Peptide filtering differently affects the performances of XIC-based quantification methods

Isma Belouah, Melisande Blein-Nicolas, Thierry Balliau, Yves Gibon, Michel Zivy, Sophie S. Colombie

Peptide filtering differently affects the performances of XIC-based quantification methods

Isma Belouah[1,Θ], Mélisande Blein-Nicolas[2, Θ], Thierry Balliau[2], Yves Gibon[1], Michel Zivy[2,▲], Sophie Colombié[1]

[Θ] These authors contributed equally to this work

[1]UMR 1332 BFP, INRA, Univ Bordeaux, F33883 Villenave d'Ornon, France

[2]PAPPSO, GQE - Le Moulon, INRA, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, 91190 Gif-sur-Yvette, France.,

[▲]Corresponding author

zivy@moulon.inra.fr

UMR Génétique Quantitative et Evolution - le Moulon, PAPPSO,

Ferme du Moulon,

91190 Gif sur Yvette

01.69.33.23.65

ABSTRACT

In bottom-up proteomics, data are acquired on peptides resulting from proteolysis. In XIC-based quantification, the quality of the estimation of protein abundance depends on how peptide data are filtered and on which quantification method is used to express peptide intensity as protein abundance. So far, these two questions have been addressed independently. Here, we studied to what extent the relative performances of the quantification methods depend on the filters applied to peptide intensity data. To this end, we performed a spike-in experiment using Universal Protein Standard to evaluate the performances of five quantification methods in five datasets obtained after application of four peptide filters. Estimated protein abundances were not equally affected by filters depending on the computation mode and the type of data for quantification. Furthermore, we found that filters could have contrasting effects depending on the quantification objective. Intensity modeling proved to be the most robust method, providing the best results in the absence of any filter. However, the different quantification methods can achieve similar performances when appropriate peptide filters are used. Altogether, our findings provide insights into how best to handle intensity data according to the quantification objective and the experimental design.

INTRODUCTION

In bottom-up proteomics, proteins are digested into peptides which are subsequently separated by liquid chromatography (LC), ionized by electrospray and analyzed by tandem mass spectrometry (MS/MS). Peptide ions, and consequently the proteins from which they originate, can be quantified by integrating the signal intensities obtained from extracted ion currents (XIC;). This protein quantification approach, referred to as XIC-based quantification, is highly sensitive. It provides as many measurements as there are quantified peptide ions, so that in a given sample, each protein is measured as many times as it has peptide ions that have been assigned to it. These multiple measurements per protein allow robust quantification but they also represent a major difficulty. Not all the peptide intensities associated with a protein are equivalent for the following reasons: i) not all the peptides bear the same information (*e.g.* peptides shared by several proteins *vs* proteotypic peptides); ii) the ionization efficiency varies according to the peptide, so peptides belonging to a same protein will display different intensity levels ; iii) some peptide ions may be incorrectly identified; iv) some peptide ions may be incorrectly quantified due to mis-cleavages or other technical issues; and v) the abundances of some peptide ions do not reflect the abundance of their corresponding proteins because of post-translational modifications. Therefore, if not properly considered, peptide ions can introduce errors when computing protein abundances.

To reduce these errors, different approaches have been proposed. The statistical and probabilistic approaches rely on a modeling framework for computing protein abundances from quantified peptides. These approaches have been used to include shared peptides to improve protein quantitation (e.g.) and to handle missing data and/or outlying measurements (e.g.). Although they allow to fully exploit the information collected by the mass spectrometers, these approaches have not been widely used by the proteomics community so far, probably because of their complexity and of their requirement in computing time to analyze large datasets. As an alternative, several authors filter the peptide data before computing protein abundances. There are four types of filter. First, there is the *shared peptide filter*. Although they constitute a valuable source of information , shared peptides are generally discarded because it is difficult to properly deconvolve the information they carry. Second, there is *the retention time (RT) filter,* which aims to remove peptide ions showing highly variable RT potentially arising from mis-identifications. Various methods have been used, based on the standard deviation of RT or on RT clustering . Third, there is the *occurrence filter*, which aims to remove peptide ions exhibiting many missing values. These

peptide ions may be associated with dubious intensities if missing values are due to problems in RT alignment or in peak detection. However, they may also be associated with valuable intensities if missing values arise from biological mechanisms (for example if the protein is not expressed) or from technical limitations (if intensities are below the detection threshold). As for shared peptides, rarely observed peptide ions are difficult to handle so one way around this problem is to remove them. Generally, a threshold is chosen arbitrarily, *e.g.* a peptide ion should be observed in at least three injections . More refined approaches have also been proposed, taking experimental groups into account so that statistical tests can be performed properly or based on a model filtering routine to select peptide ion sets that produce optimal information content [7]. Fourth, there is the *outliers filter*, which aims to exclude peptide ions showing inconsistent intensity profiles. Several approaches have been proposed based on Grubbs' test , the coefficient of variation , the peptide ion correlation or covariation .

To obtain a final protein abundance value, the intensities of the peptide ions remaining after filtering must be summed. In the case of data-dependent analysis where intensity data are collected in MS1, several quantification methods have been proposed in the last fifteen years (methods employed in acquisition approaches where intensity data are collected in MS2 such as data-independent analysis or targeted quantification are outside the scope of the present study). Six of them are commonly used: i) *Average*, which is the mean of intensities of all the peptide ions; ii) *iBAQ*, which is the sum of intensities of all peptide ions matching to a protein divided by the number of theoretically observable peptides ; iii) *TOP3*, which is the mean of intensities of the three most intense peptide ions ; iv) *Average-Log*, which is the mean of log-intensities of all the peptide ions [18]; v) *Model*, which is the adjusted mean of intensities of all the peptide ions computed using linear models and vi) *maxLFQ* implemented in maxQuant , which computes protein abundances based on a system of equations built from pair-wise peptide intensity ratios. *TOP3* and *iBAQ* were more specifically developed for absolute quantification while *Average, Average-Log* and *maxLFQ* are widely used for relative quantification. *Model* is recommended by some authors as the most adequate method to infer and quantitatively compare protein abundances . Although the relative performances of these quantification methods have been evaluated repeatedly, no clear consensus has emerged so far.

To explain this lack of consensus, we assume that the relative performances of quantification methods depend on the quality of the dataset considered and that similar performances can be achieved by using peptide filters appropriate to each method. As the

weight of a peptide ion in the computation of a protein abundance depends on the computation mode used and thus on the quantification method, one may expect peptide filters to have different effects depending on the method. To confirm these assumptions, we performed a spike-in experiment using UPS1 standard to evaluate the performances of five quantification methods in different datasets combining zero to four of the filter types previously mentioned. The five quantification methods included those mentioned above except maxLFQ, as it required the use of a non-open source program, which precluded the analysis of the effect of the different filters.

MATERIAL AND METHODS

Yeast growth

Saccharomyces cerevisiae strain S288C was inoculated in 5 ml YPD (Yeast extract Peptone Dextrose) medium containing yeast extract (10 g l^{-1}; Difco Laboratories, Detroit, Michigan), bacteriological peptone (20 g l^{-1}; Difco) and glucose (20 g l^{-1}). After 24 h of growth at 30 °C under agitation, the culture medium was centrifuged (2 750 g, 10 °C, 3 min) and the supernatant was discarded. The remaining yeast cells pellet was rinsed twice with 5 ml cold distilled water, frozen in liquid nitrogen and stored at -80 °C for subsequent protein extraction.

Yeast protein extraction

Proteins were extracted by suspending the pellet of yeast cells in 500 µl of an ice-cold extraction/precipitation solution of acetone containing trichloroacetic acid (10%) and β2-mercaptoethanol (0.07%). To promote cell wall disruption, cells were ground for 5 min with 200 µl of glass beads. The protein extract was then shortly vortexed for homogenization and immediately transferred to new vials to remove glass beads. 750 µl of the extraction/precipitation solution were added to the protein extract before incubation (-20 °C for 90 min) and centrifugation (19 283 g, 0 °C, 15 min). The supernatant was removed, and the remaining protein extract was re-suspended in 1.8 ml cold washing acetone solution containing 0.07% β2-mercaptoethanol, incubated (1 h at -20 °C) and then centrifuged (19 283 g, 0 °C, 10 min). This step was repeated twice. After the last washing, the protein pellet was

dried in a vacuum centrifuge, weighed and solubilized by adding 15 µl per mg of pellet of a solubilization buffer (6M urea, 2M thiourea, 10mM dithiothreitol (DTT), 30 mM Tris-HCl at pH 8.8, 0.1% zwitterionic acid labile surfactant (ZALS)). Remaining cellular debris was segregated from soluble proteins by centrifugation (15 000 g, 25 °C, 25 min). Protein concentration was determined using the PlusOne 2-D Quant Kit (GE Healthcare, Little Chalfont, UK) and adjusted with the solubilization buffer to 0.887 µg μl^{-1}.

