputs a list of differentially expressed proteins. Secondly, there is the issue of a many-to-many relation between feature groups and spectrum clusters. This requires a choice of feature group for each spectrum cluster, and, vice versa, a choice of spectrum clusters for each feature group. To resolve this, Triqler first selects the feature group with the best search engine score. Subsequently, if a feature group still has multiple peptides associated with it, Triqler chooses the peptide with the best combined PEP of the search engine PEP and the feature-feature match PEPs. Finally, to deal with open modification search results, we have to guard ourselves against assigning correct peptides with incorrect (small) modifications to feature groups. To prevent this, we only select the best peptide identification per protein for each spectrum cluster, under the assumption that the search engine will score the true peptide sequence, with or without modification, the highest.

As a comparison, we also analyzed the datasets with MaxQuant v1.6.1.0 [5], starting from the RAW files, followed by differential expression analysis with Perseus v1.6.1.3 [32]. We used the default parameters in MaxLFQ, except for allowing up to 2 oxidations and allowed the use of these modified peptides for quantification. For the differential expression analysis with Perseus, we filtered out decoy proteins and proteins with more than M missing values per dataset as stated above. We then $\log_2$ transformed the intensities, used missing value imputation with the default parameters and used Welch's t-test with different values of S_0 $(0.0, 0.3, 0.7, 1.0)$, where higher values of S_0 will increasingly prevent small fold changes from being selected as significant [31]. The results reported in the main text are for $S_0 = 0.3$, unless stated otherwise, as these generally gave the best trade-off between sensitivity and specificity.

Finally, we used DAVID 6.8 [12] to find significant functional annotation terms for the sets of differentially expressed proteins found by Quandenser/Triqler and MaxQuant/Perseus. We used the proteins identified at 5% protein-level identification FDR as the background set and thresholded the significant terms at a 5% Benjamin-Hochberg corrected FDR.

Results

We evaluated the performance of our quantification-first method, Quandenser, with different search engines followed by differential expression analysis with Triqler. We used one regular search engine, Tide, one open modification search engine, MODa and subsequently, we investigated a combination of them both, i.e. Tide and MODa in a cascade search setting. We compared these approaches to MaxQuant/Perseus with matches-between-runs (MBR) and also to applying Triqler directly to the MS2 search results, without clustering on MS1 and MS2 level with Quandenser but with feature detection using Dinosaur.

UPS-Yeast dataset

The UPS-Yeast dataset consisted of a total of 535k MS2 spectra. Assigning the MS1 features detected by Dinosaur to the MS2 spectra resulted in 934k mass-charge states, that were subsequently used as input to MaRaCluster. Allowing up to $M = 3$ missing values, resulted in 132k feature groups of which 49k were assigned to at least one spectrum cluster. Due to the possibility of multiple feature groups being assigned to a spectrum cluster, 109k mass-charge states (12% of the original number of mass-charge states), corresponding to 61k consensus spectra, remained to be searched by the search engine. Without the intensity scoring thresholds, 98k feature groups were assigned to at least on spectrum cluster and 238k mass-charge states would have to be searched.

Processing the Quandenser output with Triqler (`--fold_change_eval=0.8`) resulted in higher sensitivity compared to applying Triqler directly on the search results without clustering and at the same time controlled the FDR below the reported FDR of 5%, which MaxQuant/Perseus with MBR failed to do (Table 1). We also observed that MaxQuant/Perseus had trouble controlling the FDR, regardless of the value for the S_0 parameter (Supplementary Table S1).

UPS1 spike-in concentration [fmol]	25 vs 10		25 vs 5		10 vs 5	
Max. true positives	48		48		48	
Method	tp	fp	tp	fp	tp	fp
Quandenser/Triqler						
- Tide	43	1	45	1	35	0
- MODa	40	0	41	2	29	0
- Tide + MODa cascade	43	0	45	0	35	0
Tide + Triqler (without Quandenser/MBR)	40	0	40	0	34	0
MaxQuant/Perseus MBR ($S_0 = 0.3$)	39	9	42	15	36	1

Table 1: **Quandenser combined with Triqler achieves a high sensitivity on the UPS-Yeast dataset, while still maintaining control of the FDR.** The table lists the number of true and false positive significantly differentially expressed proteins at a 5% reported FDR threshold. Other values for S_0 for Perseus ($S_0 = 0.0, 0.7, 1.0$) resulted in inferior results.

