Benchmarking quantitative label-free LC–MS data processing workflows using a complex spiked proteomic standard dataset

Claire Ramus, Agnès Hovasse, Marlène Marcellin, Anne-Marie Hesse, Emmanuelle Mouton-Barbosa, David Bouyssié, Sebastian Vaca, Christine Carapito, Karima Chaoui, Christophe Bruley, et al.

▶ **To cite this version:**

Claire Ramus, Agnès Hovasse, Marlène Marcellin, Anne-Marie Hesse, Emmanuelle Mouton-Barbosa, et al.. Benchmarking quantitative label-free LC–MS data processing workflows using a complex spiked proteomic standard dataset. Journal of Proteomics, 2016, 132, pp.51-62. 10.1016/j.jprot.2015.11.011 . hal-02083890

HAL Id: hal-02083890
https://hal.science/hal-02083890

Submitted on 19 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ACCEPTED MANUSCRIPT

Benchmarking quantitative label-free LC-MS data processing workflows using a complex spiked proteomic standard dataset.

Claire Ramus[1,4,5,6 *], Agnès Hovasse[1,7 *], Marlène Marcellin[1,2,3 *], Anne-Marie Hesse[1,4,5,6 *], Emmanuelle Mouton-Barbosa[1,2,3], David Bouyssié[1,2,3], Sebastian Vaca[1,7], Christine Carapito[1,7], Karima Chaoui[1,2,3], Christophe Bruley[1,4,5,6], Jérôme Garin[1,4,5,6], Sarah Cianférani[1,7], Myriam Ferro[1,4,5,6], Alain Van Dorssaeler[1,7], Odile Burlet-Schiltz[1,2,3], Christine Schaeffer[1,7], Yohann Couté[1,4,5,6], Anne Gonzalez de Peredo[1,2,3 §]

[1]ProFi, Proteomic French Infrastructure

[2]CNRS UMR5089 Institut de Pharmacologie et de Biologie Structurale, 118 route de Narbonne 31077 Toulouse, France

[3]Université de Toulouse, 205, route de Narbonne, 31077 Toulouse, France

[4] CEA, DSV, iRTSV, Laboratoire de Biologie à Grande Echelle, Grenoble, F-38054, France

[5] INSERM U1038, Grenoble, F-38054, France

[6] Université Grenoble, F-38054, France

[7] Laboratoire de Spectrométrie de Masse BioOrganique (LSMBO), IPHC, Université de Strasbourg, CNRS, UMR7178, 25 Rue Becquerel 67087 Strasbourg, France

* These authors contributed equally to this work

[§]Corresponding author contact information:

gonzalez@ipbs.fr, tel: +33 (0)5 61 17 55 41; fax: (33) (0) 5 61 17 59 00

Email addresses:

ACCEPTED MANUSCRIPT

CR: claire.ramus@cea.fr

AH: ahovasse@unistra.fr

MM:marlene.marcellin@ipbs.fr

AMH: anne-marie.hesse@cea.fr

EMB:emmanuelle.mouton@ipbs.fr

DB: bouyssie@ipbs.fr

SV: sebastian.vaca@etu.unistra.fr

CC: ccarapito@unistra.fr

KC: karima.chaoui@ipbs.fr

CB: christophe.bruley@cea.fr

JG: jerome.garin@cea.fr

SC: sarah.cianferani@unistra.fr

MF: myriam.ferro@cea.fr

AVD: vandors@unistra.fr

OBS: schiltz@ipbs.fr

CS: christine.schaeffer@unistra.fr

YC: yohann.coute@cea.fr

Keywords: proteomic standard, nanoLC-MS/MS, label-free quantification, computational proteomics, spectral counting, MS signal analysis

Abstract

Proteomic workflows based on nanoLC-MS/MS data-dependent-acquisition analysis have progressed tremendously in recent years. High-resolution and fast sequencing instruments have enabled the use of label-free quantitative methods, based either on spectral counting or on MS signal analysis, which appear as an attractive way to analyze differential protein

ACCEPTED MANUSCRIPT

expression in complex biological samples. However, the computational processing of the data for label-free quantification still remains a challenge. Here, we used a proteomic standard composed of an equimolar mixture of 48 human proteins (Sigma UPS1) spiked at different concentrations into a background of yeast cell lysate to benchmark several label-free quantitative workflows, involving different software packages developed in recent years. This experimental design allowed to finely assess their performances in terms of sensitivity and false discovery rate, by measuring the number of true and false-positive (respectively UPS1 or yeast background proteins found as differential). The spiked standard dataset has been deposited to the ProteomeXchange repository with the identifier PXD001819 * and can be used to benchmark other label-free workflows, adjust software parameter settings, improve algorithms for extraction of the quantitative metrics from raw MS data, or evaluate downstream statistical methods.

* dataset accessible during peer review process at:
http://www.ebi.ac.uk/pride/archive/users/profile
Username: reviewer40987@ebi.ac.uk
Password: mWsW9Tcw

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Introduction

Label-free quantitative methods based on LC-MS/MS have become increasingly popular in proteomic studies, as an attractive and powerful way to analyze differential protein expression in complex biological samples [1-3]. They can be based either on the measurement of the MS/MS sampling rate for a particular protein (spectral counting), or on the MS chromatographic peak area of its corresponding peptides in the survey MS scan (MS trace analysis), both values being directly related to protein abundance. Both approaches have benefited from tremendous improvements in instrumentation, namely increased sequencing speed for spectral counting approaches (up to 15-20Hz in recent orbitrap or Q-TOF mass spectrometers) and higher resolution allowing more accurate MS signal analysis and improved matching of complex LC-MS maps. These methods have concomitantly gained in analytical depth, and can now routinely be used to profile the expression of thousands of proteins from biological systems submitted to different conditions. An important point is however to be able to assess, minimize, and eventually correct the variability associated to the LC-MS/MS analytical workflow, to ensure sufficient repeatability of the measurements and provide robust relative quantification of proteins across samples. To this respect, the development of proteomic standards has proved to be essential to assess the performances of LC-MS platforms, provide a quality control of the system and identify potential sources of variability. Importantly, they are also needed to evaluate the downstream elements of the analytical pipeline, i.e. bioinformatics processing and statistical analysis, which represent critical steps to generate the final comparative results.

The yeast Saccharomyces cerevisiae proteome has been used in many studies as a test sample to illustrate the benefits of various technological optimizations in the LC-MS/MS workflow. Due to its wide availability and relatively high complexity and dynamic range, it can be considered as a good surrogate to many real biological samples, both for method development

ACCEPTED MANUSCRIPT

and quality control. In previous studies, yeast samples have been used to establish and demonstrate the efficiency of a wide range of metrics to evaluate the LC-MS/MS performances[4, 5]. These metrics were directly related to the LC system, the MS instrument (electrospray source, MS1 and MS2 intensities), the dynamic sampling, but also the first steps of data processing, i.e. peptide identification results. They were applied by Paulovich et al. for LC-MS benchmarking of several instrumental systems operated in different laboratories [6]. Instead of focusing on specific proteins or peptides, the monitoring proposed in this study allowed to give a global and very exhaustive view of the quality of the analysis through general metrics reflecting for example the median peak FWHM on the whole peptide population, the number of MS1 or MS2 scans triggered over various portions of the chromatogram, the level of TIC, the median MS1 signal for the population of identified precursors, or the number of peptides and proteins identified.

However, the final objective of most label-free studies is to measure quantitative levels, and detect variation of some proteins across samples. To evaluate the performances of a workflow in this respect, it is relevant to use a standard spiked with known amounts of some peptides or proteins, which can then be specifically monitored to assess the ability of the analysis in detecting relative quantitative changes. Controlled datasets based on spike-in experiments thus represent a useful tool to objectively assess the performances of quantitative methods for differential analysis. Paired comparison between spiked versus non-spiked samples can be performed to benchmark analytical and computational pipelines for biomarker discovery. Such controlled datasets with known “ground truth” have been for example generated in the past in the field of microarray analysis, by spiking at different concentrations a panel of 100-200 specific RNAs into a well-defined, constant background of RNA species [7], and was then widely used as a gold standard to evaluate various data processing methods [8-12]. In the proteomics field, spiked samples are also often used to evaluate MS methods or data

ACCEPTED MANUSCRIPT

processing tools, although generally the number of spiked proteins or peptides is relatively low [13-16]. Interestingly, as exemplified in the report from Paulovich et al, the use of a more complex spiked material, such as the UPS1 standard containing 48 well-characterized purified proteins, allows to compute more extensively the exact proportion of false discoveries (number of yeast false positives relative to the total number of proteins declared as variant) and of true discoveries (number of true positives out of the 48 real variant UPS1 proteins). As a proof of concept of the kind of benchmarking that can be done with this spiked standard, they showed the performances of a spectral count approach (the SASPECT method) for detection of biomarkers when comparing the spiked sample (simulating a case sample) and the pure yeast reference sample (control sample).

In the present study, we wanted to extend this concept and use the yeast+UPS1 standard to benchmark several tools developed in recent years for relative quantification, including widely used software such as MaxQuant and Skyline. Indeed, while numerous software tools have been developed and are more and more routinely used for label-free quantitation, stringent and side-by-side evaluations have to be performed to prove the efficiency of the quantification. In addition, proper tuning and parameter settings in each of these software tools are also important for optimal downstream analysis. We thus generated a dataset from yeast samples spiked with 9 different concentrations of UPS1, analyzed in triplicate on an Orbitrap-Velos mass spectrometer. Starting from this dataset, different data processing workflows were implemented to perform relative quantification of proteins. Common statistical tests and fold-change criteria were used to identify differential peptides and proteins, for several theoretical fold variations of the spiked UPS1 standard. This experimental design allowed us to assess the performances of several workflows (4 based on spectral-count analysis and 4 based on MS signal analysis) in discovering true positive (UPS1 proteins successfully classified as variant) and avoiding false positive (yeast proteins

ACCEPTED MANUSCRIPT

erroneaously detected as variant). Overall, this study allowed to objectively evaluate label-free quantitative methods and concretely illustrate what one can expect from these approaches in terms of false discovery proportion and sensitivity for the detection of variant proteins.