Spike-in UPS1 preparation

Dried UPS1 proteins (Sigma-Aldrich) were solubilized in the buffer containing yeast proteins to a final concentration of 0.75 µg μl^{-1} (0.625 fmol μl^{-1} of each UPS1 protein) so that the total protein (yeast + UPS) concentration was 1.637 µg μl^{-1}. Proteins were incubated for one hour at room temperature for reduction by the 10 mM DTT present in the buffer. Thereafter, proteins were alkylated for one hour in the presence of 50 mM iodoacetamide and diluted with 50 mM ammonium bicarbonate to decrease the total urea and thiourea concentration to 3.6 M before being twice digested. A first 4-hour digestion was performed with 1/32 (w/w) rLysC protease (Promega). After dilution with a solution of 50 mM ammonium bicarbonate to decrease the total urea and thiourea concentration to 0.77 M, a second overnight digestion was performed with 1/32 (w/w) trypsin (Promega). Both rLysC and trypsin digestion were performed at 37 °C. Trypsin digestion was stopped by acidification (1% total volume trifluoroacetic acid). The resulting peptides were purified on solid-phase extraction using a polymeric C18 column (Phenomenex) with a washing solution containing 0.06% acetic acid and 3% acetonitrile (ACN). After elution with 0.06% acetic acid and 40% ACN, peptides were speedvac-dried and suspended in a solution containing 2% ACN, 0.06% trifluoroacetic acid and 0.06% formic acid so that the concentration of each UPS1 peptide was 141.1 fmol μl^{-1} and the total concentration of yeast peptides was 200 ng μl^{-1}. A serial 2.25-fold dilution was prepared by mixing 6.7 µl of UPS1-yeast peptide mix with 8.3 µl of solubilized yeast peptides at 200 ng μl^{-1} until a UPS1 peptide concentration of 0.04 fmol μl^{-1} was reached. Eleven samples were thus obtained, containing 141.1, 62.8, 27.9, 12.4, 5.5, 2.2, 1.1, 0.5, 0.2, 0.09 and 0.04 fmol μl^{-1} of each UPS1 peptide. This serial dilution was performed in three replicates from aliquots of the same yeast culture, thus producing 33 samples.

LS-MS/MS analyses

LC-MS/MS analyses were performed using a NanoLC-Ultra System (nano2DUltra, Eksigent, Les Ulis, France) connected to a Q-Exactive mass spectrometer (Thermo Electron, Waltham, MA, USA). For each sample, 4 µl of protein digest were loaded onto a Biosphere C18 precolumn (0.1 × 20 mm, 100 Å, 5 µm; Nanoseparation) at 7.5 µl min^{-1} and desalted with 0.1% formic acid and 2% ACN. After 3 min, the pre-column was connected to a Biosphere C18 nanocolumn (0.075 × 300 mm, 100 Å, 3 µm; Nanoseparation). Electrospray ionization was performed at 1.3 kV with an uncoated capillary probe (10 µm tip inner diameter; New Objective, Woburn, MA, USA). Buffers were 0.1% formic acid in water (A) and 0.1% formic acid and 100% ACN (B). Peptides were separated using a linear gradient from 5 to 35% buffer B for 110 min at 300 nl min^{-1}. One run took 120 min, including the regeneration step at 95% buffer B and the equilibration step at 100% buffer A.

Peptide ions were analyzed using Xcalibur 2.1 (Thermo Electron) with the following data-dependent acquisition steps: (1) MS scan (mass-to-charge ratio (m/z) 300 to 1 400, 70 000 resolution, profile mode), (2) MS/MS (17 500 resolution, normalized collision energy of 30, profile mode). Step 2 was repeated for the eight major ions detected in step (1). Dynamic exclusion was set to 30 seconds. Xcalibur raw datafiles were transformed to mzXML open source format using msconvert software in the ProteoWizard 3.0.3706 package . During conversion, MS and MS/MS data were centroided. The raw MS output files and protein abundances were deposited on-line using PROTICdb database at the following URL: http://moulon.inra.fr/protic/filtering_quanti_methods. They are currently available with the following username: filtering and password: review. The mass spectrometry proteomics data have also been deposited with the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD009740. They are currently available with the following username: reviewer32109@ebi.ac.uk and password: JH5JcHXE. They will be made freely available after publication.

Protein identification

Protein identification was performed using the protein sequence database of *S. cerevisiae* strain S288c downloaded from the Saccharomyces Genome Database (SGD project, http://www.yeastgenome.org/, version dated 13/01/2015) and the sequences of UPS1 proteins available at http://www.sigmaaldrich.com/content/dam/sigma-aldrich/life-science/proteomics-and-protein/ups1-ups2-sequences.fasta. A contaminant database

containing the sequences of standard contaminants was also interrogated. The decoy database comprised the reverse sequences of yeast and UPS1 proteins. Database search was performed with X!Tandem (version 2015.04.01.1;) using the following settings. Carboxyamidomethylation of cysteine residues was set to static modification. Oxidation of methionine residues, N-terminal acetylation with or without excision of the N-terminal methionine, deamination of N-terminal glutamine and of carbamidomethylated cysteines and loss of H_2O from N-terminal asparagins were set to possible modifications. In the refine mode, excision of signal peptides was searched on the 50 first N-terminal amino acids possibly acetylated. Precursor mass precision was set to 10 ppm. Fragment mass tolerance was 0.02 Thomson (Th, unit of m/z). Only peptides with an E-value smaller than 0.05 were reported.

Identified proteins were filtered and sorted by using X!TandemPipeline (version 3.4.0,). Criteria used for protein identification were (i) at least two different peptides identified with an E-value smaller than 0.01 and (ii) a protein E-value (product of unique peptide E-values) smaller than 10^{-5}. Using these criteria, peptide and protein false discovery rates were 0.034% and 0 %, respectively.

Peptide ion quantification and intensity data filtering

Peptide ions were quantified according to extracted ion chromatograms (XIC) using MassChroQ software version 2.2 with the following parameters: "ms2_1" alignment method, tendency_halfwindow of 10, MS1 smoothing halfwindow of 0, MS2 smoothing halfwindow of 15, "quant1" quantification method, XIC extraction based on max, min and max ppm range of 10, anti-spike half of 5, background half median of 5, background half min max of 20, detection thresholds on min and max at 30 000 and 50 000, respectively, peak post-matching mode, ni min abundance of 0.1. The peptide intensities thus obtained constituted the initial dataset (Dataset 0), which was used to derive five datasets combining zero to four filters (Figure 1).

In the first dataset (Dataset 1), no filter was applied. Yeast peptide intensities were normalized to take possible global quantitative variations between LC-MS runs into account. For this, we used a local normalization method adapted from Lyutvinskiy *et al.* [30] and described in Millan-Oropeza *et al.* . In the second dataset (Dataset 2) one filter was applied: after normalization of yeast peptide intensities as described above, shared peptides

were removed (*shared peptide filter*). The third dataset (Dataset 3) comprised two filters. Peptides with a standard deviation of retention time higher than 30 seconds were first removed (*RT filter*). Since these peptides were considered as dubious, this filter was applied before normalization of yeast peptide intensities. Then, shared peptides were removed. The fourth dataset (Dataset 4) comprised three filters. It was obtained by applying an *occurrence filter* to Dataset 3, which resulted in the selection of peptide ions quantified in at least 28 samples, with no more than one missing value per UPS1 concentration. Thus, a maximum of 15.15% of missing values per peptide ion was tolerated and the selected peptide ions were quantified in at least two replicates for each UPS1 concentration. To ensure the quality of normalization, which depends on the number of peptide ions quantified both in a sample chosen as reference and in a sample to be normalized, we decided to apply this filter after normalization. Several peptide ions removed by the *occurrence filter* are good quality peptides whose intensities may fall below the detection threshold because their ionization efficiency is low. The fifth dataset (Dataset 5) comprised four filters and was obtained by applying an *outliers filter* to Dataset 4. To this end, Pearson correlations between log10-transformed intensities were computed for each pair of peptide ions belonging to the same protein. To avoid bias induced by outlier values due to individual technical variations, the correlations were computed on mean values of peptide ion intensities per concentration. The peptide ion with the highest number of coefficients of correlation greater than or equal to the mean of the positive coefficients of correlation was chosen as a reference for the protein. The peptide ions showing a non-significant correlation to the reference (p-value >= 0.01) or whose coefficients of correlation to the reference were lower than 0.8 were considered as outliers and were removed (*outliers filter*). Proteins quantified by fewer than two peptide ions were removed from all the datasets. Missing intensity values were not imputed. Consequently, the number of peptide ions used to compute protein abundances could vary from one sample to another.

Protein quantification

For each protein, five methods were used to compute abundances: i) *iBAQ* : the sum of peptide ion intensities was divided by the theoretical number of tryptic peptides; ii) *TOP3* : the three most intense peptide ions in median were selected and their mean intensity was computed. When one of the three most intense peptide ions was missing in a sample, *TOP3* was computed from the two remaining ones; iii) *Average*: the mean of all peptide ion

intensities was computed, iv) *Average-Log* [18]: peptide ion intensities were log10-transformed before their mean was computed; v) *Model*: log10-transformed intensities were modeled using a mixed effects model derived from Blein-Nicolas *et al.*:

$$I_{ijk} = \mu + A_i + R_j + P_k + \theta_{ij} + \varepsilon_{ijk}$$

where $\theta_{ijk} \sim N(0, \sigma\theta^2)$

$\varepsilon_{ijk} \sim N(0, \sigma\varepsilon^2)$

where I_{ijk} is the intensity measured for peptide ion k in serial dilution j (with j = 1, 2 or 3) at UPS1 concentration i; μ is the overall mean; the terms A_i R_j and P_k represent the effect due to UPS1 concentration i; serial dilution j and ionization efficiency of peptide k (also called peptide effect) respectively; Θ_{ij} represents the technical variation due to sample handling and injection in the mass spectrometer; $\boldsymbol{\varepsilon}_{ijk}$ is the residual error. *Model* was fitted with sum contrasts by maximizing the restricted log-likelihood. This allowed us to estimate the effects of P_k and Θ_{ij} and to subtract them from log10-transformed intensities. By doing so, we could subsequently compute protein abundances as adjusted mean intensities whose undesirable effects (P_k, Θ_{ij},) were removed. Log-abundances obtained by *Average-Log* and *Model* were converted to abundances for further analyses. All data analyses and graphical representations were performed using R version 3.3.2 . R scripts as well as quantification data are available at http://moulon.inra.fr/protic/filtering_quanti_methods (temporary username: filtering and password: review).