To demonstrate the advantages of reducing the number of spectra and mass-charge states that need to be searched, we ran the unidentified consensus spectra through an open modification search with MODa using the cascade search approach [14]. We can see a clear increase in the number of feature groups that were assigned a peptide and more modest, but still significant, increases in the number of unique peptides and proteins (Supplementary Figure S2). However, the cascade search did not result in an increased sensitivity on the spiked-in proteins, as the newly discovered peptides were either modified versions of already identified peptides or came from already identified proteins (Table 1).

Interestingly, searching with MODa without searching with Tide first actually decreased the sensitivity on the spiked-in proteins relative to only searching with Tide, even though more unique peptides were identified than by Tide. This is likely a result of the lower sensitivity of open modification searches on unmodified peptides, as a result of the increased search space. We indeed discovered several unmodified peptides from UPS proteins that were confidently identified by Tide but not picked up by MODa. In this engineered dataset the modified peptides did seem to follow the correct expression pattern in the vast majority of the cases. In general, however, we should be careful about using modified peptides for quantification, as they are not guaranteed to follow the protein's expression pattern. On the other hand, quantifying modified peptides can be of great interest for understanding biological processes.

We also tried out MSFragger with its large precursor tolerance (±500 Da) as an alternative open modification search engine. It produced more identifications than MODa, reducing the number of unidentified consensus spectra to around 40%, but also produced several dubious modifications. MSFragger could, therefore, be a good source for finding candidate peptide identifications, but some extra verification seems to be required for the moment.

To illustrate the utility of MS2 clustering, we used Novor on the 20 most frequently occurring unidentified spectra, followed by a BLAST search [19]. Using this approach, we found 2 distinct peptides from a capsid of a known yeast virus (UniProtKB: P32503 / GAG_SCVLA) and another 2 distinct peptides from lysyl endopeptidase (UniProtKB: Q9HWK6 / LYSC_PSEAE), the latter of which might have been used for improved protein digestion, although this was not mentioned in the original manuscript. All but one of these largest 20 unidentified spectrum clusters were identified as peptides from the 2 above-mentioned proteins or as modifications of already identified peptides of high-abundant proteins (Supplementary Table S2).

Furthermore, the benefit of having clustered on MS1 level allowed us to zoom in on feature groups without peptide identifications, but with the same expression pattern as the UPS proteins. For this, we calculated the cosine distance between the expected expression pattern and the observed expression pattern, omitting missing values from the calculation, and selected the 200 feature groups with the smallest cosine distance (Figure 2, Supplementary Table S3). Of these 200 feature groups, 58 were identified through closer inspection as, often modified, UPS peptides and often came from chimeric spectra. One helpful approach in identifying these chimeric spectra was by filtering out the fragment peaks of an already identified peptide species and applying another open modification search [7].

Interestingly, we also found 68 feature groups which had consensus spectra that contained many fragments (≥ 25) between 100 and 200 Da that, based on their accurate masses, were carbohydrates or hydrocarbons and did not contain any nitrogen atoms (Supplementary Table S4). In total, we found 1 724 of these types of spectrum clusters, which mainly eluded towards the end of the runs. Furthermore, their expression pattern typically seemed to follow the UPS expression pattern (Figure 2) and MS2 spectra associated with these feature groups generally remained unidentified. Based on their precursor mass differences and late retention times, these feature groups most likely originated from polyethylene glycols (PEG), which might have been present as a contaminant in the UPS samples. The putative identifications of a yeast virus, lysyl endopeptidase, and PEGs are examples that demonstrate that Quandenser gives its users the capability to identify unknown either abundant or differentially abundant compounds in their samples.

Of the remaining feature groups, 50 were from analytes with low precursor mass (< 1000 Da), mostly charge 1 ions, which are generally hard to identify. For 8 feature groups, the UPS expression pattern was a result of deisotoping errors where isotopes of a UPS peptide were incorrectly counted towards the intensity of the feature. Finally, 13 feature groups remained unidentified and did not fit into any of the above categories, but usually had spectra that showed clear signs of chimericity or only had fragment ions spanning less than half of the peptide backbone making them hard to identify.