Experimental procedures

Sample preparation. A yeast cell lysate was prepared in 8M urea / 0.1M ammonium bicarbonate buffer, protein concentration was adjusted at 8µg/µL after Bradford assay, and this lysate was used to resuspend and perform a serial dilution of the UPS1 standard mixture (Sigma). Twenty µL of each of the resulting samples, corresponding to 9 different spiked levels of UPS1 (respectively 0.05 – 0.125 – 0.250 – 0.5 – 2.5 - 5 – 12.5 - 25 - 50 fmol of UPS1 /µg of yeast lysate), were reduced with DTT and alkylated with iodoacetamide. The urea concentration was lowered to 1M by dilution, and proteins were digested in solution by addition of 2% of trypsin overnight. Enzymatic digestion was stopped by addition of TFA (0.5% final concentration).

NanoLC-MS/MS analysis. Samples (2µg of yeast cell lysate + different spiked level of UPS1) were analyzed in triplicate by nanoLC-MS/MS using a nanoRS UHPLC system (Dionex, Amsterdam, The Netherlands) coupled to an LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). 2 µL of each sample were loaded on a C-18 precolumn (300 µm ID x 5 mm, Dionex) at 20 µL/min in 5% acetonitrile, 0.05% TFA. After 5 minutes desalting, the precolumn was switched online with the analytical C-18 column (75 µm ID x 15 cm, in-house packed with C18 Reprosil) equilibrated in 95% solvent A (5% acetonitrile, 0.2% formic acid) and 5% solvent B (80% acetonitrile, 0.2% formic acid). Peptides were eluted using the following gradient of solvent B at 300 nL/min flow rate: 5 to 25% gradient during 75 min; 25 to 50% during 30min; 50 to 100% during 10min. The LTQ-

ACCEPTED MANUSCRIPT

Orbitrap Velos was operated in data-dependent acquisition mode with the XCalibur software. Survey scan MS were acquired in the Orbitrap on the 300-2000 m/z range with the resolution set to a value of 60000. The 20 most intense ions per survey scan were selected for CID fragmentation and the resulting fragments were analyzed in the linear trap (LTQ). Dynamic exclusion was employed within 60 seconds to prevent repetitive selection of the same peptide.

MS data processing. The dataset was processed according to different workflows listed in Table 1, consisting in the following steps: peaklist generation, database search, validation of the identified proteins and extraction of quantitative metric (spectral count or MS signal). According to the different tools used for each step, eight distinct workflows were evaluated. The same databases were used for peptide identifications: yeast database from UniprotKB (S_cerevisiae_ 20121108.fasta, 7798 sequences) and a compiled database containing the UPS1 human sequences (48 sequences).

Workflow 1: ExtractMSn / Mascot / MFPaQ / Spectral Counting. The Mascot Daemon software (version 2.4; Matrix Science, London, UK) was used to perform database searches, using the Extract_msn.exe macro provided with Xcalibur (version 2.0 SR2; Thermo Fisher Scientific) to generate peaklists. Parameters used for creation of the peaklists were: parent ions in the mass range 400–4500, no grouping of MS/MS scans, and threshold at 1000. Peaklists were submitted to Mascot database searches (version 2.4.2). ESI-TRAP was chosen as the instrument, trypsin/P as the enzyme and 2 missed cleavages were allowed. Precursor and fragment mass error tolerances were set at 5 ppm and 0.8 Da, respectively. Peptide variable modifications allowed during the search were: acetyl (Protein N-ter), oxidation (M), whereas carbamidomethyl (C) was set as fixed modification. To calculate the false discovery rate (FDR), the search was performed using the "decoy" option in Mascot. Validation was performed with an in-house developed module associated to MFPaQ [17] (http://mfpaq.sourceforge.net/), based on the target-decoy strategy, as described before [18].

ACCEPTED MANUSCRIPT

Briefly, FDR at peptide level was calculated as described in [19] and set at 5% by adjusting peptide p-value threshold. Validated peptides were assembled into protein groups following the principle of parsimony (Occam's razor) [20]. Protein groups were then validated to obtain a FDR of 1% at the protein level, by adjusting the threshold on a protein group score defined as the sum of peptide score offsets (difference between each peptide Mascot score and its homology or identity threshold). The total spectral count metric was extracted for each protein group by MFPaQ in each analytical run.

Workflow 2: Andromeda / MaxQuant / Spectral Counting. Acquired MS data were processed using MaxQuant version 1.3.0.5 [21]. Derived peak lists were submitted to the Andromeda search engine [22]) (www.maxquant.org). For database searches, the precursor mass tolerance was set to 20 ppm for first searches and 6ppm for main Andromeda database searches. The fragment ion mass tolerance was set to 0.5 Da. Trypsin/P was chosen as the enzyme and 2 missed cleavages were allowed. Oxidation of methionine and protein N-terminal acetylation were defined as variable modifications, and carbamidomethylation of cysteine was defined as a fixed modification. Minimum peptide length was set to six amino acids. Minimum number of unique peptides was set to one. Maximum FDR – calculated by employing a reverse database strategy – were set to 1% for peptides and proteins. Proteins identified as "reverse" and "only identified by site" were discarded from the list of identified proteins. In this particular workflow, total spectral count for each validated protein group was computed from msms.txt table.

Workflow 3: Mascot Distiller / Mascot / IRMa-hEIDI / Spectral Counting. Data were processed automatically using Mascot Distiller software (version 2.4.3.0, Matrix Science). ESI-TRAP was chosen as the instrument, trypsin/P as the enzyme and 2 missed cleavages were allowed. Precursor and fragment mass error tolerances were set at 5 ppm and 0.8 Da, respectively. Peptide variable modifications allowed during the search were: acetylation

ACCEPTED MANUSCRIPT

(Protein N-ter), oxidation (M), whereas carbamidomethyl (C) was set as fixed modification. The IRMa software v1.31 [23] was used to filter the results. Filters used were : (1) peptides whose score ≥ query homology threshold ($p<0.5$) and rank ≤ 1 are marked as significant; (2) Single match per query filter was: Move to ambiguous all peptides which aren't assigned to best protein for this query (best is higher protein score); (3) FDR seeker filter : Seek a 1% FDR based on score filtering; (4) Accession filter : Delete proteins coming from reverse database ; (5) Specific peptide filter : Accept only protein hits whose specific peptides count >= 1. The filtered results were then compiled and structured within dedicated relational Databases and a homemade tool (hEIDI) was used for the compilation, grouping and comparison of the proteins from the different samples, analytical replicates and conditions to compare (Hesse *et al.*, in preparation). In such workflow, total spectral count values calculated for each protein groups are used for quantification.

Workflow 4: ExtractMSn / Mascot / Scaffold / Spectral Counting. Peaklists generation and protein identifications were made as detailed in workflow 1. Mascot results were loaded into the Scaffold software (Version 3.6.5, Proteome Software, Portland, USA). To minimize false positive identifications, results were subjected to very stringent filtering criteria as follows. For the identification of proteins, a Mascot ion score had to be minimum 30 and above the 95% Mascot significance threshold ("Identity score"). The target-decoy database search allowed us to control and estimate the false positive identification rate of our study, and the final catalogue of proteins presented an estimated false discovery rate (FDR) below 5%. The spectral count metric used for quantitation corresponds to the Unweighted Spectrum Count values in Scaffold.

Workflow 5: ExtractMSn / Mascot / MFPaQ / MS Signal analysis. The first steps (peaklist creation, database search, validation) were the same than in workflow 1. Quantification of proteins was then performed using the label-free module implemented in the MFPaQ v4.0.0

ACCEPTED MANUSCRIPT

software, as previously described [18, 24]. Briefly, the software uses the validated identification results and retrieves the XIC of the identified peptide ions in the corresponding raw nanoLC-MS files, based on their experimentally measured RT and monoisotopic *m*/*z* values. Peptide ions identified in all the samples to be compared are used to build a retention time matrix and re-align in time LC-MS runs. For peptides not identified by MS/MS in a particular run, this re-alignment matrix is used to perform cross-assignment and extract their XIC signal starting from a predicted RT. Normalization across conditions is performed based on the median of XIC area ratios for all the extracted peptide ions. Protein quantification is based on a protein abundance index calculated as the average of XIC area values for at most three intense reference tryptic peptides per protein.

Workflow 6 and 7: Andromeda / MaxQuant / MS Signal analysis. The first steps (database search with Andromeda and validation) were the same as in workflow 2. For quantification purposes, either Intensities (workflow 6) or LFQ [25] (workflow 7) calculated by MaxQuant were used. The LFQ metric, as described in [25], is derived from the raw intensities by the MaxLFQ algorithm, which uses a specific normalization procedure, as well as a particular aggregation method to calculate protein intensities, by taking into account, for each protein, all the peptide ratios measured in all pair-wise comparisons of the different quantified samples. “Match between run” time window was set to 2 minutes. For LFQ quantification, only protein ratios calculated from at least two unique peptides ratios (min LFQ ratio count=2) were considered for calculation of the LFQ protein intensity.

Workflow 8: Mascot Distiller / Mascot / Skyline / MS Signal analysis. Peaklist creation was performed with Mascot Distiller as described in workflow 3, then database searches were performed with Mascot and validated with Scaffold as described for workflow 4. XIC signal corresponding to all validated peptides were extracted using the Skyline software [26]

ACCEPTED MANUSCRIPT

(Skyline version v2.5, daily updates of April 2014, https://skyline.gs.washington.edu). This method was well described by Schilling et al (Schilling et al, MCP, 2012). Total areas, corresponding to the sum of the 3 extracted isotopes areas, were used for statistical analysis.

Statistical analysis. For pairwise comparisons of samples spiked at different concentrations of UPS1, same statistical tests and fold-change criteria were applied to the quantitative data obtained from each workflow, as follows:

When working on spectral count metrics (workflows 1-2-3-4), a beta-binomial test was performed based on triplicate MS/MS analyses. p-values were calculated with the software package BetaBinomial_1.2 [27] implemented in R. Fold change was calculated as ratio of average spectral counts from both conditions. For proteins absent in all replicates of one specific condition, their spectral count values were modified by adding 1 spectrum to all 6 samples in order to be able to calculate a fold change for these particular proteins. To classify proteins as variant and non-variant and plot ROC curves, different combinations of criteria were tested ($|\log_2$ fold change$| > x$, from 0.8 to 3 ; p-value $< y$, from 0.05 to 0.0001).