RESULTS AND DISCUSSION

We evaluated the crossed effects of peptide filters and quantification methods on the performances of protein quantification using a spike-in experiment where different concentrations of UPS1 proteins were added to a constant yeast background. Four filters were used: the *shared peptide filter*, the *RT filter*, the *occurrence filter* and the *outliers filter.* When applied separately, the filters exhibited some overlap since a number of peptide ions were removed by both the *shared peptide* and the *outliers filters*, the *RT* and the *occurrence filters* or the *occurrence* and the *outliers filters.* However, each filter also allowed us to remove many peptides (Figure S1). To take advantage of their complementarity, we applied these filters in cascade as described in Figure 1 (see Material & Methods for details), thus obtaining five datasets combining zero to four filters.

As the *shared peptide*, the *RT* and the *occurrence filters* discard peptide ions on the basis of their own characteristics, which do not depend on other peptide ions, the order in which these three filters are applied does not change the composition of the final dataset. This is not the case for the *outliers filter*, whose criterion of exclusion is based on the correlation with the other peptides of the same protein: the result of this filter can thus be influenced by the application of prior filters. As it is not appropriate to define outliers on the basis of peptide ions that will finally be discarded by other filters, we applied the *outliers filter* at the end. For each of the five datasets, five quantification methods, referred to as *iBAQ*, *TOP3*, *Average*, *Average-Log* and *Model*, were used to compute protein abundances.

1. The amplitude of peptide filtering affects protein data composition

Yeast and UPS1 datasets were differently affected by the filters. The proportion of shared peptides removed was much higher for yeast than for the UPS1 standard (-4.2% *vs* -0.8%, respectively). Although the UPS1 standard was designed to contain few proteins with similar sequences, yeast is a living organism that contains many duplicated genes resulting from whole genome duplication and other small-scale duplications .

The *occurrence* and *outliers filter*s were those that most drastically reduced the whole dataset (-38% and -64% peptide ions, respectively; -26.9% and -32.4% proteins, respectively). At the peptide level, the *occurrence filter* removed two-fold more UPS1 peptide ions than the yeast peptide ions (77.1% *vs* 35.9%, respectively). This is because UPS1 proteins have a wide dynamic range while yeast proteins are in constant amounts. The detectability of a peptide at a given protein concentration depends on its ionization efficiency: a peptide with a high ionization efficiency can be detectable even at low protein concentration, while a peptide with a low ionization efficiency will be detectable only if the protein concentration is high enough. Consequently, when the protein dynamic range is wide, peptides with low ionization efficiency are more subject to qualitative variations than those with high ionization efficiency. At the protein level, the *occurrence filter* also had a high impact on the number of quantified UPS1 proteins (-12.2%), mainly excluding small proteins quantified with few peptide ions (Figure S2). These proteins were probably represented mostly by peptides with a low ionization efficiency. Although these proteins were removed from the quantitative analysis, the information they carry was not completely lost as their

abundance variations can still be analyzed semi-quantitatively by using a spectral counting approach.

The *outliers filter* reduced yeast data more drastically than UPS1 data, both at the peptide level (-65% yeast peptide ions *vs* -12.6% UPS1 peptide ions, respectively) and at the protein level (-33.1% yeast proteins *vs* -2.8% UPS1 proteins, respectively). This was expected because the *outliers filter* is based on the correlation between peptide ions. Since the amount of yeast peptide ions was constant across the samples, they necessarily exhibited poor correlations. This is why the *outliers filter* not only has the advantage of removing peptide ions with dubious intensity profiles; it also allows proteins showing abundance variations in response to a treatment of interest (here the UPS1 concentration) to be selected. However, this characteristic can become a disadvantage if the objective is to obtain abundance values for all the proteins, including those in constant amounts, as is the case when protein abundances are used to feed metabolic models. Since the *outliers filter* implicitly makes it possible to select proteins showing abundance variations across UPS1 concentrations, we could have expected all yeast proteins to be removed. This was not the case, however, because the relative proportion of yeast proteins in the total protein pool actually decreased with increasing UPS1 concentration. This variation in the total abundance of yeast proteins was subtle and barely detectable until the highest concentration of UPS1 (Figure S3).

Altogether, these results show that the effects of the *occurrence* and *outliers filter*s on the amount of data depend greatly on the dynamics of protein abundance in the experiment. If these dynamics are large, the *occurrence filter* will not only remove dubious peptide ions associated with alignment or peak detection problems, but also many peptides with low ionization efficiencies that could be valuable for protein quantification. To further test the extent to which the severity of the *occurrence filter* can affect the performance of quantification, we also decided to use a restrained setup with a smaller UPS1 concentration range (0.5 to 27.9 fmol μl^{-1}), which was more representative of a natural dynamic range as the distribution of UPS1 intensities fitted that of yeast better (Figure S4). In this restrained setup, the UPS1 peptides with low ionization efficiencies had much fewer missing values, with the result that the *occurrence filter* affected the amount of data less severely (-12.2% and -7.3% of UPS1 proteins in the whole and restrained setup, respectively; Supplemental Table S1). In addition, 91.4% of the yeast peptide ions and 71.6% of the yeast proteins were removed by the *outliers filter*, confirming the efficiency of this filter for removing proteins showing no abundance variations.

2. Quantification methods do not respond equally to peptide filters

For each quantification method, the effects of peptide filters were evaluated in terms of precision, accuracy and linearity of response to increasing UPS1 concentrations. To determine to what extent these quality criteria can be affected by the severity of the *occurrence filter* (see above), we computed them for both the whole and restrained experimental setup. Precision, accuracy and linearity were evaluated on the UPS1 proteins detected in the five datasets (*i.e.* 35 and 37 UPS1 proteins in the whole and restrained experimental setups, respectively).

For each UPS1 protein, precision was computed as the median of the coefficients of variation (CV) determined between replicates of each UPS1 concentration. Results are presented in Figure 2 as boxplots showing the dispersion of CVs in each dataset. They show that none of the filters had a clear global effect on the precision of quantification for UPS1 proteins either in the whole experimental setup or in the restrained setup (Figure 2). Since the serial dilutions included only a few technical variations, we assumed that the number of UPS1 proteins was not high enough to observe a global effect of the filters on precision. Precision was slightly improved on yeast proteins by the *occurrence filter* when the *Average* or *Average-Log* method was employed, while the *outliers filter* decreased the precision with all methods except *TOP3* (Figure S5). Note that precision was similar regardless of the quantification method used (Figure 2).

Then, to estimate accuracy in the absence of a reference indicating the theoretical protein abundances expected at each UPS1 concentration, we used the equimolarity of the UPS1 proteins. If accuracy is high, the estimated abundances within the set of UPS1 proteins should present little dispersion. We therefore used the inverse of the CVs of protein abundances across UPS1 proteins as a proxy for accuracy, with protein abundances averaged across serial dilutions. Accuracy measurements thus obtained at each UPS1 concentration in each dataset are summarized as boxplots in Figure 3, showing that protein quantifications by *iBAQ* and *Average* were particularly improved by the *shared peptide filter*, which was not the case for the other quantification methods (Figure 3). This result, observed in the two experimental setups, is explained by the fact that both *iBAQ* and *Average* are based on untransformed intensities: in the computation of their sum or average, peptides of high intensity weigh more than peptides of low intensity. As their intensities correspond to the sum

of abundances of the proteins they belong to , shared peptides are globally more intense than proteotypic peptides. When taken into account, they can therefore lead to strongly overestimating protein abundances, especially when computed by *iBAQ* and *Average* (Figure 4A). These results indicate that in the case of these two quantification methods, it is important to filter not only shared peptides but also all types of dubious peptide ions of high intensity (see for example Figure 4B). By contrast, *Average-Log* and *Model* were only slightly improved by the *shared peptide filter*: both methods are based on log-transformed intensities, where the difference between peptide ions of high and low intensity is reduced. In addition, the *Model* discards the peptide ion effect, which results in a similar weight of all peptides in the computation of protein abundance.