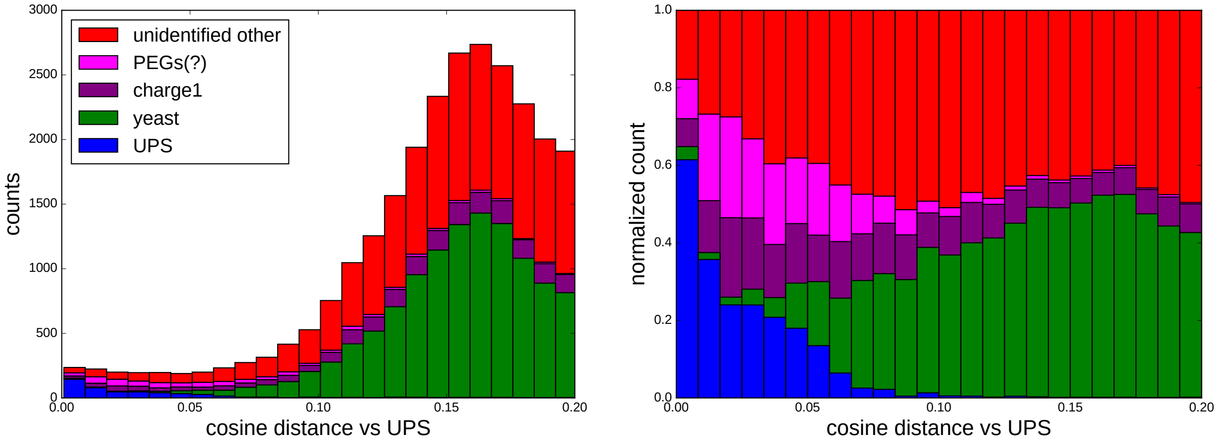

Figure 2: **The number of feature groups and their origin as a function of dissimilarity to the UPS1 spike-in concentrations.** The histogram displays the number (left pane) and relative number (right pane) of feature groups as a function of the cosine distance relative to the UPS1 spike-in concentrations. The vast majority of the identified feature groups had an expression pattern that conformed to their origin. Still, a large proportion of feature groups remained unidentified, including a group that was tentatively identified as polyethylene glycols (PEGs), which exhibited an expression pattern similar to the UPS1 proteins. The cosine distance between the yeast and UPS concentrations was 0.16 and we can indeed observe that the majority of the yeast peptides center around that value.

Clinical datasets

The Latosinska dataset consisted of 413k MS2 spectra, resulting in 991k mass-charge states after MS1 feature detection by Dinosaur. Allowing up to $M = 4$ missing values and filtering based on the intensity score, we were left with 83k feature groups, of which 47k had at least one spectrum cluster associated with them. This corresponded to 122k consensus spectra and 183k mass-charge states to be searched, just 18% of the original number of mass-charge states. This dataset contained a relatively large number of singleton clusters, i.e. clusters with only one spectrum, that could be identified. Such singleton clusters illustrate the benefit of quantification across runs as a criterion, opposed to requiring one fragment spectrum per run.