When working on MS signal intensity-based metrics (workflows 5-6-7-8), proteins were filtered out if they were not quantified in at least all replicates from one condition. Missing protein intensity values were replaced by a constant value calculated independently for each sample as the 5-percentile value of the total population. A welch t-test (two-tailed t-test, unequal variances) based on triplicate MS analyses was then performed on $\log_2$ transformed values using the Perseus toolbox (version 1.4.0.11; http://141.61.102.17/perseus_doku). Criteria used to classify the proteins were the Welch t-test difference calculated by Perseus (difference between the two compared conditions of the mean $\log_2$ transformed value for triplicate MS/MS analyses), and the Welch t-test p-value. Results were filtered using different combinations of these criteria: |welch t-test difference| > x (from 0 to 7) and p-value < y (from 0.3 to 0.0001). z-score was also calculated as z-score= {(Welch t-test difference) - Median

ACCEPTED MANUSCRIPT

[(Welch t-test difference) for all quantified proteins] } / Standard deviation [(Welch t-test difference) for all quantified proteins].

Results

Experimental design, sample preparation and analysis. In order to evaluate different quantitative workflows in their ability to correctly detect known variant proteins in complex samples, we prepared a series of 9 yeast lysate samples spiked with growing concentrations of the Sigma UPS1 standard composed of an equimolar mixture of 48 human proteins. To that aim, UPS1 lyophilized proteins were directly resuspended using the yeast lysate prepared in urea buffer, and a serial dilution of this initial mixture was then performed using the same yeast lysate, resulting in spiked UPS1 concentrations ranging from 50amol/µg up to 50fmol/µg of yeast lysate. Protein samples were digested with trypsin, and resulting peptides were analyzed by nanoLC-MS/MS on a LTQ Velos-Orbitrap instrument, using routine chromatographic conditions (15cm C18 reverse-phase column, 2 hours gradient) and data-dependent acquisition MS parameters (resolution 60000 for MS survey scan, top 20 CID sequencing in the ion trap). Triplicate MS analysis was performed for each sample, resulting in 27 raw data files that were subsequently processed in different ways, using several computational workflows (**Table 1**). Two different softwares were used for protein identification (Mascot and Andromeda) and 5 solutions were employed for protein quantification (Scaffold, IRMa/hEIDI (Hesse et al, in preparation), MaxQuant [21, 22, 28], MFPaQ [17, 24] and Skyline [26, 29]), some of them generating a unique quantitative output, either spectral counting or MS signal extraction data, and some of them generating both types

ACCEPTED MANUSCRIPT

of quantitative data. Finally, 8 different quantitative datasets were obtained, as indicated in **Table 1** and described in details in the Experimental section.

We first evaluated the identification datasets in a qualitative way by simply reporting the number of identified and validated proteins for both the background (yeast proteins) and the spiked standard (UPS1 proteins) in each sample. **Sup data 1** shows the number of proteins identified by MS/MS sequencing and validated by various bioinformatics workflows. As expected, the total number of proteins, reflecting mainly the constant yeast background proteome, was fairly reproducible across triplicate MS analysis and across the series of spiked samples, whereas the number of identified UPS1 proteins increased with the spiked amount. While no UPS1 protein was correctly identified at a concentration of 500amol/µg (as none of the peptide sequence matches could be validated at this concentration), all 48 human proteins were sequenced and correctly identified at 50fmol/µg. From these results, we decided to select different concentration levels of UPS1 to perform pairwise quantitative comparisons of samples, trying to mimic distinct biochemical situations, as illustrated in **Figure 1**. Comparison A (500amol/µg versus 50fmol/µg) should reflect a case were a protein is typically under the detection level of the instrument in one condition, and strongly expressed in the other condition with a fold change of 100. In comparison B (5fmol/µg versus 50fmol/µg), the protein may be in turn detectable in both conditions, and strongly up-regulated with a fold change of 10. Finally, comparison C (12.5fmol/µg versus 25fmol/µg) should simulate a situation where the protein is detectable in both conditions, but only slightly up-regulated with a fold change of 2. Because "real-life" biological samples usually contain many proteins with a differential abundance, encompassing a wide range of absolute expression levels and fold change values, we tried to approximate such a situation by gathering together the quantitative data obtained for each binary comparison, after computational processing. Using this post-processing assembly of the 3 individual datasets,

ACCEPTED MANUSCRIPT

we composed a global quantitative dataset containing theoretically 144 variant proteins (UPS1 proteins issued from the 3 relative quantitative analyses, and thus expected to vary with a fold change of 100, 10 or 2), and a background of around 2500 non-variant yeast proteins (measured and quantified in the different pairwise comparisons) (see ref [30], **Sup Table 1)**. The generation of this synthetic dataset allowed us to illustrate, in a single representation, the performance of quantitative proteomic tools and methods, challenged with different situations.

The final aim of relative quantitative analysis in discovery proteomics is usually to identify differentially expressed proteins. Therefore, the tested informatics workflows were mainly evaluated in their ability to correctly detect the expected variants, rather than in the accuracy of the measured fold change. The experimental design and the spiked standard used here allowed us to unambiguously assess such performances by counting the number of true-positives (TP) and false-positives (FP), respectively UPS1 or yeast background proteins found to be differentially expressed. Clearly, the classification of proteins as variant (positive hits) or non-variant (negative hits) both relies on the one hand, on the accuracy of the quantitative metrics generated by the bioinformatics software, and on the other hand, on the performance of the statistical test and criteria used to discriminate the positive and negative populations. Here, we mainly tried to benchmark the former step of the workflow (extraction of quantitative metrics by informatics tools), and we didn't aim to evaluate statistical methods. We thus used a common, simple statistical test for protein classification, based either on the beta-binomial method for spectral count datasets [27], or on a modified t-test for datasets containing peptide intensity-based values (see Experimental section and below). Proteins were classified as variant or non-variant by a combined filtering on the p-value of this statistical test and on the fold change value, as very often performed in "real life " biological studies [31-34]. Following such classification, the sensitivity of the workflows for the detection of

ACCEPTED MANUSCRIPT

variant proteins (number of true positive hits relative to the real total number of variant proteins, i.e. TP/144), and false discovery proportion (FDP, defined as the number of false positive hits relative to the total number of proteins found as variant, i.e. FP/(TP+FP)) could easily be computed.

Performances of spectral counting for discrimination of variant proteins. **Figure 2A** shows the volcano plots obtained by applying spectral counting quantification methods, in which the $\log_{10}$(p-value) (calculated from the results of the BetaBinomial R package) is plotted against the calculated protein $\log_2$(fold change). As illustrated on these graphs, the majority of UPS1 proteins from comparison A and B (green and red populations, theoretical fold changes of respectively 100 and 10) were easily discriminated from the background of yeast proteins (grey), by both their p-values and fold changes. This was particularly the case with software tools such as IRMa/hEIDI and Scaffold. These results indicated the ability of the spectral count-based quantitative approaches to confidently detect protein variations of high to medium amplitude while minimizing the level of false discoveries. However, it can be noted that the UPS1 proteins quantified in the comparison C (12.5fmol/µg versus 25fmol/µg, yellow dots) were not well segregated from the background independently of the software used. Overall, these observations pointed out some limitations of quantification with spectral count data when dealing with low fold change variations or weakly concentrated proteins.

From these data, we tried to determine which criterion was best suited to retrieve significantly variant proteins. Sensitivity-FDP curves were plotted for the data obtained from the different workflows by using either the fold change or the p-value as a unique criterion to classify the proteins, and we further wanted to apply combinations of these filters to improve the classification. Resulting curves (**Sup data 2A**) show that the beta-binomial test was *per se* more efficient than a simple fold change to discriminate the TP from the TN. However, applying an additional fixed fold change cutoff improved significantly the results, as could be

ACCEPTED MANUSCRIPT

anticipated already from the volcano plots. On the dataset presented here, the best classification was obtained for all the workflows by applying this double-filtering approach with a threshold of 2 (or ½) on the fold change. Therefore sensitivity-FDP curves were plotted this way (variation of the p-value combined with a fixed threshold of 1 on the absolute $\log_2$(fold change)) for the different spectral count workflows as shown in **Figure 2B**. Globally, the best results were obtained with workflow 3 (Mascot/IRMa-hEIDI) which allowed for example to obtain a reasonable sensitivity (62%) with a very low FDP (4%) when setting a stringent p-value threshold of 0.001. Leveraging the p-value threshold at 0.0025 led to a slightly better sensitivity (67%) at the cost of a FDP increase to about 10%. Interestingly, in the case of workflows 3 (Mascot/IRMa-hEIDI) and 4 (Mascot/Scaffold), it was possible to reach really low FDP values by increasing the stringency on the p-value, showing the efficiency of these data processing tools for the exclusion of FP. Altogether, it turns out that spectral count approaches were very efficient for detecting high levels of variations on relatively abundant proteins, but tends to fail to reach high sensitivity on the present dataset which includes a population with moderate fold change variations. Markedly, very low levels of FDP can be reached with appropriate filtering.

Performances of MS intensity-based methods. **Figure 3A** shows the volcano plots from data obtained using different MS feature extraction tools ($-\log_{10}$(p-value) - calculated with the two-samples welch t-test from Perseus - plotted against the $\log_2$(fold change)). Conversely to what we observed with spectral-counting, the plots obtained with MS intensity-based techniques show that a large majority of UPS1 proteins quantified in the different pair-wise comparisons (green, red, and yellow populations) can be visually discriminated from the background of yeast proteins. While proteins with high signal levels and high theoretical fold changes were most often easily classified as variant (good p-values and high calculated fold

ACCEPTED MANUSCRIPT

changes), it can be noticed that even the UPS1 proteins quantified in the comparison C can be segregated from background, although with a partial overlap.