Note that Figure 3 indicates that accuracy for *TOP3* was not as improved by the *shared peptide filter* as for *iBAQ* and *Average*. Nonetheless, Figures 4A and 4B, illustrating the effects of the *shared peptide filter* and the *RT filter* on peptide data and on estimated protein abundances for two proteins, show that as for *iBAQ* and *Average*, *TOP3* may be strongly biased by peptide ions of high intensity. Therefore, it is difficult to globally compare the effects of filters on quantification performances between *TOP3* and the other quantification methods because in the case of *TOP3*, the effects of the filters are highly dependent on the proteins used. If the peptide ions that are filtered are not among the three most intense ones, the filter will have no effect on *TOP3* (for instance, see Figure 4D). By contrast, if the peptide ions that are filtered are among the three most intense ones, the filter will necessarily have a large effect because the bias introduced by the irrelevant peptide ion before filtering is poorly buffered by the other two peptide ions. ~~*TOP3* is therefore an "all-or-nothing" method in the sense that depending on their ionization potential, irrelevant peptide ions can either have no effect or introduce a strong bias in protein quantification.~~

Regarding the *occurrence filter*, we observed that for *Average*, *Average-Log* and *Model*, it had contrasting effects on accuracy depending on the experimental setup (Figure 3). Accuracy was clearly improved in the restrained setup, especially for *Model* and *Average-Log*, while it was slightly degraded in the whole setup (Figure 3). This result was unexpected since in the whole setup, the *occurrence filter* allowed us to select peptide ions with high ionization efficiencies (Figure 4C). These peptides are indeed commonly admitted as being the most representative of the protein abundances (*e.g.*) based on the observation that the average intensity of the three most intense peptides per mole of protein was constant within a CV less than 10% . This observation has led to the development of *TOP3* for absolute

quantification . As previously mentioned, many peptide ions removed by the *occurrence filter* in the whole experimental setup were valuable peptide ions with low ionization efficiency but with nice linear responses to increasing UPS1 concentrations (Figure 4C). By contrast, the proportion of valuable peptide ions removed by the *occurrence filter* in the restrained setup was lower than in the whole setup. These results therefore indicate that decreasing the number of valuable peptide ions to compute protein abundance negatively affects the accuracy of *Average*, *Average-Log* and *Model*. This may seem contradictory with the principle of *TOP3*, but it can be easily explained since peptides have unequal ionization efficiencies. To reach high accuracy, proteins must be quantified with peptide ion sets representing, on average, equivalent ionization efficiencies. This is what *TOP3* does when selecting the three most intense peptide ions: it levels the average ionization efficiencies associated with the proteins upwards. In the case of *Average*, *Average-Log* and *Model*, the set of peptide ions used to compute a protein abundance can be viewed as a sampling of the diversity of the peptide ionization efficiencies. This sampling must be large enough to be representative. To confirm this hypothesis, we separated the UPS1 proteins into two groups depending on their number of quantified peptides, thus showing that accuracy was much higher for proteins quantified by many peptide ions, particularly in the case *Average-Log* and *Model* (Figure S6). Therefore, we conclude that by removing too many valuable peptide ions in the whole experimental setup, the *occurrence filter* affected the representativeness of the peptide ion sets associated with proteins, which consequently led to a lower accuracy for *Average*, *Average-Log* and *Model*.

Unlike *Average*, *Average-Log* and *Model*, the effect of the *occurrence filter* on the accuracy of *iBAQ* was the same in the two experimental setups and led to a loss of accuracy (Figure 3). This is because the number of peptide ions associated with a protein in *iBAQ* is *per se* an indication of abundance. To compute *iBAQ*, peptide data should ideally be filtered to remove peptide ions with missing values due to problems in RT alignment or in peak detection, but not peptide ions with missing values due to low ionization efficiency. However, in real experiments, the proportion of these two types of peptide ions is not known. We therefore recommend not applying the *occurrence filter* in the case of *iBAQ* if high accuracy is the objective.

Linearity was evaluated by using the coefficients of determination (R^2) of linear regressions calculated between the log-transformed abundances obtained experimentally for UPS1 proteins and their spiked log-transformed concentrations. Abundance and concentrations were log-transformed for the sake of clarity. The R^2 values obtained for each

UPS1 protein in each dataset are summarized as boxplots in Figure 5. Filters improved the linearity by removing peptide ions displaying non-linear responses to increasing UPS1 concentrations (Figure 4). In the case of *iBAQ* and *Model*, a good linearity was obtained without using any filter in both experimental setups. By contrast, *TOP3* linearity was clearly improved by the *RT filter*. The effect of the *occurrence filter* was globally the same in the two experimental setups in which it greatly improved linearity for *Average* and *Average-Log* (Figure 5) Therefore, using the two latter methods, linearity was strongly affected by missing data because it led to high between-sample variability. Of note, linearity of *Average* was less affected than *Average-Log* by the *occurrence filter* because peptides with low ionization efficiency had less weight in the former (Figure 5).

Interestingly, when no filter was used *iBAQ* and *Model*, the slope of the regression between the log-transformed abundances obtained experimentally for UPS1 proteins and their spiked log-transformed concentrations was close to their optimal value and the theoretically expected value of 1. For *Average* and *Average-Log*, slopes similar to those of *iBAQ* and *Model* (close to expected value of 1) were obtained with the *occurrence filter* (Figure S7). This indicates, that filters not only improved the linearity of the response but also made it possible to obtain abundance-concentration relationships close to that theoretically expected. This was especially the case for *Average* and *Average-Log*.

3. The performances of one quantification method over another depend on how the data were filtered

To summarize the absolute and relative quantification performances of the quantification methods tested in this study, we plotted accuracy versus linearity obtained for each method in the two experimental setups. When the objective is absolute quantification, high accuracy is essential for reliably estimating intracellular protein concentrations. However, if the objective is relative quantification, accuracy can be neglected as long as the errors between the observed and theoretical values are similar in all samples. If this is not the case, the linearity of the response to increasing UPS1 concentrations would be affected.

Figure 6 clearly shows that the absolute and relative quantification performances of the methods depend on the quality of the dataset, and that filtering made it possible to reduce their differences in performance for all the experimental setups. Interestingly, *Model* gave the best performance in terms of linearity and accuracy in the two

experimental setups in absence of any filter, indicating that it is the most robust method. This result is in agreement with a previous study showing that statistical modeling of protein abundances is the most adequate method to infer and quantitatively compare protein abundances . Figure 6 also shows that the filtering procedure should be chosen according to the quantification objective, since filters increasing performance in relative quantification may degrade performance in absolute quantification. For example, with *Average*, the *occurrence* and *outlier filters* improved linearity at the expense of accuracy in the whole experimental setup.

CONCLUSION

Owing to their different properties related to the computation modes used to estimate protein abundances, quantification methods do not respond similarly to peptide filters. Therefore, filters should be chosen carefully according to a) the quantification method, b) the quantification objective (absolute or relative), and c) the experimental design. We make the following recommendations: data should be filtered to remove shared peptides, especially when using *iBAQ* or *Average* because they are susceptible to high intensity peptide ions. First, missing data should be handled carefully when using *Average* and *Average-Log* because they are a potential source of between-sample variability that affects relative quantification. Second, the *occurrence filter* can be used to manage missing data but it is to be used with caution: depending on the experimental design, it may remove many valuable peptide ions that present qualitative variations due to the large dynamics of protein expression. In addition, if the filter is too stringent, it may degrade accuracy in the case of *Average*, *Average-Log* and *Model*. Carefully combining the *occurrence filter* with missing data imputation would probably be a good alternative. In the case of *iBAQ*, the *occurrence filter* degraded accuracy, so if absolute quantification is the objective, we recommend not applying it when using *iBAQ*, even if it means keeping some dubious peptides. For the same reason, the *outliers filter* should be used with caution. However, these two filters improved *iBAQ* linearity, so they are relevant if relative quantification is the objective. Finally, we confirmed our hypothesis that by appropriately using peptide filters, good performances could be reached in both relative and absolute quantification, regardless of the quantification method. *Model* proved to be the most efficient method and may be used for absolute quantification when proteins are quantified by a sufficient number of peptides.

ACKNOWLEDGEMENTS

We acknowledge funding from ANR (ANR-15-CE20-0009-01 FRIMOUSS). We also acknowledge the Saclay Plant Science Labex and thank Dr Ray Cooke for copyediting the manuscript.

FIGURE LEGENDS

Figure 1 Schema of peptide ion filtering workflow. Dataset 1 derived from normalization of raw dataset (Dataset 0), Dataset 2 derived from normalized Dataset 0 without shared peptides (*shared peptide filter*). To produce Dataset 3, peptides with a standard deviation of retention time higher than 30 seconds were removed (*RT filter*) before normalizing and filtering shared peptides. To produce Dataset 4, peptide ions presenting more than 15.15% of missing values were filtered out from Dataset 3 (*occurrence filter*). To produce Dataset 5, uncorrelated peptide ions belonging to same protein (Pearson, $R^2 > 0.8$, p-value < 0.01) were filtered out (*outliers filter*).

Figure 2 Effect of peptide filters on precision of UPS1 protein abundance estimation by five methods of quantification (*iBAQ*, *TOP3*, *Average*, *Average-Log* and *Model*). For each UPS1 protein, precision was calculated as median CV (%) of protein abundance between three technical replicates determined at each UPS1 protein concentration. Only UPS1 proteins detected in the five datasets were used (*i.e.* 35 UPS1 proteins in whole experimental setup (red boxplots) and 37 in restrained setup (blue boxplots)). Only medians were plotted to compare all methods (bottom right).

Figure 3 Effect of peptide filters on accuracy of UPS1 protein abundance estimation by five methods of quantification (*iBAQ*, *TOP3*, *Average*, *Average-Log* and *Model*). Accuracy was computed at each UPS1 concentration as the inverse of the coefficient of variation (CV) between the average abundances of UPS1 proteins (n = 3 replicates). Only UPS1 proteins detected in the five datasets were used (*i.e.* 35 UPS1 proteins in whole experimental setup (red boxplots) and 37 in restrained setup (blue boxplots)). Only medians were plotted to compare all methods (bottom right).