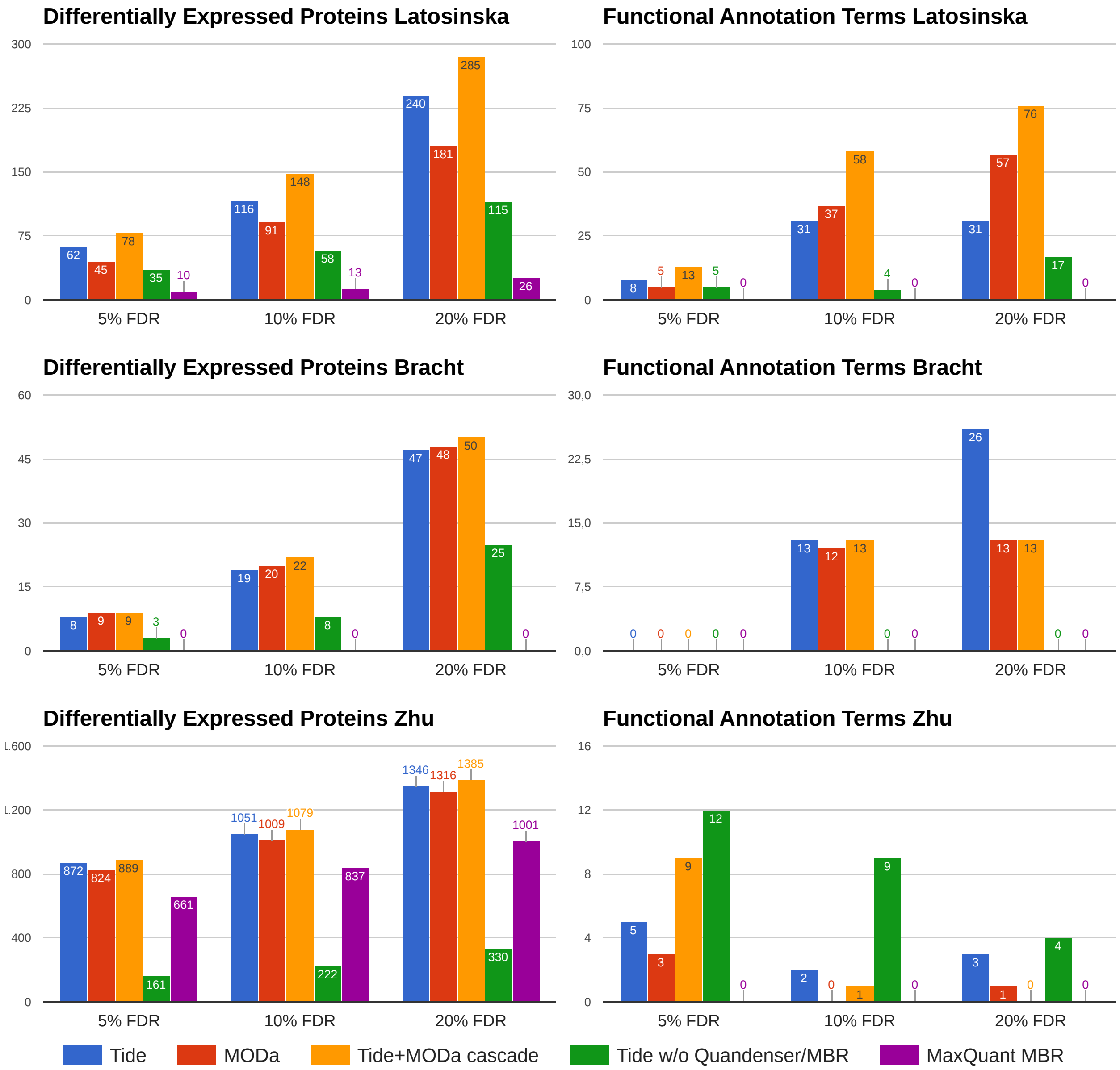

Figure 3: **With Quandenser/Triqler we generally discovered more differentially expressed proteins and enriched functional annotation terms than when not using Quandenser. Notably, we found enriched functional annotation terms for the Bracht set for which no enrichments were previously found.** The left plots show the number of differentially expressed proteins at 3 FDR thresholds. The plots on the right show the number of significant functional annotation terms we discovered with DAVID using the sets obtained in the left plots. Note that the FDR reported in the plots on the right refer to the differential expression FDR and not the functional annotation term FDR, which was kept fixed at 5%. The MaxQuant MBR series refer to the analyses done with MaxQuant with matches-between-runs followed by statistical analysis with Perseus.

Searching the consensus spectra with Tide and/or MODa resulted in more identifications compared to MaxQuant on all levels (Supplementary Figure S2). Subsequent processing with Triqler (`--fold_change_eval=0.8` resulted in more enriched functional annotation terms than applying Triqler directly on the MS2 search results (Figure 3, Supplementary Table S5). There was a noticeable advantage for a cascaded Tide+MODa search, both in terms of the number of differentially expressed proteins, as well as in the number of enriched functional annotation terms. The original study found only a single protein at a 5% differential expression FDR and 77 proteins with a p value below 0.05. These 77 proteins did not show any term enrichment in DAVID. The significant proteins called by MaxQuant/Perseus at the three FDR thresholds nor the 142 proteins with a p value below 0.05 showed any enriched terms either.