Here again we plotted different sensitivity-FDP curves by classifying the proteins either on their absolute fold change, on their welch t-test p-value, or by a combination of these criteria (setting up a fixed threshold for one of them and varying the other) (**Sup data2B**). In the case of MS intensity values obtained in our dataset, the fold change appeared to be generally a more efficient filter to discriminate TP from background than a simple statistical test based on the variance of the protein intensities. Indeed, the modified Welch t-test may produce a high number of FP hits on this particular dataset containing only three analytical replicates, finally leading to a high FDP after multiple testing. For example, on the MaxQuant LFQ dataset (workflow 7), filtering the proteins at a 0.05 cutoff only on the Welch p-value allowed to efficiently retrieve almost all UPS1 variant proteins (134 out of 144, e.g. 94% sensitivity), but with as many as 387 FP yeast proteins declared as variant (i.e. a final FDP of 74%). On the other hand, correction of the p-values for multiple-testing with methods such as the Benjamini-Hochberg (BH) procedure can be used to limit the number of FP and control the final FDR, but at the cost of a much lower sensitivity. For example, applying this correction on the same dataset and filtering afterwards with a BH adjusted p-value cutoff of 0.05 led to only 3 FP yeast proteins, but the number of TP UPS1 proteins also dropped to 50 (i.e. a calculated final FDP of 6%, close to the desired theoretical value, but a sensitivity of only 35%, see ref [30], **Sup Table 1**). Finally, combining fold change and Welch t-test p-value criteria emerged as the most discriminant approach, and allowed to reach good sensitivity with relatively low FDP. It has to be noticed that, unlike with the statistical t-test, setting a fold change threshold was quite sensitive to any shift in the population fold change distribution and to the optional normalization procedure applied in the workflows. Since some of the used methods contained a normalization step (e.g. MFPaQ or MaxQuant with the LFQ

ACCEPTED MANUSCRIPT

metric) and others not (e.g. Skyline or MaxQuant based on summed peptide intensity values), we used a z-score to avoid possible discrepancies between quantitative data depending on their origin. This z-score reflects, for each protein, the distance between the protein fold change and the mean of the population fold changes, relative to the standard deviation of this population (see Experimental procedures for calculation of the z-score). The combination of z-score and p-value criteria gave efficient discrimination results, as shown in **Sup data 2B**. For example, in the case of the MaxQuant LFQ workflow, we obtained a sensitivity of 94% and a calculated FDP of 8% when combining a |z-score| threshold of 1 and a Welch t-test p-value threshold of 0.05.

Figure 3B shows the sensitivity-FDP curves obtained for the MS intensity based workflows by varying the welch t-test p-value filter, with a fixed |z-score| cut-off of 1. Altogether, it appeared that the tested label-free tools based on MS signal analysis have the potential to be globally very sensitive (detect a large proportion of the true variant UPS1 proteins), with sensitivity values up to 94% when setting a p-value of 0.05. Comparative results for the different software are shown in **Table2** with sensitivity and FDP for this specific p-value. It has to be noticed however that all workflows produced still relatively elevated FDP values, that may be related to signal extraction errors by the softwares. The best compromise between sensitivity and FDP was obtained using the LFQ metric from MaxQuant [25] and the Top3 metric from MFPaQ [24].

Use of the spiked standard dataset to highlight data processing problems and optimize the workflows.

We next wanted to take advantage of this model dataset to identify quantification errors associated to the generation of false-negative (FN) and false-positive (FP) proteins, and illustrate a number of possible mistakes introduced by the different MS intensity based

ACCEPTED MANUSCRIPT

workflows. Protein quantification is a multi-step process, and possible errors associated to each of these steps may influence the final result. Obviously, processing steps based on peptide validation, grouping, and peptide-to-protein inference are important for final protein quantification. **Sup data 3A** illustrates a case where quantification based on non-specific peptides, shared between a stable yeast protein and a UPS1 variant protein (Ubiquitin-40S ribosomal protein S27a), compromised the result and led to classification of the spiked protein as a FN. Most of the time however, errors seem to take place at the signal extraction step itself. **Sup data 3B** shows a situation with overlapping isotopic patterns from several coeluting species, in which the MFPaQ software wrongly picked, in addition to the monoisopic peak of the correct peptide, the third and second isotope peaks from other species, as well as the monoisopic peak of a closely eluting isobaric peptide. Such errors could be avoided through a better recognition by the algorithms of peptide isotopic patterns. In addition, in the cases illustrated here, 16 peptides were correctly quantified for the protein, while signal extraction error occurred occasionally on a single peptide. Enabling the detection and elimination of outlier peptides with adequate testing procedures (option not enabled in that case) would alleviate such problems. Good alignment of LC-MS runs in retention time is also important for correct peak picking when cross assignment between runs is implemented. Some errors in Skyline could be attributed to wrong selection of a particular peptide in one of the runs in which the peptide was not sequenced by MS/MS, and in which XIC extraction was thus performed based on the RT of the peptide in another run (not shown). It must be noticed that tracking and eventually correcting these signal extraction errors is quite dependent on the software interface. To this respect, a software like MFPaQ offers a visualization interface that enables a rapid inspection of the XICs extracted for each peptide in the different conditions, and possibly unselects some of them to eliminate these peptides from the final quantification of the protein. However, it does not allow going back to raw MS data and correct for example

ACCEPTED MANUSCRIPT

the selection of the integration area directly on the chromatogram. This in turn is possible in Skyline, which really offers an interactive interface to efficiently review the results and manually correct possible mistakes. We thus wanted to take advantage of this feature and evaluate whether manual validation of the entire dataset was practically possible and how efficient it could be to improve the quantitative results. It took around 15 hours to manually check all the peptide ions from the dataset and either validate or correct the integration of the corresponding XIC. **Figure 4** shows the result of this exhaustive reviewing of the data on the accuracy of the quantitative result. While relatively time consuming, the manual correction clearly reduces the number of both false positive and false negative. The sensitivity was thus improved (from 88% using raw data to 97% after manual correction) and the FDP was significantly reduced (from 22% to around 9%) (sensitivity and FDP values calculated by filtering proteins based on a welch t-test p-value<0.05 calculated with the two-samples test from Perseus and |z-score|>1). In addition, the calculated fold changes were closer to the expected theoretical values. It appeared that most of the extraction errors generating false positive hits were related to low intensity signals, as illustrated in **Figure 4**. Finally, after manual correction, no more than 8 yeast proteins were classified as variant. Out of these 8 false positive hits, 3 contained peptides that were clearly "contaminated" with UPS1 peptides, 4 had very low intensity signals, and one of them was detected as variant while the expression profile of the related peptides did not follow that of UPS1 peptides. Altogether, these residual mistakes remaining after in-depth manual validation may reflect the minimal margin of error of the label-free, MS intensity-based quantification process, which may be difficult to reduce even by improving the automatic signal extraction algorithms of the software.

Discussion

ACCEPTED MANUSCRIPT

In this study, we generated a complex, spiked proteomic standard dataset, in which the ground truth is well characterized, and showed its utility for benchmarking label-free relative quantification computational workflows. Different protein standards have been used in the past to measure the performances of such software and data processing methods, ranging from simple mixtures of recombinant proteins, to complex cellular extracts spiked with a known amount of exogenous proteins. In the design of such a standard, it is important to be able to easily differentiate the spiked proteins from the background after the database search and identification process, in order to perform a correct classification of spiked (TP) and background (TN) molecules. The most straightforward approaches are either to apply some isotopic labeling on the background or the spiked samples, or to use sets of proteins from different species. Ideally, the number of spiked molecules should be large enough to provide a relevant statistical estimation of the sensitivity and FDP of the quantitative methods. Typically, samples can be spiked with recombinant purified proteins added in known quantities to the background, or with a much more complex sample, such as a biological extract from another species. In recent studies aiming at benchmarking software tools, such "double-proteome" samples have been used. For example, a mixture of lysates from human cells and from the *Streptococcus pyrogenes* bacterium at different ratios was used in a comparative study to show the performances of the OpenMS software [35]. Similarly, Cox *et al* used a complex digest of Hela cells, spiked with an *E.coli* digested cellular extract at two different amounts, creating a 3 fold variation of the *E.coli* proteins in the quantitative comparison [36]. In that later case, the spiked population represents a significant portion of the total sample (about one third of the identified proteins). Such a dataset may simulate particular biological experiments where a stimulation could for example induce a massive variation of the proteome, or some interaction proteomics experiments where a control is compared to an affinity purified sample containing many up-regulated proteins. However,

ACCEPTED MANUSCRIPT

normalization of such datasets may be difficult, because the usual hypothesis underlying normalization procedures is that the major part of the protein population remains stable, and the median of the fold change distribution should be 1. On the other hand, spiking a proteome background with a calibrated set of recombinant purified proteins is statistically less representative, as the number of TP decreases, but allows to simulate easily a classical expression proteomics experiment, in which a very minor part of the proteome will undergo a fold change. The UPS1 commercial standard, containing an equimolar mixture of 48 purified human proteins, represents a convenient sample for a spiking scheme experiment, and offers already a significant number of TP that allows to get an estimation of the sensitivity and FDP of the computational methods.

As software tools are expected to perform unequally depending on the fold change and amount of the spiked proteins, producing signals that will be more or less difficult to extract from the raw data according to their intensities, it is important to challenge them with different simulated variations. In a previous study, Cox et al spiked UPS1 in combination with the UPS2 standard, which contains the same proteins than UPS1, but distributed into 6 groups of decreasing concentration, spanning 5 orders of dynamic range [36]. By adding respectively these two standards into a background *E.coli* proteome, the authors simulated a situation where groups of proteins vary with different ratios, in a single pairwise comparison (6 analytical runs corresponding to 2 conditions with 3 technical replicates). However, in that case, only a small number of proteins are representative of each ratio, and many highly diluted UPS2 proteins are hardly detectable, creating a significant set of proteins which are differentially expressed but not really quantifiable.

In the present study, we chose to spike the UPS1 mixture at 9 different concentrations in a background yeast proteome, as described previously in Paulovich et al [6], and analyzed these samples in triplicate, resulting in a dataset of 27 runs. In order to artificially recreate a

ACCEPTED MANUSCRIPT

simulated dataset containing TP with different intensities and fold change values, we performed several pairwise comparisons by label-free quantification, and then combined the quantitative outputs.This approach has the benefit to illustrate the performances of the computational and statistical methods in a more comprehensive way. As a proof of principle, we show here the results obtained by simulating 3 kinds of variations (comparisons A, B and C: detection in only one condition; high fold change; moderate fold change). In principle, more comparisons could be performed and gathered to better approximate the inherent complexity of the variations that take place in a real biological experiment. For example, we didn't challenge here the software tools with comparisons involving only the more diluted spikes of the UPS1 concentration range, which would simulate variations of lower abundance proteins. Nevertheless, the different UPS1 spikes considered here could represent different types of biological samples, notably affinity purifications for large fold change analyses, or more classical proteome-wide analyses including moderate but significant expression fold change for some regulated proteins.