Figure 4 Effect of filters on peptide ion selection (left panel) and estimation of protein abundance (right panel) illustrated on four UPS1 proteins in whole experimental setup. Effect of *shared peptide filter* on P62988 protein (A), effect of *RT filter* on P63165 protein (B), effect of *occurrence filter* on P02144 protein (C) and effect of *outliers filter* on P02787 protein (D). Estimated protein abundances were averaged across technical replicates (n=3). In Figure 4D, protein abundance estimated before and after *outliers filter* by *TOP3* are confused. Protein abundances estimated after *outliers filter* by *Model* and *Average*d-Log are superimposed.

Figure 5 Effect of peptide filters on linearity between spiked UPS1 proteins concentrations and their abundances based on the five methods of quantification (*iBAQ*, *TOP3*, *Average*, *Average-Log* and *Model*). Linearity was evaluated by the coefficients of determination (R^2) of linear regressions between the log-transformed abundances obtained experimentally for UPS1 proteins and their spiked log-transformed concentrations. Protein abundances obtained experimentally were averaged across replicates (n=3). Only UPS1 proteins detected in the five datasets were used (*i.e.* 35 UPS1 proteins in whole experimental setup (red boxplots) and 37 in restrained setup (blue boxplots)). Only medians were plotted to compare all methods (bottom right).

Figure 6 Overall effect of peptide filters on performances (accuracy versus linearity) of five methods of quantification in whole experimental (A) and restrained (B) setup. For each quantification method, the third quartile (75% of UPS1 proteins) was used to sum up accuracy and linearity values displayed in Figure 3 and 5.

Figure S1 A four-set Venn diagram showing number of peptide ions removed by each filter applied separately for UPS1 (A) and yeast (B) proteins.

Figure S2 Relationship between number of quantified peptide ions and sequence length (in amino acids) for each UPS1 protein in the different datasets.

Figure S3 Distribution of log-transformed intensities of yeast peptide ions at each UPS1 concentration in dataset 0 (raw data). Number of yeast proteins quantified at each UPS1 concentration is shown above boxplots.

Figure S4 Distribution of log-transformed intensities of yeast (blue) and UPS1 (red) peptide ions in dataset 1 in whole experimental (A) and restrained setup (B). 2039 and 2033 yeast proteins were detected in whole experimental and restrained setup, respectively, and 41 UPS proteins were detected in both experimental setups (Table 1, Table S1).

Figure S5 Effect of four filters on precision of yeast protein quantification based on *iBAQ*, *TOP3*, *Average*, *Average-Log* and *Model* methods. For each yeast protein and at each UPS1 protein concentration, CV (%) of protein abundance between replicates (n= 3) was determined. Then, precision for each protein was calculated as median across serial dilutions of CVs (%). Only yeast proteins detected in the five datasets were used (*i.e.* 973 yeast proteins in whole experimental setup (red boxplots) and 518 in restrained setup (blue boxplots)).

Figure S6 Accuracy of UPS1 protein abundance according to number of peptides. For the whole experimental setup and the five methods of quantification, 35 UPS1 proteins (detected in the five datasets) were split into two groups -lower (orange boxplots) and higher (green boxplots)- according to median of peptide number determined in each dataset (Dataset 1 and 2: 18 peptides, Dataset 3: 17 peptides, Dataset 4 and 5: 6 peptides). Accuracy was computed at each UPS1 concentration as inverse of coefficient of variation (CV) between average abundances of UPS1 proteins (n = 3 replicates).

Figure S7 Effect of four filters on slope of linear regression calculated between spiked UPS1 protein concentrations and their abundances based on the five methods of quantification (*iBAQ*, *TOP3*, *Average*, *Average-Log*, *Model*). Linear regressions were performed between log10-transformed concentrations and averaged protein abundances (n= 3 replicates) log10-transformed. Only UPS1 proteins detected in the five datasets were used (*i.e.* 35 UPS1 proteins in whole experimental setup (red boxplots) and 37 in the restrained setup (blue boxplots)).

609 REFERENCES

610

611 [2] D. Chelius, P. V Bondarenko, Quantitative profiling of proteins in complex mixtures
612 using liquid chromatography and mass spectrometry., J. Proteome Res. 1 (2002) 317–
613 23. http://www.ncbi.nlm.nih.gov/pubmed/12645887 (accessed October 19, 2017).

614 [3] D.S. Daly, K.K. Anderson, E.A. Panisko, S.O. Purvine, R. Fang, M.E. Monroe, S.E.
615 Baker, Mixed-Effects Statistical Model for Comparative LC−MS Proteomics Studies, J.
616 Proteome Res. 7 (2008) 1209–1217. doi:10.1021/pr070441i.

617 [4] M. Blein-Nicolas, H. Xu, D. de Vienne, C. Giraud, S. Huet, M. Zivy, Including shared
618 peptides for estimating protein abundances: A significant improvement for quantitative
619 proteomics, Proteomics. 12 (2012) 2797–2801. doi:10.1002/pmic.201100660.

620 [5] S. Gerster, T. Kwon, C. Ludwig, M. Matondo, C. Vogel, E.M. Marcotte, R. Aebersold,
621 P. Bühlmann, Statistical approach to protein quantification., Mol. Cell. Proteomics. 13
622 (2014) 666–77. doi:10.1074/mcp.M112.025445.

623 [6] L. Jacob, F. Combes, T. Burger, PEPA test: fast and powerful differential analysis from
624 relative quantitative proteomics data using shared peptides, Biostatistics. (2018).
625 doi:10.1093/biostatistics/kxy021.

626 [7] Y. Karpievitch, J. Stanley, T. Taverner, J. Huang, J.N. Adkins, C. Ansong, F. Heffron,
627 T.O. Metz, W. Qian, H. Yoon, R.D. Smith, A.R. Dabney, A statistical framework for
628 protein quantitation in bottom-up MS-based proteomics, 25 (2009) 2028–2034.
629 doi:10.1093/bioinformatics/btp362.

630 [8] K. Richardson, R. Denny, C. Hughes, J. Skilling, J. Sikora, M. Dadlez, A. Manteca,
631 H.R. Jung, O.N. Jensen, V. Redeker, R. Melki, J.I. Langridge, J.P.C. Vissers, A
632 Probabilistic Framework for Peptide and Protein Quantification from Data-Dependent
633 and Data-Independent LC-MS Proteomics Experiments, Omi. A J. Integr. Biol. 16
634 (2012) 468–482. doi:10.1089/omi.2012.0019.

635 [9] M. Blein-Nicolas, W. Albertin, T. Da Silva A, B. Valot, T. Balliau, I. Masneuf-Pomarè
636 De C, M. Bely, P. Marullo, D. Sicard, C. Dillmann, D. De Vienne, M. Zivy, A Systems
637 Approach to Elucidate Heterosis of Protein Abundances in Yeast, Mol. Cell.
638 Proteomics. 14 (2015) 2056–71. doi:10.1074/mcp.M115.048058.

[10] X. Lai, L. Wang, H. Tang, F.A. Witzmann, A Novel Alignment Method and Multiple Filters for Exclusion of Unqualified Peptides To Enhance Label-Free Quantification Using Peptide Intensity in LC—MS/MS, J. Proteome Res. 10 (2011) 759–785. doi:10.1146/annurev-cellbio-092910-154240.Sensory.

[11] B.J.M. Webb-Robertson, L.A. McCue, K.M. Waters, M.M. Matzke, J.M. Jacobs, T.O. Metz, S.M. Varnum, J.G. Pounds, Combined statistical analyses of peptide intensities and peptide occurrences improves identification of significant peptides from MS-based proteomics data, J. Proteome Res. 9 (2010) 5748–5756. doi:10.1021/pr1005247.

[12] A.D. Polpitiya, W.-J. Qian, N. Jaitly, V.A. Petyuk, J.N. Adkins, D.G. Camp, G.A. Anderson, R.D. Smith, DAnTE: a statistical tool for quantitative analysis of -omics data, Bioinformatics. 24 (2008) 1556–1558. doi:10.1093/bioinformatics/btn217.

[13] J. Forshed, H.J. Johansson, M. Pernemalm, R.M.M. Branca, A. Sandberg, J. Lehtiö, Enhanced Information Output From Shotgun Proteomics Data by Protein Quantification and Peptide Quality Control (PQPQ), Mol. Cell. Proteomics. 10 (2011) M111.010264. doi:10.1074/mcp.M111.010264.

[14] B. Zhang, M. Pirmoradian, R. Zubarev, L. Käll, Covariation of Peptide Abundances Accurately Reflects Protein Concentration Differences, Mol. Cell. Proteomics. (2017) 1–42.

[15] M. Blein-Nicolas, M. Zivy, Thousand and one ways to quantify and compare protein abundances in label-free bottom-up proteomics, Biochim. Biophys. Acta - Proteins Proteomics. 1864 (2016) 883–895. doi:10.1016/j.bbapap.2016.02.019.

[16] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, M. Selbach, Global quantification of mammalian gene expression control, Nature. 473 (2011) 337–342. doi:10.1038/nature10098.

[17] J.C. Silva, M. V. Gorenstein, G.-Z. Li, J.P.C. Vissers, S.J. Geromanos, Absolute Quantification of Proteins by LCMS ^{E}, Mol. Cell. Proteomics. 5 (2006) 144–156. doi:10.1074/mcp.M500230-MCP200.

[18] R.E. Higgs, M.D. Knierman, V. Gelfanova, J.P. Butler, J.E. Hale, Comprehensive label-free method for the relative quantification of proteins from biological samples, J. Proteome Res. 4 (2005) 1442–1450. doi:10.1021/pr050109b.