The Bracht dataset contained 1.01M MS2 spectra, which were assigned a total of 1.47M mass-charge states after feature detection. We allowed up to $M = 7$ missing values, which resulted in 69k feature groups in total and 45k with at least one spectrum cluster. This left 106k consensus spectra, with 150k mass-charge states to be searched, which was only 10% of the original number of mass-charge states.

Again, we observed an increase in the number of identifications on all levels compared to MaxQuant (Supplementary Figure S2). Analysis of the Tide search results with Triqler (`--fold_change_eval=0.5`; the fold change threshold used in the original study was $\log_2(1.5) = 0.58$) resulted in multiple differentially expressed proteins for all three searches, even at as low as 5% FDR (Figure 3). Conversely, neither the original study nor MaxQuant/Perseus found any differentially expressed proteins at 5% FDR. Moreover, functional annotation analysis with DAVID actually resulted in several significant terms for the 10% and 20% FDR thresholds (Supplementary Table S6).

Notably, in the original paper, 7 proteins with a p value below 0.05 and high fold change differences were subjected to verification through gene expression analysis, as well as targeted analysis with MRM. The 4 proteins that showed a consistent relationship with increasing fibrosis stages in both experiments (FBLN5, LUM, COL14A1 and MFAP4) were all discovered at 10% FDR, as was one protein that showed significance in the gene expression analysis but only partial statistical significance in the MRM analysis (TAGLN). The 2 proteins with the least consistent relation of expression levels with increasing fibrosis stages (CSRP2 and CNN2) were not discovered at 10% FDR, though CSRP2 was called significant at 20% FDR. In the original study, these 2 proteins actually obtained a lower p value than the other 5 proteins, and Quandenser/Triqler, thus, seemed to give a better ordering of these 7 proteins.

Finally, we wanted to test the applicability of the method to the Zhu dataset, that contained samples of a small number (10–100) of cells. This dataset comprised 593k MS2 spectra, assigned to 1.02M mass-charge states. We allowed up to $M = 11$ missing values, resulting in 73k feature groups of which 52 have at least one spectrum cluster. This left 117k consensus spectra, corresponding to 187k mass-charge states to be searched, 18% of the original number. We set the $\log_2$ fold change threshold `--fold_change_eval=1.0` in Triqler, corresponding to the threshold employed in the original study.

The number of identified feature groups and unique peptides were much closer between Quandenser and MaxQuant for this particular set (Supplementary Figure S2). Nevertheless, Quandenser/Triqler managed to discover more significant proteins across all tested FDR thresholds compared to MaxQuant/Perseus (Figure 3). Using Tide as a search engine, we found 703 significant proteins at 2% FDR, considerably more than the 304 proteins at the same FDR found in the MaxQuant/Perseus analysis of the original study. Again, we found enriched functional annotation terms associated to several sets of significant proteins using Quandenser/Triqler (Supplementary Table S7). No enriched terms could be found for MaxQuant/Perseus analysis, neither from the original study nor from our own reanalysis. Interestingly, using Triqler without Quandenser appeared more sensitive in the functional annotation enrichment analysis, as did using stricter FDR thresholds.

Discussion

Here, we demonstrated the utility of a quantification-first approach in which we cluster both MS1 features and MS2 spectra prior to identifying spectra, peptides and proteins of interest. Not only does this approach provide several new ways of obtaining unidentified analytes of interest, but by combining Quandenser with Triqler, we discovered substantially more differentially expressed proteins and enriched functional annotation terms than MaxQuant/Perseus.

The clustering steps further unveil the potential of integrated quantification and identification error models such as Triqler. Using matches-between-runs with feature-feature match error rates substantially increases the proteome coverage while still controlling the differential expression FDR. Quandenser/Triqler showed remarkable increases in the number of differentially expressed proteins which compellingly enriched for functional annotation terms. Moreover, the clustering of MS2 spectra prior to identification allows a wide variety of, potentially computationally expensive, search strategies to be applied to the quantification data without the need for repeating the quantification steps. Especially with regards to the constant stream of new search engines, this provides an easy way for users to efficiently interrogate their data.