While label-free methods are more and more used for quantification of complex protein mixtures in biological studies, they are sometimes still considered as less accurate and reliable than label-based approaches. In addition, while many software tools for label-free quantification have been developed and are available, it may be difficult for an unexperienced user to choose a particular workflow. Finally, the quality of the results may be influenced by the parameter settings and the user's expertise with the programs. Consequently, test datasets are really needed to assess the performances of a given label-free workflow, adjust the parameters of a particular algorithm, and optimize post-processing methods such as missing value imputation, normalization, and statistical tests. The dataset presented here offers such possibilities, as illustrated on 8 different label-free pipelines which were objectively evaluated, and for which the number of FN and FP could be easily measured. The results

ACCEPTED MANUSCRIPT

obtained here show that label-free approaches are indeed efficient to detect variant proteins on the standard dataset. Globally, compared to signal extraction procedures, spectral counting workflows exhibited limited sensitivity (see **Sup data 4A**, showing overlaid ROC curves for both type of approaches). Even with lenient p-value cutoff, spectral count methods could only reach sensitivity levels up to 70-80%, mainly due to inefficiency to classify low abundance proteins with moderate fold change (comparison C). However, it must be noticed that they are easier to implement (shorter data processing time), and work quite well to sort out proteins with medium to high fold change (comparison A and B). Noticeably, they also proved to be quite specific, with the possibility to reach low level of FDP. Indeed, with data from such workflows, it was possible to set stringent filtering criteria and to almost completely avoid the detection of false positive yeast proteins, whereas this was much more difficult with MS intensity based methods (see below). Thus, as illustrated in **Sup data 4B**, at a given FDP level of e.g. 5%, spectral count approaches globally provided better sensitivity levels than MS intensity based approaches. In other words, if one is interested in the generation of a very "clean" and reduced list of differentially expressed proteins, the analysis of spectral count data with stringent filtering may represent a safe way to sort out very confident hits – probably with some compromise on sensitivity. Among spectral count workflows, coupling Mascot peptide identification with IRMa validation and hEIDI grouping and comparison ended up with the best compromise between sensitivity and FDP (**Fig 2B**). Indeed, even if retrieving the spectral count metric could *per se* be seen as a basic process which is not error-prone, depending on the workflow used, some differences in FDP were observed at the same sensitivity levels. In fact, spectral count approaches are still dependent on the quality of peptide validation, selection and grouping, which may directly influence the performances of the different software tools tested here.

ACCEPTED MANUSCRIPT

On the other hand, our results indicated that workflows based on signal extraction clearly have the potential to be globally very sensitive, and are effective in detecting large variations as well as accurately measuring moderate fold changes. Sensitivity levels up to 90-100% could be reached by relaxing filtering criteria. Thus, when admitting FDP levels higher than 10%, such workflows outperformed spectral count methods for the classification of differentially expressed proteins in the dataset (**Sup data 4B**). They represent promising approaches to detect variations even on minor proteins expressed at low level in the sample, and/or showing subtle changes. However, it has to be noticed that at present, software tools based on MS intensity analysis still generate a significant number of FN and FP. The presence of false positive hits (type I error) associated to statistical tests in multiple comparisons is a well documented problem when using high-throughput analytical methods which enable the quantification of hundreds or thousands of species. When a large number of statistical tests are performed, the final proportion of false discoveries (FDP) is actually larger than the user-specified p-value cutoff used for each individual test. Multiple testing correction procedures are classically used to adust the individual p-values of each gene or protein, and to control the final FDR, such as the Benjamini-Hochberg method. Interestingly, spiked datasets, such as the yeast-UPS1 dataset provided here, allow to experimentally measure this FDP rate as well as the associated sensitivity, and could represent a useful tool for optimization of statistical processing steps for proteomic data. The Benjamini-Hochberg adjustment, while very effective for controlling the final FDR of the process, appeared to be very conservative and reduced strongly the sensitivity of the workflows. In our hands, empiric filtering based on the combination of p-value and fold change (or z-score) cutoffs offered a more efficient compromise to obtain good sensitivity with relatively low levels of experimentally measured FDP, although this FDP was not formally controlled through the statistical process. Clearly, further studies will be needed to implement statistical methods allowing to control the FDR

ACCEPTED MANUSCRIPT

rate when looking for differentially expressed proteins in proteomic experiments. For example, while we used here arbitrary, fixed fold- change and p-value cutoffs, other approaches have been described in which the fold-change cutoff can be modulated as a function of the t-test P-value, to increase sensitivity for a given FDR after Benjamini-Hochberg correction [37]. Additionnally, pre-filtering can also be implemented to eliminate lowly abundant proteins which tend to give artificially high fold change values after spectral count quantification, and create false positives [37]. Finally, other statistical methods have bee proposed previously for microarray data in order to take into account a fold-change threshold of interest in a formal hypothesis test with FDR control [38-40].

The occurrence of FP and FN hits is also a problem that has to be tackled upstream of statistical processing, at the level of quantitative analysis and raw data processing tools, as these false hits are very often associated to signal extraction or matching problems (**Sup data 3**). Indeed, extraction of peptide intensity values is a complex process based on MS peak picking, isotope pattern and chromatographic peak recognition, and association of peptide features with MS/MS identification results, which can be complicated by the frequent occurrence of overlapping peptides in the LC-MS space. In our comparison, the MaxQuant software performed the best when using the LFQ metric (**Fig 3B**). In MaxQuant, the data analysis starts from the detection of features in the LC-MS map, based on recognition of elution peaks and peptide isotope profiles. In contrast, the processing in MFPaQ and Skyline is based on direct XICs extraction, using as a starting point m/z and RT coordinates derived from MS/MS identification results. Our study indicates that the later approach can however also produce good results, as illustrated by the good sensitivity and FDP obtained from MFPaQ quantification. A higher number of false positive hits were obtained with Skyline, which could be attributed in most instances to the absence of realignment procedure in the version of Skyline used for this study, and incorrect retrieving of peptide signals at deviated

ACCEPTED MANUSCRIPT

RT in some of the conditions. On the other hand, the interactive interface of Skyline allowed to efficiently check the signal extraction, and enabled an in–depth manual verification which clearly improved the final quantitative results, and particularly allowed to reduce the number of false-positive. The reduction of false-positive is an important challenge in label-free based discovery proteomic approaches, as it will directly influence the success of further validation steps, based on the selection of protein candidates from the first quantitative analysis. Although manual validation of the whole population of peptide ions, as performed in this study, is certainly overly long and impracticable in "real-life" biological studies, the ability to go back to the raw data for manual inspection of some specific proteins is probably an important feature for a label-free quantitative software. Indeed, the user can in this way really check the evidence for the differential expression of a protein, directly on the XIC and MS spectra of the different peptides. This manual verification can be performed on specific proteins that make biological sense (e.g. on some expected markers which would not be found as variants, due to signal extraction errors by the software, but also on new candidate proteins that will be subsequently selected for further validation studies, to ensure that these are not false positive).

In summary, our study on the presented standard dataset indicates that 1/ the number of false-positive hits from label-free quantitative analysis is still significant, even with the best performing workflows, 2/ that manual verification by the expert allows to reduce it, illustrating that there is still some margin of improvement for the automatic signal extraction step by label-free software, and 3/ that a residual number of errors remain inherently difficult to avoid, independently of the quality of the signal extraction procedure, particularly in the case of co-elution and overlapping peptide features, which would in turn require better resolution of both chromatographic and MS instruments. Ideally, label-free software should offer good performances in order to keep this number of FP relatively low, but also offer a

ACCEPTED MANUSCRIPT

user-friendly interface allowing to efficiently going back to the raw data and check the MS signal extraction on all the peptides of a particular candidate protein.

Conclusion

As outlined in previous reports, benchmark datasets are really needed to evaluate software algorithms in mass spectrometry–based protein analysis, and should be made freely available [41]. All raw MS data generated from the spiked standard presented here have been deposited to the ProteomeXchange Consortium [42] via the PRIDE partner repository with the dataset identifier PXD001819, and quantitative outputs from the different workflows tested are given in ref [30], **Sup Table 1**. It must be noticed that all these results are dependent on the parameter settings used for each computational workflow, and to this respect, one main utility of this model dataset may be to help the users in optimizing the tuning and finding the best parameters for a particular tool. Additionally, we hope that such spiked datasets could be useful for developers in order to efficiently test algorithms and improve the extraction of intensity metrics for protein quantitation. Finally, post-processing steps such as possible normalization, imputation of missing values, and downstream statistical analysis will also strongly influence the results. The use of spiked datasets could be beneficial to objectively evaluate their performances and their ability to reduce the level of FP and correctly classify variant proteins in large-scale studies.

ACCEPTED MANUSCRIPT

Notes:

The authors declare no competing financial interest.

Acknowledgements:

This work was funded through the French National Agency for Research (ANR) (grant ANR-10-INBS-08; ProFI project, "Infrastructures Nationales en Biologie et Santé"; "Investissements d'Avenir" call).