669 [19] T. Clough, M. Key, I. Ott, S. Ragg, G. Schadow, O. Vitek, Protein Quantification in
670 Label-Free LC-MS Experiments, J. Proteome Res. 8 (2009) 5275–5284.
671 doi:10.1021/pr900610q.

672 [20] J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized
673 p.p.b.-range mass accuracies and proteome-wide protein quantification, Nat.
674 Biotechnol. (2008). doi:10.1038/nbt.1511.

675 [21] J. Cox, M.Y. Hein, C.A. Luber, I. Paron, N. Nagaraj, M. Mann, Accurate Proteome-
676 wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio
677 Extraction, Termed MaxLFQ, Mol. Cell. Proteomics. (2014).
678 doi:10.1074/mcp.M113.031591.

679 [22] D. Kessner, M. Chambers, R. Burke, D. Agus, P. Mallick, ProteoWizard: open source
680 software for rapid proteomics tools development, Bioinformatics. 24 (2008) 2534–
681 2536. doi:10.1093/bioinformatics/btn323.

682 [23] H. Ferry-Dumazet, G. Houel, P. Montalent, L. Moreau, O. Langella, L. Negroni, D.
683 Vincent, C. Lalanne, A. de Daruvar, C. Plomion, M. Zivy, J. Joets, PROTICdb: A web-
684 based application to store, track, query, and compare plant proteome data, Proteomics.
685 5 (2005) 2069–2081. doi:10.1002/pmic.200401111.

686 [24] O. Langella, M. Zivy, J. Joets, The PROTICdb Database for 2-DE Proteomics, in: Plant
687 Proteomics, Humana Press, New Jersey, 2007: pp. 279–304. doi:10.1385/1-59745-227-
688 0:279.

689 [25] O. Langella, B. Valot, D. Jacob, T. Balliau, R. Flores, C. Hoogland, J. Joets, M. Zivy,
690 Management and dissemination of MS proteomic data with PROTICdb: Example of a
691 quantitative comparison between methods of protein extraction, Proteomics. 13 (2013)
692 1457–1466. doi:10.1002/pmic.201200564.

693 [26] J.A. Vizcaíno, A. Csordas, N. Del-Toro, J.A. Dianes, J. Griss, I. Lavidas, G. Mayer, Y.
694 Perez-Riverol, F. Reisinger, T. Ternent, Q.W. Xu, R. Wang, H. Hermjakob, 2016 update
695 of the PRIDE database and its related tools, Nucleic Acids Res. (2016).
696 doi:10.1093/nar/gkv1145.

697 [27] R. Craig, R.C. Beavis, TANDEM: matching proteins with tandem mass spectra,
698 Bioinformatics. 20 (2004) 1466–1467. doi:10.1093/bioinformatics/bth092.

699 [28] O. Langella, B. Valot, T. Balliau, M. Blein-Nicolas, L. Bonhomme, M. Zivy,
700 X!TandemPipeline: A Tool to Manage Sequence Redundancy for Protein Inference and
701 Phosphosite Identification, J. Proteome Res. 16 (2017) 494–503.
702 doi:10.1021/acs.jproteome.6b00632.

703 [29] B. Valot, O. Langella, E. Nano, M. Zivy, MassChroQ: A versatile tool for mass
704 spectrometry quantification, Proteomics. 11 (2011) 3572–3577.
705 doi:10.1002/pmic.201100120.

706 [30] Y. Lyutvinskiy, H. Yang, D. Rutishauser, R.A. Zubarev, *In Silico* Instrumental Response
707 Correction Improves Precision of Label-free Proteomics and Accuracy of Proteomics-
708 based Predictive Models, Mol. Cell. Proteomics. 12 (2013) 2324–2331.
709 doi:10.1074/mcp.O112.023804.

710 [31] A. Millan-Oropeza, C. Henry, M. Blein-Nicolas, A. Aubert-Frambourg, F. Moussa, J.
711 Bleton, M.-J. Virolle, Quantitative Proteomics Analysis Confirmed Oxidative
712 Metabolism Predominates in *Streptomyces coelicolor* versus Glycolytic Metabolism in
713 *Streptomyces lividans*, J. Proteome Res. 16 (2017) 2597–2613.
714 doi:10.1021/acs.jproteome.7b00163.

715 [32] RStudio Team, RStudio: Integrated Development for R. RStudio, Inc., Boston, MA
716 URL, (2015). http://www.rstudio.com/.

717 [33] M. Kellis, B.W. Birren, E.S. Lander, Proof and evolutionary analysis of ancient
718 genome duplication in the yeast Saccharomyces cerevisiae, Nature. 428 (2004) 617–
719 624. doi:10.1038/nature02424.

720 [34] Y. V Bukhman, M. Dharsee, R.O.B. Ewing, P. Chu, T. Topaloglou, T.L.E. Bihan, T.
721 Goh, H. Duewel, I.A.N.I. Stewart, J.R. Wisniewski, N.F. Ng, Design and analysis of
722 quantitative differential proteomics investigations using LC-MS technologies, J.
723 Bioinform. Comput. Biol. 6 (2008) 107–123.

724 [35] J.D. Worboys, J. Sinclair, Y. Yuan, C. Jørgensen, Systematic evaluation of quantotypic
725 peptides for targeted analysis of the human kinome, Nat. Methods. 11 (2014) 1041–
726 1044. doi:10.1038/nmeth.3072.

727

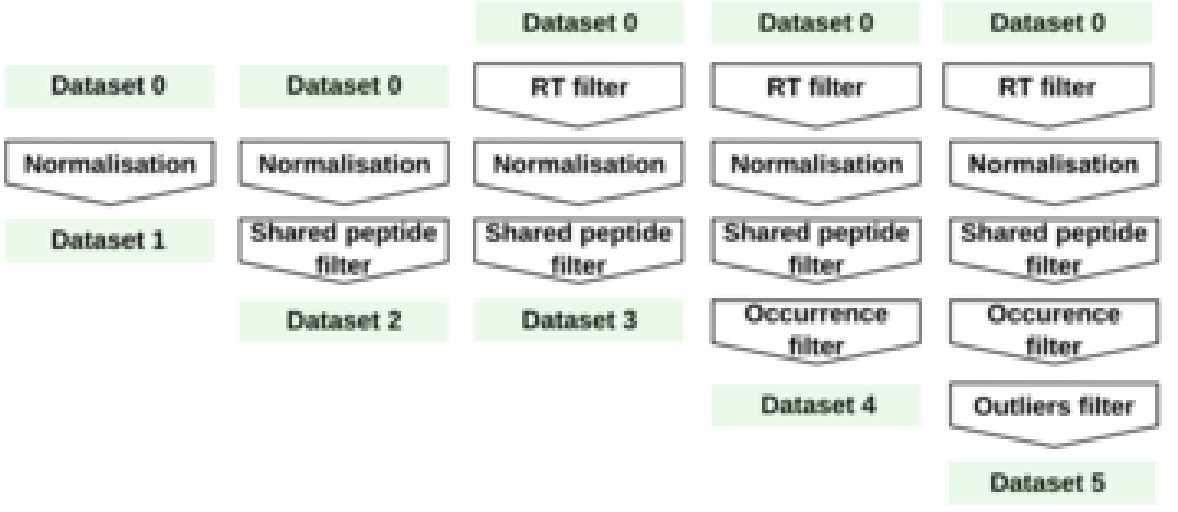
728 **Table 1** Effect of filters on number of peptide ions and proteins. Numbers in parenthesis
729 indicate percentage of data removed by filter from previous dataset.

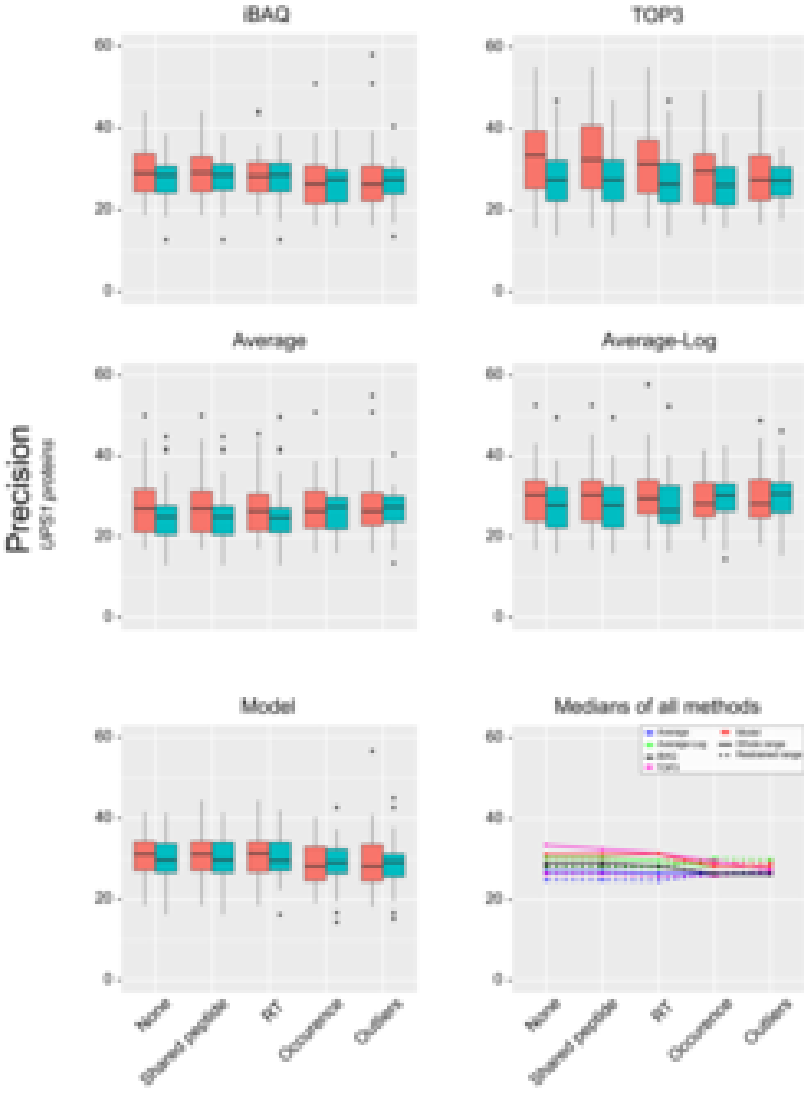
		No filter	*Shared peptide filter*	*RT filter*	*Occurrence filter*	*Outliers filter*
Peptide ions	Total	22 950	22 044 (-3.9%)	21 857 (-0.8%)	13 561 (-38.0%)	4 882 (-64.0%)
	Yeast	21 820	20 915 (-4.2%)	20 778 (-0.7%)	13 314 (-35.9%)	4 666 (-65.0%)
	UPS1	1 138	1 129 (-0.8%)	1 079 (-4.4%)	247 (-77.1%)	216 (-12.6%)
Proteins	Total	2 080	2 046 (-1.6%)	2 041 (-0.2%)	1 491 (-26.9%)	1 008 (-32.4%)
	Yeast	2 039	2 005 (-1.7%)	2 000 (-0.3%)	1 455 (-21.3 %)	973 (-33.1%)
	UPS1	41	41 (-0%)	41 (-0%)	36 (-12.2%)	35 (-2.8%)