We showed that clustering of MS2 spectra was effective in increasing the coverage of the proteome, both by means of reducing of the number of spectra that need to be searched by one order of magnitude, as well as by focusing on the largest clusters that remained unidentified. By using MODa as an open modification search in a cascaded search after a regular database search with Tide, the number of identified consensus spectra was increased by 47% for the UPS-Yeast dataset, 14% for the Latosinska dataset, 20% for the Bracht dataset and 17% for the Zhu dataset. Not surprisingly, across all datasets, the identification rate for large clusters ($> 80\%$ for cluster size ≥ 16) was much higher than for small clusters ($< 50\%$ for cluster size ≤ 7). Furthermore, de novo searches on large MS2 spectrum clusters of the UPS-Yeast dataset resulted in the identification of peptides and proteins not present in the database.

One of the benefits of combining clustering on MS1 level with clustering on MS2 level is that we can include quantification information in the selection of MS2 spectrum clusters of interest. Specifically, this addresses a frequently observed phenomenon in MS2 spectrum clustering in which the majority of the clusters only contain a single spectrum, known as singleton clusters [8, 10, 29]. On the one hand, these could be the result of poor quality MS2 spectra or spurious MS2 fragmentation events, which are often irrelevant and should preferably not be matched by a search engine. On the other hand, they could be low abundant peptides rarely selected for fragmentation. Doing the quantification prior to identification allowed us to get rid of the vast majority of uninteresting singleton clusters, due to them not having an MS1 feature in most of the runs. Instead, the Quandenser workflow retained the fragment spectra from the analytes that were quantified across several runs, that is, it extracted the data points the experimenter would care about most.

Alternatively, we could use similarity of feature groups' expression patterns as an indication that they originated from the same group of proteins. For instance, in the UPS-Yeast dataset, we identified a major part of the unidentified analytes that had a similar expression to the spiked-in UPS proteins as modified UPS proteins. This was made possible by the fact that the number of hypotheses was drastically reduced to just the set of 48 UPS proteins. Unfortunately, this technique cannot directly be used to increase the number of peptides available for protein quantification, as we would introduce a bias by only identifying peptides that already have a similar quantification pattern to the proteins that we searched for. It is, however, still a way to obtain more knowledge about peptides and modifications we might be missing out on. Simultaneously, this approach revealed an interesting group of analytes that covaried with the UPS proteins and most likely originated from polyethylene glycols. Although further investigation is required to confirm this, these analytes would not have come to our attention if we had used the traditional identification-first approach.

Several improvements could still be made to increase the sensitivity of Quandenser and the quantification-first pipeline in general. For example, in a small number of cases ($< 5\%$) the intensity score filter removed the correct feature group from a spectrum cluster due to a high-intensity analyte in the neighborhood. In most of the cases, the peptide was still identified through other spectrum clusters and no sensitivity was lost. However, we could use retention time differences between the MS2 spectra in the spectrum cluster and the MS1 features in the feature group to mitigate this. Furthermore, almost half of the spectrum clusters were assigned to at least two feature groups, even after the intensity score filter. This suggests that a large proportion of these spectra contain fragments ions from multiple peptide species. Indeed, by filtering out the fragment peaks of an already identified peptide species, we identified a second peptide species in a number of cases. Applying this technique to entire datasets will be a subject for further investigation.

The quantification-first approach in label-free protein quantification, thus, provides an attractive alternative to the traditional identification-first approach. Through the use of unsupervised clustering, we condensed the data into a comprehensive format that retained the relevant information and thereby allow the researcher to spend more time on a reduced set of hypotheses. By subsequently propagating the feature-level error rates to probabilistic protein quantification methods, the bounds of sensitivity and specificity in LFQ are extended considerably. Already, we can see the benefits of this approach in terms of coverage and sensitivity using the techniques presented here, but many more modes of interpretation are available, ready to be applied.

References

[1] Andrea Argentini, Ludger JE Goeminne, Kenneth Verheggen, Niels Hulstaert, An Staes, Lieven Clement, and Lennart Martens. moFF: A robust and automated approach to extract peptide ion intensities. *Nature Methods*, 13(12):964–966, 2016.