References

[1] Nahnsen S, Bielow C, Reinert K, Kohlbacher O. Tools for label-free peptide quantification. Molecular & cellular proteomics : MCP. 2013;12:549-56.
[2] Neilson KA, Ali NA, Muralidharan S, Mirzaei M, Mariani M, Assadourian G, et al. Less label, more free: approaches in label-free quantitative mass spectrometry. Proteomics. 2011;11:535-53.
[3] Sandin M, Teleman J, Malmstrom J, Levander F. Data processing methods and quality control strategies for label-free LC-MS protein quantification. Biochimica et biophysica acta. 2014;1844:29-41.
[4] Beasley-Green A, Bunk D, Rudnick P, Kilpatrick L, Phinney K. A proteomics performance standard to support measurement quality in proteomics. Proteomics. 2012;12:923-31.
[5] Rudnick PA, Clauser KR, Kilpatrick LE, Tchekhovskoi DV, Neta P, Blonder N, et al. Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. Molecular & cellular proteomics : MCP. 2010;9:225-41.
[6] Paulovich AG, Billheimer D, Ham AJ, Vega-Montoto L, Rudnick PA, Tabb DL, et al. Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. Molecular & cellular proteomics : MCP. 2010;9:242-54.
[7] Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. Genome biology. 2005;6:R16.
[8] Dabney AR, Storey JD. A reanalysis of a published Affymetrix GeneChip control dataset. Genome biology. 2006;7:401.
[9] De Hertogh B, De Meulder B, Berger F, Pierre M, Bareke E, Gaigneaux A, et al. A benchmark for statistical microarray data analysis that preserves actual biological and technical variance. BMC bioinformatics. 2010;11:17.
[10] Irizarry RA, Cope LM, Wu Z. Feature-level exploration of a published Affymetrix GeneChip control dataset. Genome biology. 2006;7:404.
[11] Irizarry RA, Wu Z, Jaffee HA. Comparison of Affymetrix GeneChip expression measures. Bioinformatics. 2006;22:789-94.
[12] Pearson RD. A comprehensive re-analysis of the Golden Spike data: towards a benchmark for differential expression methods. BMC bioinformatics. 2008;9:164.

ACCEPTED MANUSCRIPT

[13] Hoekman B, Breitling R, Suits F, Bischoff R, Horvatovich P. msCompare: A Framework for Quantitative Analysis of Label-free LC-MS Data for Comparative Candidate Biomarker Studies. Molecular & Cellular Proteomics. 2012;11.
[14] Tsou CC, Tsai CF, Tsui YH, Sudhir PR, Wang YT, Chen YJ, et al. IDEAL-Q, an automated tool for label-free quantitation analysis using an efficient peptide alignment approach and spectral data validation. Molecular & cellular proteomics : MCP. 2010;9:131-44.
[15] Zhang R BA, Brittenden J, Huang JT, Crowther D. Evaluation for Computational Platforms of LC-MS Based Label-Free Quantitative Proteomics: A Global View. J Proteomics Bioinform. 2010;3:260-5.
[16] Christin C, Hoefsloot HC, Smilde AK, Hoekman B, Suits F, Bischoff R, et al. A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. Molecular & cellular proteomics : MCP. 2013;12:263-76.
[17] Bouyssie D, Gonzalez de Peredo A, Mouton E, Albigot R, Roussel L, Ortega N, et al. Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: application to the proteomics study of membrane proteins from primary human endothelial cells. Molecular & cellular proteomics : MCP. 2007;6:1621-37.
[18] Gautier V, Mouton-Barbosa E, Bouyssie D, Delcourt N, Beau M, Girard JP, et al. Label-free quantification and shotgun analysis of complex proteomes by one-dimensional SDS-PAGE/NanoLC-MS: evaluation for the large scale analysis of inflammatory human endothelial cells. Molecular & cellular proteomics : MCP. 2012;11:527-39.
[19] Navarro P, Vazquez J. A refined method to calculate false discovery rates for peptide identification using decoy databases. Journal of proteome research. 2009;8:1792-6.
[20] Zhang B, Chambers MC, Tabb DL. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. Journal of proteome research. 2007;6:3549-57.
[21] Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nature biotechnology. 2008;26:1367-72.
[22] Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. Journal of proteome research. 2011;10:1794-805.
[23] Dupierris V, Masselon C, Court M, Kieffer-Jaquinod S, Bruley C. A toolbox for validation of mass spectrometry peptides identification and generation of database: IRMa. Bioinformatics. 2009;25:1980-1.
[24] Mouton-Barbosa E, Roux-Dalvai F, Bouyssie D, Berger F, Schmidt E, Righetti PG, et al. In-depth exploration of cerebrospinal fluid by combining peptide ligand library treatment and label-free protein quantification. Molecular & cellular proteomics : MCP. 2010;9:1006-21.
[25] Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Molecular & cellular proteomics : MCP. 2014;13:2513-26.
[26] MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics. 2010;26:966-8.
[27] Pham TV, Piersma SR, Warmoes M, Jimenez CR. On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. Bioinformatics. 2010;26:363-9.
[28] Cox J, Matic I, Hilger M, Nagaraj N, Selbach M, Olsen JV, et al. A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. Nature protocols. 2009;4:698-705.
[29] Schilling B, Rardin MJ, MacLean BX, Zawadzka AM, Frewen BE, Cusack MP, et al. Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: application to protein acetylation and phosphorylation. Molecular & cellular proteomics : MCP. 2012;11:202-14.

ACCEPTED MANUSCRIPT

[30] Ramus C, Hovasse A, Marcellin M, Hesse AM, Mouton-Barbosa E, Bouyssié D, et al. Spiked proteomic standard dataset for testing label-free quantitative software and statistical methods. Data in Brief.submitted.
[31] Albrethsen J, Agner J, Piersma SR, Hojrup P, Pham TV, Weldingh K, et al. Proteomic profiling of Mycobacterium tuberculosis identifies nutrient-starvation-responsive toxin-antitoxin systems. Molecular & cellular proteomics : MCP. 2013;12:1180-91.
[32] Bell C, English L, Boulais J, Chemali M, Caron-Lizotte O, Desjardins M, et al. Quantitative proteomics reveals the induction of mitophagy in tumor necrosis factor-alpha-activated (TNFalpha) macrophages. Molecular & cellular proteomics : MCP. 2013;12:2394-407.
[33] Ichikawa H, Yoshida A, Kanda T, Kosugi S, Ishikawa T, Hanyu T, et al. Prognostic significance of promyelocytic leukemia expression in gastrointestinal stromal tumor; integrated proteomic and transcriptomic analysis. Cancer science. 2015;106:115-24.
[34] Zhang L, Wang Z, Chen Y, Zhang C, Xie S, Cui Y, et al. Label-free proteomic analysis of PBMCs reveals gender differences in response to long-term antiretroviral therapy of HIV. Journal of proteomics. 2015;126:46-53.
[35] Weisser H, Nahnsen S, Grossmann J, Nilse L, Quandt A, Brauer H, et al. An automated pipeline for high-throughput label-free quantitative proteomics. Journal of proteome research. 2013;12:1628-44.
[36] Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. MaxLFQ allows accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction. Molecular & cellular proteomics : MCP. 2014.
[37] Carvalho PC, Yates JR, Barbosa VC. Improving the TFold test for differential shotgun proteomics. Bioinformatics. 2012;28:1652-4.
[38] Dembele D, Kastner P. Fold change rank ordering statistics: a new method for detecting differentially expressed genes. BMC bioinformatics. 2014;15:14.
[39] McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics. 2009;25:765-71.
[40] Vaes E, Khan M, Mombaerts P. Statistical analysis of differential gene expression relative to a fold change threshold on NanoString data of mouse odorant receptor genes. BMC bioinformatics. 2014;15:39.
[41] Yates JR, 3rd, Park SK, Delahunty CM, Xu T, Savas JN, Cociorva D, et al. Toward objective evaluation of proteomic algorithms. Nature methods. 2012;9:455-6.
[42] Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nature biotechnology. 2014;32:223-6.

ACCEPTED MANUSCRIPT

Tables

Table1: LC-MS quantification workflows evaluated. Combinations of tools were used for peaklist creation, database search, validation and quantification, resulting in 8 different workflows based on either spectral counting or MS signal extraction procedures, as described in details in the Experimental methods. The software tools used for spectral count quantification were Scaffold, IRMa/hEIDI, MaxQuant and MFPaQ. In the case of MS intensity-based quantification, protein intensity metrics were obtained from MFPaQ, MaxQuant or Skyline.

Workflow number	Peaklist creation device	Database search engine	Validation of identified proteins /spectral counting device	MS signal extraction device	Quantification method
1	ExtractMSn	Mascot	MFPaQ		Spectral counting
2	Andromeda	Andromeda	MaxQuant		Spectral counting
3	Mascot Distiller	Mascot	IRMa/hEIDI		Spectral counting
4	ExtractMSn	Mascot	Scaffold		Spectral counting
5	ExtractMSn	Mascot	MFPaQ	MFPaQ	MS signal analysis
6	Andromeda	Andromeda	MaxQuant	MaxQuant (Intensity)	MS signal analysis
7	Andromeda	Andromeda	MaxQuant	MaxQuant (LFQ)	MS signal analysis
8	Mascot Distiller	Mascot	Scaffold	Skyline	MS signal analysis

ACCEPTED MANUSCRIPT

Table 2: FDP and TPR obtained on the spiked dataset for different quantitative workflows.

Similar criteria were used for all workflows to classify proteins as variant (positive hits), i.e. |z-score|>1 and Welch t-test p-value<0.05. Human UPS1 proteins and yeast proteins verifying these criteria were counted respectively as True Positive and False Positive. False Discovery Proportion and True Positive Rate (sensitivity) were computed as described in the table.

	MFPaQ (workflow 5)	Maxquant Intensity (workflow 6)	Maxquant LFQ (workflow 7)	Skyline (workflow 8)
True Positive	135	130	134	126
False Positive	25	18	11	36
FDP=FP/(FP+TP)*100	16%	12%	8%	22%
TPR=TP/(TP+FN)*100	94%	90%	93%	88%

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Figure legends

Figure 1: Experimental design. A series of 9 yeast lysate samples spiked with growing concentrations of the Sigma UPS1 standard, was analyzed in triplicate by nanoLC-MS/MS mass spectrometry on a LTQ Velos-Orbitrap instrument. Different computational workflows were used to identify, validate, and quantify proteins based on spectral counting or MS signal analysis. In the present study, 3 different pairwise quantitative comparisons (A, B, and C) were performed between samples spiked with different amounts of UPS1, involving in each case the quantification of 6 raw files (2 conditions X 3 replicates), trying to mimic distinct biochemical situations. The 3 individual quantitative datasets containing protein abundance values were then gathered. This global quantitative dataset was generated for each data processing workflow, and identical downstream statistical processing methods were then applied for classification of variant proteins.

Figure 2: Quantitative results obtained with spectral counting workflows.