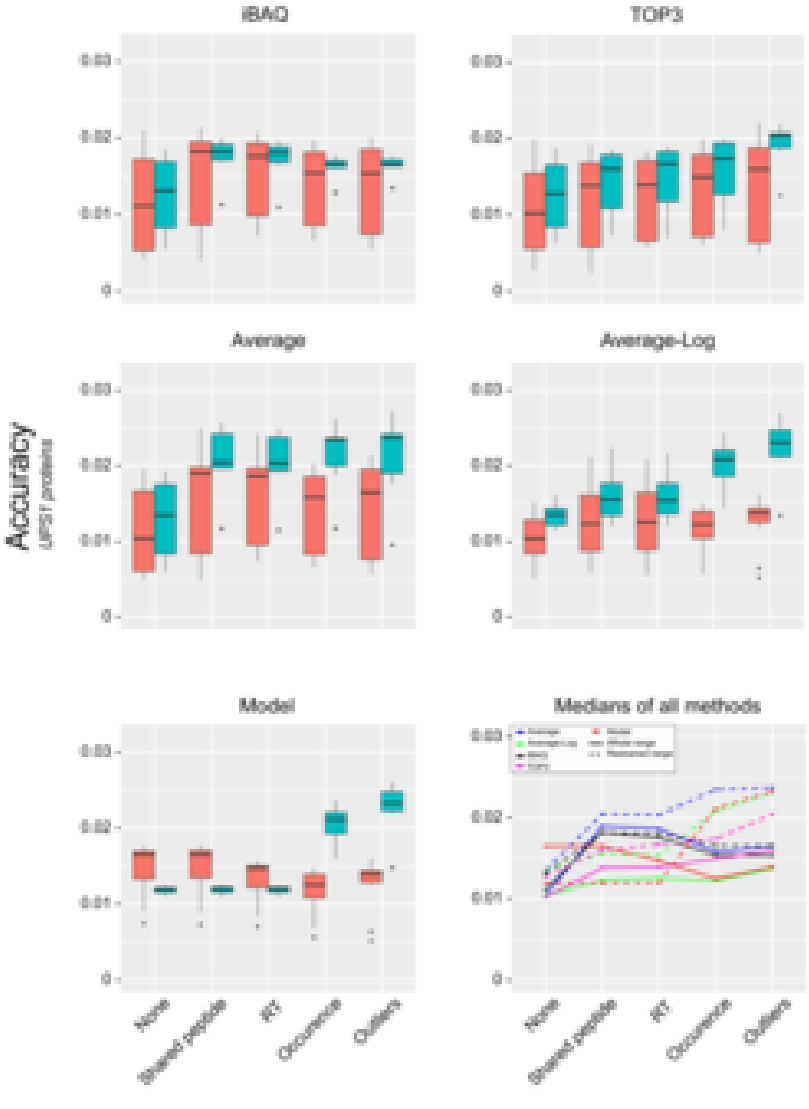
730

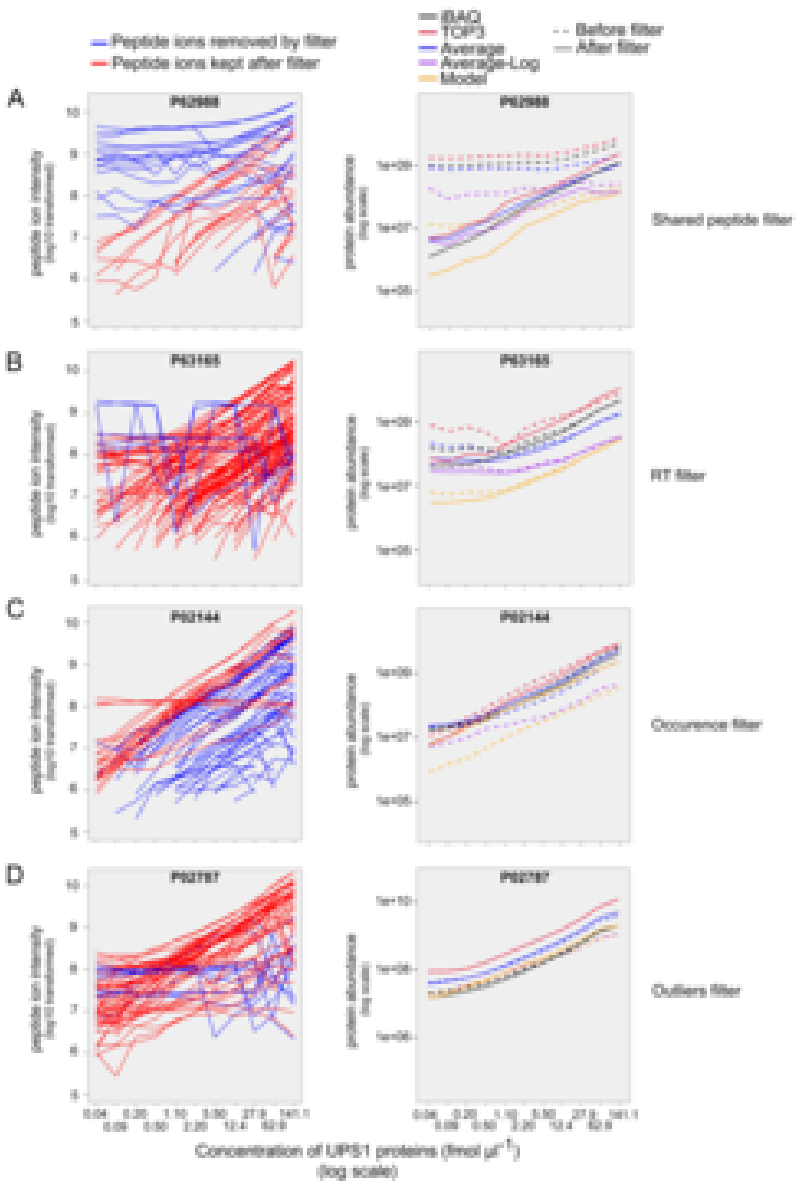
731

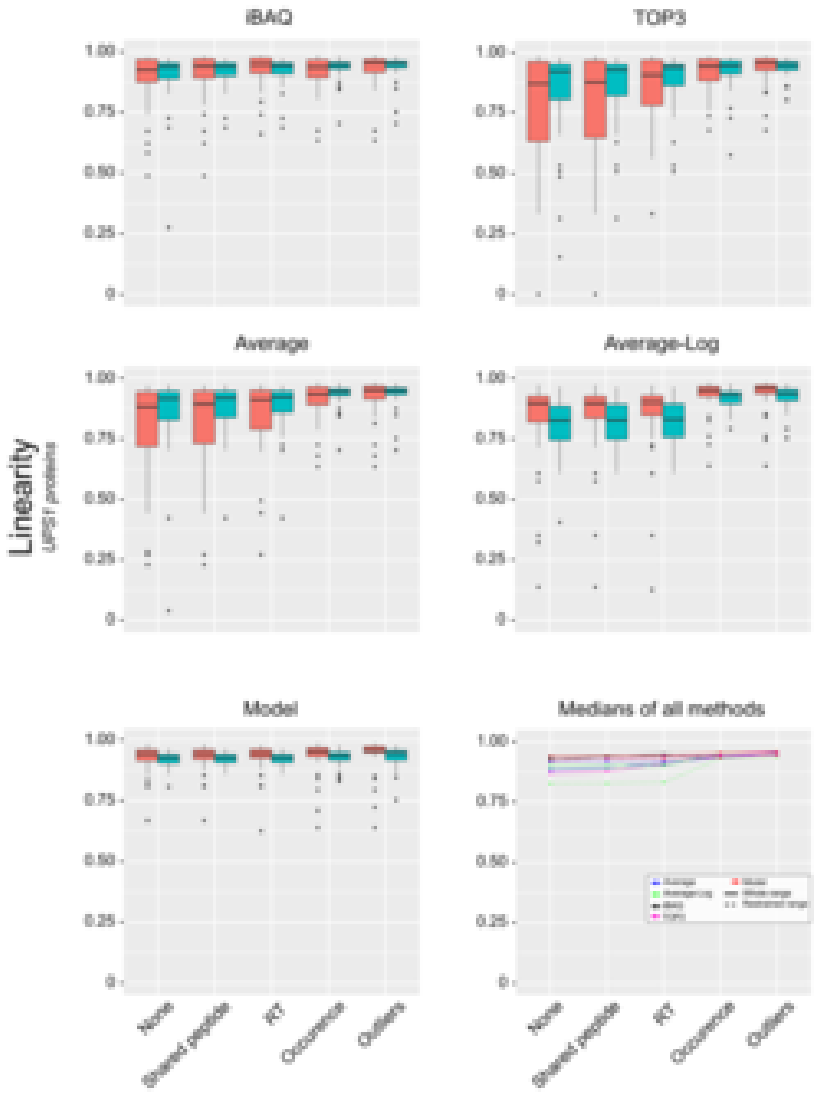
732

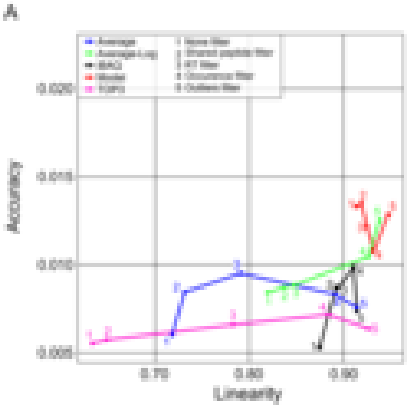

Dataset 0
Dataset 0
Dataset 0
Dataset 0
RT filter
RT filter
RT filter
Normalisation
Normalisation
Normalisation
Normalisation
Normalisation
Dataset 1
Shared peptide filter
Shared peptide filter
Shared peptide filter
Shared peptide filter
Dataset 2
Dataset 3
Occurrence filter
Occurence filter
Dataset 4
Outliers filter
Dataset 5


Precision
UPS1 proteins
iBAQ
TOP3
Average
Average-Log
Model
Medians of all methods
60
40
20
0
None
Shared peptide
IDT
Occurence
Outliers


iBAQ
TOP3
Average
Average-Log
Model
Medians of all methods
Accuracy
UPS1 proteins
0.03
0.02
0.01
0
None
Ignore peptide
IF1
Occurrence
Outliers


Peptide ions removed by filter
Peptide ions kept after filter
iBAQ
TOP3
Average
Average-Log
Model
Before filter
After filter
A
B
C
D
P62988
P63165
P62144
P62787
peptide ion intensity (log10 transformed)
protein abundance (log scale)
Shared peptide filter
RT filter
Occurence filter
Outliers filter
0.04 0.20 1.10 5.50 27.5 141.1
0.09 0.50 2.20 12.4 62.9
Concentration of UPS1 proteins (fmol µl⁻¹)
(log scale)


IBAQ
TOP3
Average
Average-Log
Model
Medians of all methods
Linearity
UPS1 proteins
1.00
0.75
0.50
0.25
0
None
Shared peptide
IDI
Occurence
Outliers


A
Average
Average-Log
BAG
Model
Cubic