[2] Marcus Bantscheff, Simone Lemeer, Mikhail M Savitski, and Bernhard Kuster. Quantitative mass spectrometry in proteomics: Critical review update from 2007 to the present. *Analytical and Bioanalytical Chemistry*, 404(4):939–965, 2012.

[3] Marcus Bantscheff, Markus Schirle, Gavain Sweetman, Jens Rick, and Bernhard Kuster. Quantitative mass spectrometry in proteomics: A critical review. *Analytical and Bioanalytical Chemistry*, 389(4):1017–1031, 2007.

[4] Thilo Bracht, Vincent Schweinsberg, Martin Trippler, Michael Kohl, Maike Ahrens, Juliet Padden, Wael Naboulsi, Katalin Barkovits, Dominik A Megger, Martin Eisenacher, et al. Analysis of disease-associated protein expression using quantitative proteomics—fibulin-5 is expressed in association with hepatic fibrosis. *Journal of Proteome Research*, 14(5):2278–2286, 2015.

[5] Jürgen Cox, Marco Y Hein, Christian A Luber, Igor Paron, Nagarjuna Nagaraj, and Matthias Mann. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & Cellular Proteomics*, 13(9):2513–2526, 2014.

[6] Benjamin J Diament and William Stafford Noble. Faster SEQUEST searching for peptide identification from tandem mass spectra. *Journal of Proteome Research*, 10(9):3871–3879, 2011.

[7] Viktoria Dorfer, Sergey Maltsev, Stephan Winkler, and Karl Mechtler. CharmeRT: Boosting peptide identifications by chimeric spectra identification and retention time prediction. *Journal of Proteome Research*, 2018.

[8] Ari M Frank, Matthew E Monroe, Anuj R Shah, Jeremy J Carver, Nuno Bandeira, Ronald J Moore, Gordon A Anderson, Richard D Smith, and Pavel A Pevzner. Spectral archives: Extending spectral libraries to analyze both identified and unidentified spectra. *Nature Methods*, 8(7):587–591, 2011.

[9] Quentin Giai Gianetto, Florence Combes, Claire Ramus, Christophe Bruley, Yohann Couté, and Thomas Burger. Calibration plot for proteomics: A graphical tool to visually check the assumptions underlying FDR control in quantitative experiments. *Proteomics*, 16(1):29–32, 2016.

[10] Johannes Griss, Joseph M Foster, Henning Hermjakob, and Juan Antonio Vizcaíno. PRIDE Cluster: Building a consensus of proteomics data. *Nature Methods*, 10(2):95–96, 2013.

[11] Johannes Griss, Yasset Perez-Riverol, Steve Lewis, David L Tabb, José A Dianes, Noemi del Toro, Marc Rurik, Mathias Walzer, Oliver Kohlbacher, Henning Hermjakob, et al. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nature Methods*, 13(8):651, 2016.

[12] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44, 2008.

[13] Yuliya Karpievitch, Jeff Stanley, Thomas Taverner, Jianhua Huang, Joshua N Adkins, Charles Ansong, Fred Heffron, Thomas O Metz, Wei-Jun Qian, Hyunjin Yoon, et al. A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics*, 25(16):2028–2034, 2009.

[14] Attila Kertesz-Farkas, Uri Keich, and William Stafford Noble. Tandem mass spectrum identification via cascaded search. *Journal of Proteome Research*, 14(8):3027–3038, 2015.

[15] Darren Kessner, Matt Chambers, Robert Burke, David Agus, and Parag Mallick. ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics*, 24(21):2534–2536, 2008.

[16] Andy T Kong, Felipe V Leprevost, Dmitry M Avtonomov, Dattatreya Mellacheruvu, and Alexey I Nesvizhskii. MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nature Methods*, 14(5):513, 2017.

[17] Agnieszka Latosinska, Konstantinos Vougas, Manousos Makridakis, Julie Klein, William Mullen, Mahmoud Abbas, Konstantinos Stravodimos, Ioannis Katafigiotis, Axel S Merseburger, Jerome Zoidakis, et al. Comparative analysis of label-free and 8-Plex iTRAQ approach for quantitative tissue proteomic analysis. *PloS one*, 10(9):e0137048, 2015.