A/ Volcano plots ($-\log_{10}$(p-value) of the beta-binomial test versus protein $\log_2$(fold change)) are shown for the different software tools tested. The graphs illustrate the quantitative results for the UPS1 proteins quantified in each binary comparison (Green: comparison A, 0,5fmol/µg versus 50fmol/µg, theoretical fold change 100; Red: comparison B, 5fmol/µg versus 50fmol/µg, theoretical fold change 10; yellow: comparison C, 12.5fmol/µg versus 25fmol/µg, theoretical fold change 2). Grey dots correspond to yeast proteins quantified in all of these comparisons. Dotted lines represent a fixed p-value threshold of 0,001 and a fixed $|\log_2$(fold change)| threshold of 1.

B/ For each spectral count workflow, proteins of the mixed dataset (comparison A+B+C) were classified as variant after application of different p-value thresholds combined to a fixed $\log_2$(fold change) threshold of 1. The number of true positives (TP) and false positives (FP)

ACCEPTED MANUSCRIPT

was retrieved, and true positive rate (TPR or sensitivity = TP/144) was plotted as a function of false-discovery proportion (FDP=FP/(TP+FP)).

Figure 3: Quantitative results obtained with MS feature extraction workflows

A/ Volcano plots ($-\log_{10}$(p-value) of the Welch t-test versus protein welch t-test difference) are shown for the different software tools tested. As in Fig2, the graphs illustrate the quantitative results for the UPS1 proteins quantified in the different binary comparison A, B and C. Grey dots correspond to yeast proteins quantified in all of these comparisons. Dotted lines represent a fixed p-value threshold of 0.05 and a fixed |welch t-test difference| threshold of 1.

B/ For each MS signal analysis workflow, proteins of the mixed dataset (comparison A+B+C) were classified as variant after application of different p-value thresholds combined to a fixed |z-score| threshold of 1. TPR (sensitivity)= TP/144) was plotted as a function of false-discovery proportion (FDP=FP/(TP+FP)).

Figure 4: Manual feature-extraction correction in Skyline. The graphs illustrate the $\log_2$(fold change) calculated from protein intensity values in each binary comparison (A, B, and C) as a function of protein intensity. Protein intensity values were calculated as the sum of all peptide area values extracted by Skyline for each protein, and fold changes were computed from the mean of triplicate protein intensity values for each spiked concentration point. Results were plotted either from the raw Skyline output, or after an extensive manual check of all the peptide ions from the dataset (leading to either validate or correct the integration of the corresponding XIC, or eliminate the peptide from quantification). UPS1 proteins quantified in each binary comparison are represented as indicated in the legend, yeast proteins are represented either as grey dots (non-variant, true negatives) or blue crosses (variant, false-positives). Tables on the right indicate the number of proteins and peptides actually quantified in each case. Proteins were classified as variant after application of a p-value thresholds of

ACCEPTED MANUSCRIPT

0,05 combined to a fixed $\log_2$(fold change) threshold of 1. TPR (TP/144) and FDP(FP/(TP+FP)) are indicated after classification of the proteins individually for each binary comparison (A, B or C), or on the mixed dataset (comparison A+B+C).

Supporting information available: This material is available free of charge via the Internet at http://pubs.acs.org."

Sup Table1: Quantitative data obtained from the 8 different workflows

Sup data 1: Identification results. A series of 9 yeast lysate samples spiked with growing concentrations of the Sigma UPS1 standard, was analyzed in triplicate by nanoLC-MS/MS mass spectrometry on a LTQ Velos-Orbitrap instrument. Graphs indicate the average number of proteins, or the average number of UPS1 proteins, identified and validated in each different spiked sample, after data processing with ExtractMSn-Mascot-MFPaQ (workflows 1 and 5), Mascot Distiller-Mascot-Irma/Heidi (workflow 3), Mascot Distiller-Mascot-Scaffold (workflows 8) and Andromeda-MaxQuant (workflows 2, 6 and 7)

Sup data 2: evaluation of different filters to retrieve significantly variant proteins.

A/ Spectral count workflows: sensitivity-FDP curves were plotted for the data obtained from the different workflows by varying either the $\log_2$(fold change) threshold (red) or the beta-binomial test p-value threshold (blue). The fold change or p-value were used respectively as a unique criterion to classify the proteins (full line curves), or a combinations of these filters were applied to improve the classification (dotted line curves).

B/ MS intensity-based workflows: sensitivity-FDP curves were plotted for the data obtained from the different workflows by varying either the welch t-test difference threshold (red), the z-score threshold (green) or the welch t-test p-value threshold (blue). The welch t-test difference, z-score or p-value were used respectively as a unique criterion to classify the

ACCEPTED MANUSCRIPT

proteins (full line curves), or a combinations of these filters were applied to improve the classification (dotted line curves).

Sup data 3: Examples of quantification errors in label-free software tools.

A/ *Protein quantification based on non-specific peptides*. Upper panel: the Ubiquitin-40S ribosomal protein S27a protein from UPS1 (RS27A_HUMAN), contains a ubiquitin domain bearing strong homology with that of ubiquitin-containing proteins from the yeast background (RS27A_YEAST and RL402_YEAST). Colour code: grey: shared peptides between all three protein sequences; green: shared peptides between the two yeast proteins; blue and yellow: specific peptides for the two respective yeast proteins; red: specific peptide for the UPS1 spiked protein. Lower panel: XIC extraction in MFPaQ for the comparison B (5fmol/µg versus 50fmol/µg). For shared peptides, both the yeast stable proteins and the spiked protein contribute to the signal, that consequently exhibits only a moderate decrease in the low-spike condition. For the UPS1 specific peptide, the 10-fold decrease is accurately measured on the XIC. The PAI calculated for the top 3 most intense peptide without elimination of non-specific peptides leads to classification of the spiked protein as a false-negative.

B/ *Peptide signal error on an outlier peptide*. The spiked UPS1 protein ANXA5_HUMAN was quantified with 16 peptides in MFPaQ. The XIC extraction for the comparison A (0.5fmol/µg versus 50fmol/µg) was correctly performed for all of them except for the ion at m/z 447.2398 attributed to the QEISAAFK peptide (Upper panel). Examination of the raw spectra in Xcalibur indicates that the software also picked consecutively for this m/z value the third and second isotope peaks from other species (1 and 2), as well as a closely eluting isobaric peptide (3). The signal most probably belonging to the QEISAAFK peptide was actually only found in the high-spiked condition (4), but was finally summed by the software with that of contaminating yeast species. Consequently the ANXA5_HUMAN protein was incorrectly quantified with a fold change of about 2.5.

ACCEPTED MANUSCRIPT

***C/** Three examples of peptides picked wrongly by Skyline*

This figure shows a visualization of results in Skyline corresponding to 3 yeast peptides (Panel A, B et C). XIC were extracted from raw data for samples spiked with 0.5, 5, 12.5, 25, 50 fmol of UPS1 per µg of yeast background, respectively injected in triplicates.

Panel A: Protein URA2, (sp|P07259,Yeast), peptide TTAVNVIR

Panel B: Fimbrin (sp|P32599, Yeast), peptide LINDSVPDTIDTR

Panel C: Flavohemoprotein (sp|P39676, Yeast), peptide ENFPAGLVSEYLHK

A1, B1, C1: The left side shows wrong peak picking

A2, B2, C2: The right side shows manually corrected integration.

Each color (blue, purple and brown) corresponds to one precursor isotope (P, P+1 and P+2). In the chromatogram view (bottom left corner of the panel), vertical light blues lines indicate retention time of MS/MS spectra used to identify the peptide.

In each case, the wrong peak picking is easily detected thanks to a non-homogeneous profile between replicates in the retention time windows (square with dashed line).

Sup data 4: Comparison of spectral count versus intensity_based workflows. A/ Overlaid ROC curves for all workflows: proteins of the mixed dataset (comparison A+B+C) were classified as variant by filtering on the p-value thresholds, combined to a fixed |log2(fold change)| threshold of 1 for spectral-count workflows (1 to 4) and to a fixed |z-score| threshold of 1 for MS-intensity-based workflows (5 to 8). In each case, TPR (sensitivity)= TP/144 was plotted as a function of false-discovery proportion (FDP=FP/(TP+FP)). B/ Histograms indicate the sensitivity (TPR) level attained for a given value of FDP by the different workflows, based either on spectral counting (blue) or MS intensity analysis (orange).