1 None filter
2 Shared peptide filter
3 RT filter
4 Occurence filter
5 Outlaws filter
Accuracy
0.020
0.015
0.010
0.005
0.70
0.80
0.90
Linearity

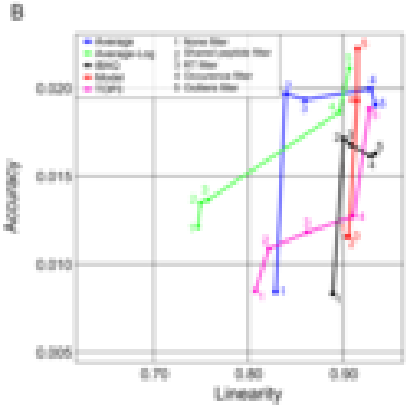
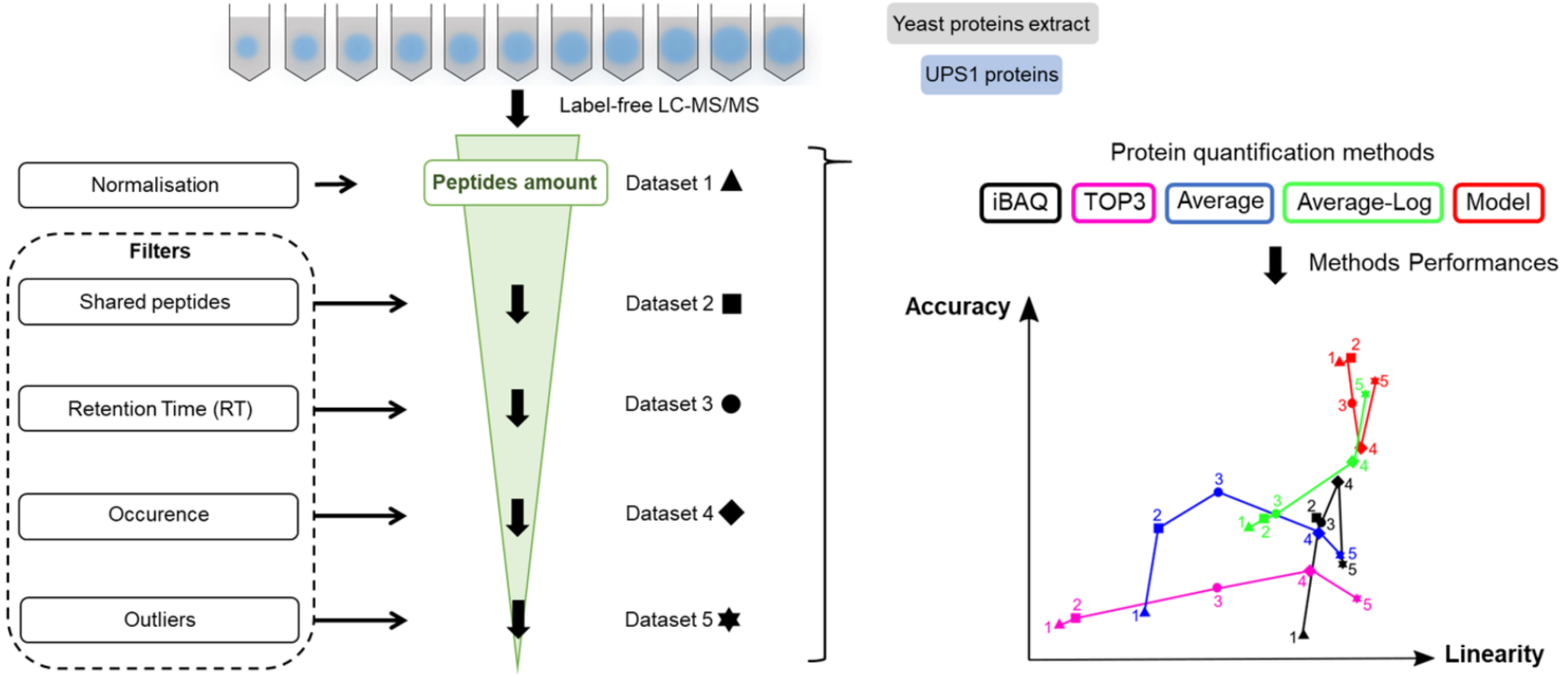

B
Average
Average-Log
BAG
Model
Cubic
1 None filter
2 Shared peptide filter
3 RT filter
4 Occurence filter
5 Outlaws filter
Accuracy
0.020
0.015
0.010
0.005
0.70
0.80
0.90
Linearity

Table 1 Effect of filters on the number of peptides ions and proteins. Numbers in parenthesis indicate the percentage of data removed by the filter from the previous dataset.

		No filter	*Shared peptide filter*	*RT filter*	*Occurrence filter*	*Outliers filter*
Peptide ions	Total	22 950	22 044 (-3.9%)	21 857 (-0.8%)	13 561 (-38.0%)	4 882 (-64.0%)
	Yeast	21 820	20 915 (-4.2%)	20 778 (-0.7%)	13 314 (-35.9%)	4 666 (-65.0%)
	UPS1	1 138	1 129 (-0.8%)	1 079 (-4.4%)	247 (-77.1%)	216 (-12.6%)
Proteins	Total	2 080	2 046 (-1.6%)	2 041 (-0.2%)	1 491 (-26.9%)	1 008 (-32.4%)
	Yeast	2 039	2 005 (-1.7%)	2 000 (-0.3%)	1 455 (-21.3 %)	973 (-33.1%)
	UPS1	41	41 (-0%)	41 (-0%)	36 (-12.2%)	35 (-2.8%)


Yeast proteins extract
UPS1 proteins
Label-free LC-MS/MS
Normalisation
Peptides amount
Dataset 1
Filters
Shared peptides
Dataset 2
Retention Time (RT)
Dataset 3
Occurence
Dataset 4
Outliers
Dataset 5
Protein quantification methods
iBAQ
TOP3
Average
Average-Log
Model
Methods Performances
Accuracy
Linearity