[18] Cosmin Lazar, Laurent Gatto, Myriam Ferro, Christophe Bruley, and Thomas Burger. Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *Journal of Proteome Research*, 15(4):1116–1125, 2016.

[19] Susanna L Lundström, Bo Zhang, Dorothea Rutishauser, Dag Aarsland, and Roman A Zubarev. SpotLight Proteomics: Uncovering the hidden blood proteome improves diagnostic power of proteomics. *Scientific Reports*, 7:41929, 2017.

[20] Bin Ma. Novor: Real-time peptide de novo sequencing software. *Journal of the American Society for Mass Spectrometry*, 26(11):1885–1894, 2015.

[21] Sean McIlwain, Kaipo Tamura, Attila Kertesz-Farkas, Charles E Grant, Benjamin Diament, Barbara Frewen, J Jeffry Howbert, Michael R Hoopmann, Lukas Käll, Jimmy K Eng, et al. Crux: Rapid open source protein tandem mass spectrometry analysis. *Journal of Proteome Research*, 13(10):4488–4491, 2014.

[22] Annette Michalski, Juergen Cox, and Matthias Mann. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *Journal of Proteome Research*, 10(4):1785–1793, 2011.

[23] Seungjin Na, Nuno Bandeira, and Eunok Paek. Fast multi-blind modification search through tandem mass spectrometry. *Molecular & Cellular Proteomics*, pages mcp–M111, 2011.

[24] Dana Pascovici, David CL Handler, Jemma X Wu, and Paul A Haynes. Multiple testing corrections in quantitative proteomics: A useful but blunt tool. *Proteomics*, 16(18):2448–2453, 2016.

[25] Hannes L Röst, Yansheng Liu, Giuseppe D'Agostino, Matteo Zanella, Pedro Navarro, George Rosenberger, Ben C Collins, Ludovic Gillet, Giuseppe Testa, Lars Malmström, et al. TRIC: An automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nature Methods*, 13(9):777, 2016.

[26] Owen S Skinner and Neil L Kelleher. Illuminating the dark matter of shotgun proteomics. *Nature Biotechnology*, 33(7):717, 2015.

[27] Johan Teleman, Aakash Chawade, Marianne Sandin, Fredrik Levander, and Johan Malmström. Dinosaur: A refined open-source peptide MS feature detector. *Journal of Proteome Research*, 15(7):2143–2151, 2016.

[28] Matthew The and Lukas Käll. Integrated identification and quantification error probabilities for shotgun proteomics. *bioRxiv*, 2018.

[29] Matthew The and Lukas Kall. MaRaCluster: A fragment rarity metric for clustering fragment spectra in shotgun proteomics. *Journal of Proteome Research*, 15(3):713–720, 2016.

[30] Matthew The, Michael J MacCoss, William S Noble, and Lukas Käll. Fast and accurate protein false discovery rates on large-scale proteomics data sets with Percolator 3.0. *Journal of the American Society for Mass Spectrometry*, 27(11):1719, 2016.

[31] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[32] Stefka Tyanova, Tikira Temu, Pavel Sinitcyn, Arthur Carlson, Marco Y Hein, Tamar Geiger, Matthias Mann, and Jürgen Cox. The Perseus computational platform for comprehensive analysis of (prote) omics data. *Nature Methods*, 13(9):731, 2016.

[33] Bobbie-Jo M Webb-Robertson, Holli K Wiberg, Melissa M Matzke, Joseph N Brown, Jing Wang, Jason E McDermott, Richard D Smith, Karin D Rodland, Thomas O Metz, Joel G Pounds, and Katrina M Waters. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *Journal of Proteome Research*, 14(5):1993–2001, 2015.

[34] Bo Zhang, Lukas Käll, and Roman A Zubarev. DeMix-Q: Quantification-centered data processing workflow. *Molecular & Cellular Proteomics*, 15(4):1467–1478, 2016.

[35] Ying Zhu, Paul D Piehowski, Rui Zhao, Jing Chen, Yufeng Shen, Ronald J Moore, Anil K Shukla, Vladislav A Petyuk, Martha Campbell-Thompson, Clayton E Mathews, et al. Nanodroplet processing platform for deep and quantitative proteome profiling of 10–100 mammalian cells. *Nature Communications*, 9(1):882, 2018.