ACCEPTED MANUSCRIPT

Figure 1: Experimental design

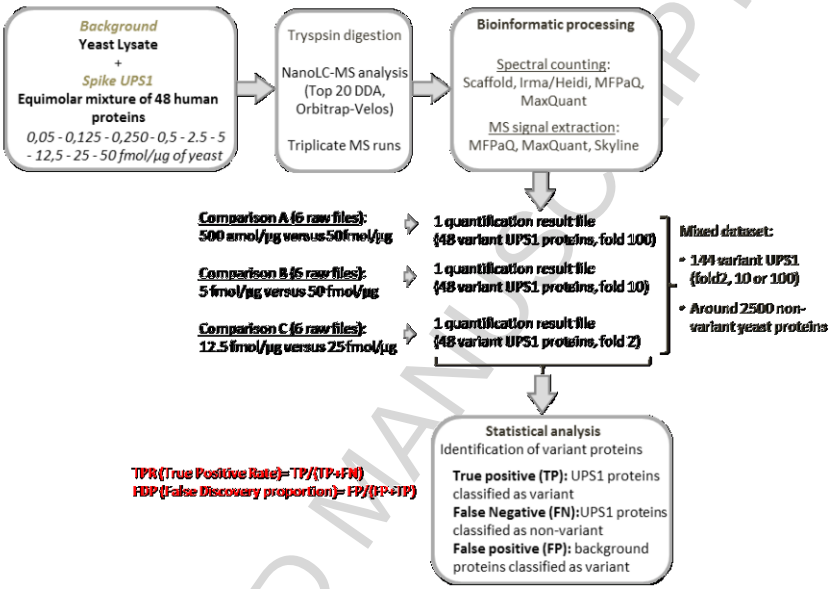
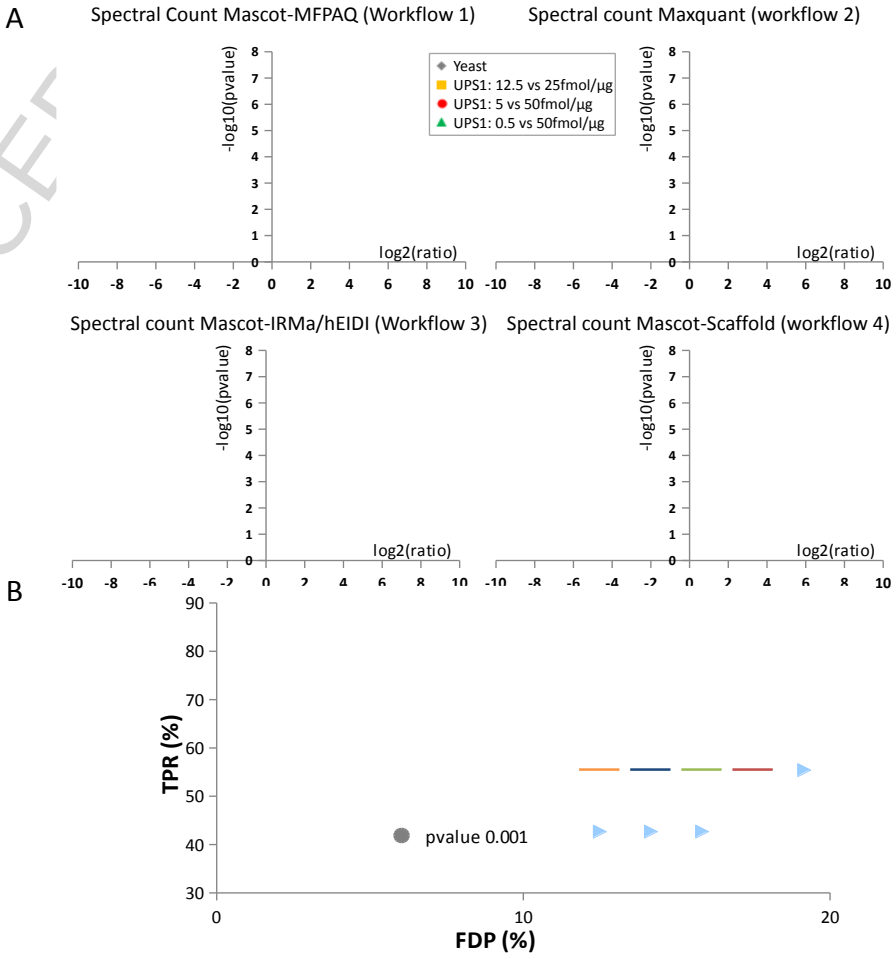


Figure 2: Quantitative results obtained with spectral count workflows



ACCEPTED MANUSCRIPT

Figure 3: Quantitative results obtained with software tools based on MS feature extraction

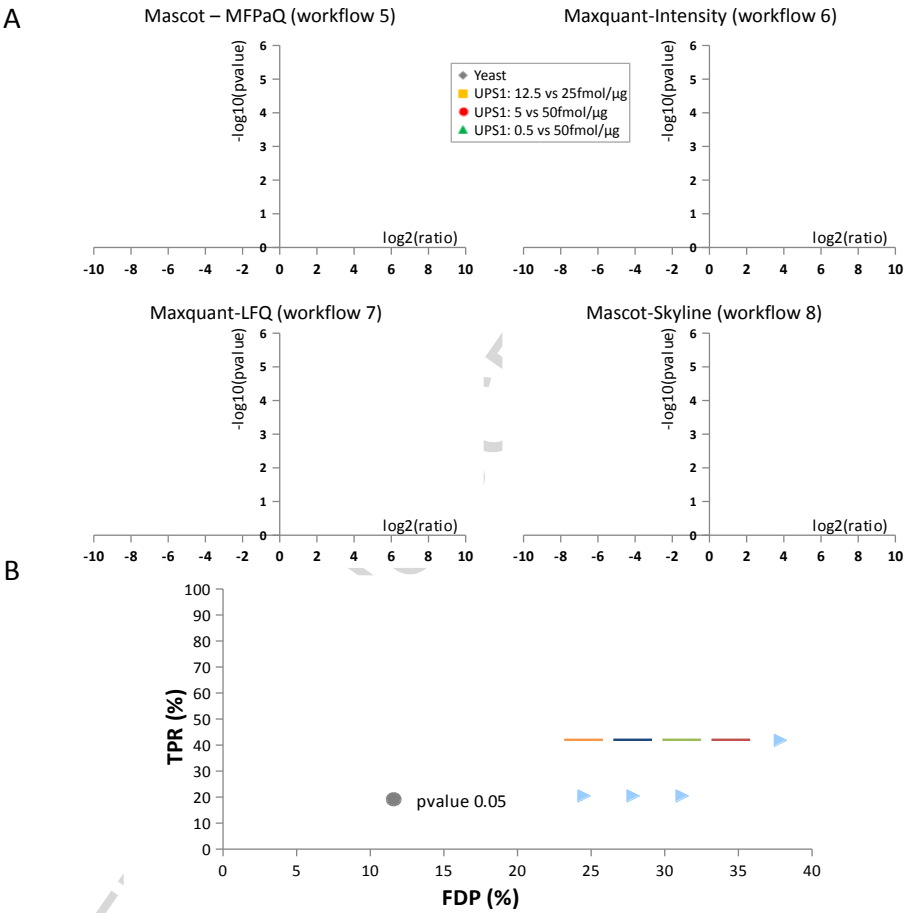
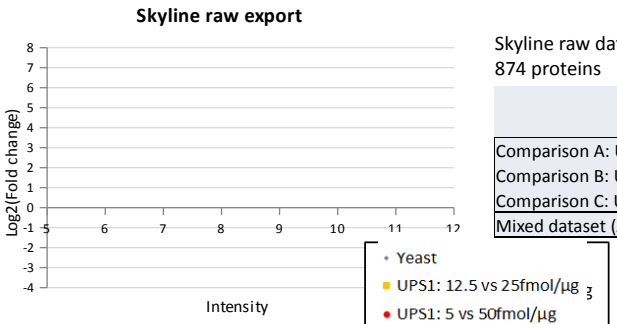
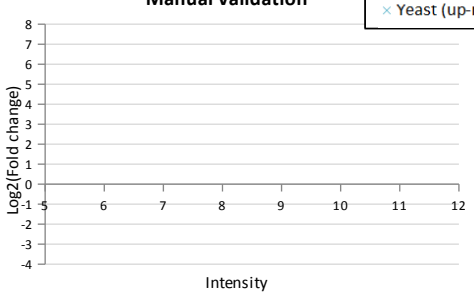


Figure 4: Manual feature-extraction correction in Skyline



Skyline raw data
874 proteins

	TPR	FDP
Comparison A: UPS1 0.5 vs 50fmol/µg	90%	23%
Comparison B: UPS1 5 vs 50fmol/µg	98%	29%
Comparison C: UPS1 12.5 vs 25fmol/µg	75%	10%
Mixed dataset (A+B+C)	88%	22%

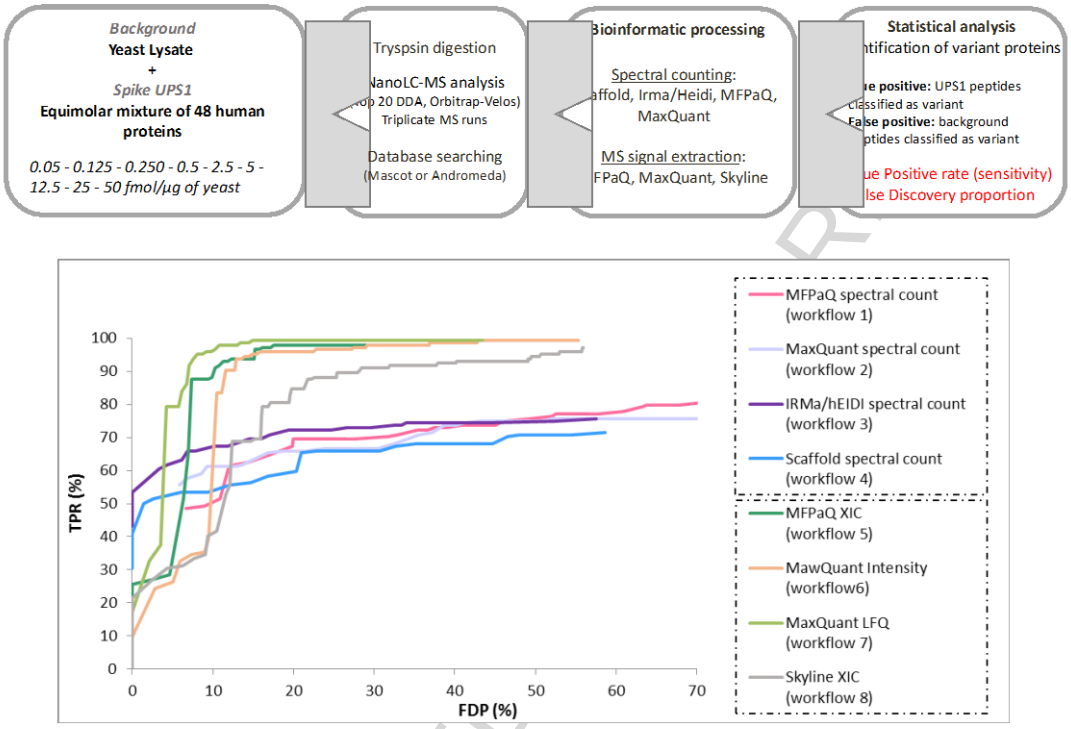


Skyline manual validation
803 proteins

	TPR	FDP
Comparison A: UPS1 0.5 vs 50fmol/µg	98%	6%
Comparison B: UPS1 5 vs 50fmol/µg	98%	11%
Comparison C: UPS1 12.5 vs 25fmol/µg	96%	10%
Mixed dataset (A+B+C)	97%	9%

ACCEPTED MANUSCRIPT

Graphical abstract



Highlights

- We provide a reference proteomic dataset, generated from a series of samples spiked with different amounts of a mixture of recombinant proteins, to test label-free quantitative methods
- We benchmarked several label-free workflows based either on spectral counting or on peptide ions MS signal analysis
- We evaluated the performances of different bioinformatic pipelines for detection of variant proteins with different absolute expression levels and fold change values