



HAL
open science

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

► **To cite this version:**

Isma Belouah, Melisande Blein-Nicolas, Thierry Balliau, Yves Gibon, Michel Zivy, et al.. Peptide filtering differently affects the performances of XIC-based quantification methods. *Journal of Proteomics*, 2019, 193, pp.131-141. 10.1016/j.jprot.2018.10.003 . hal-02327322

HAL Id: hal-02327322

<https://hal.science/hal-02327322v1>

Submitted on 19 Oct 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.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

17 **ABSTRACT**

18

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

34

35

36

37

38

39

40 INTRODUCTION

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

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

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

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

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

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

114

115 **MATERIAL AND METHODS**

116

117 **Yeast growth**

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

125

126 **Yeast protein extraction**

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

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

143

144 **Spike-in UPS1 preparation**

145 Dried UPS1 proteins (Sigma-Aldrich) were solubilized in the buffer containing
146 yeast proteins to a final concentration of 0.75 $\mu\text{g } \mu\text{l}^{-1}$ (0.625 fmol μl^{-1} of each UPS1 protein)
147 so that the total protein (yeast + UPS) concentration was 1.637 $\mu\text{g } \mu\text{l}^{-1}$. Proteins were
148 incubated for one hour at room temperature for reduction by the 10 mM DTT present in the
149 buffer. Thereafter, proteins were alkylated for one hour in the presence of 50 mM
150 iodoacetamide and diluted with 50 mM ammonium bicarbonate to decrease the total urea and
151 thiourea concentration to 3.6 M before being twice digested. A first 4-hour digestion was
152 performed with 1/32 (w/w) rLysC protease (Promega). After dilution with a solution of 50
153 mM ammonium bicarbonate to decrease the total urea and thiourea concentration to 0.77 M, a
154 second overnight digestion was performed with 1/32 (w/w) trypsin (Promega). Both rLysC
155 and trypsin digestion were performed at 37 °C. Trypsin digestion was stopped by acidification
156 (1% total volume trifluoroacetic acid). The resulting peptides were purified on solid-phase
157 extraction using a polymeric C18 column (Phenomenex) with a washing solution containing
158 0.06% acetic acid and 3% acetonitrile (ACN). After elution with 0.06% acetic acid and 40%
159 ACN, peptides were speedvac-dried and suspended in a solution containing 2% ACN, 0.06%
160 trifluoroacetic acid and 0.06% formic acid so that the concentration of each UPS1 peptide was
161 141.1 fmol μl^{-1} and the total concentration of yeast peptides was 200 ng μl^{-1} . A serial 2.25-
162 fold dilution was prepared by mixing 6.7 μl of UPS1-yeast peptide mix with 8.3 μl of
163 solubilized yeast peptides at 200 ng μl^{-1} until a UPS1 peptide concentration of 0.04 fmol μl^{-1}
164 was reached. Eleven samples were thus obtained, containing 141.1, 62.8, 27.9, 12.4, 5.5, 2.2,
165 1.1, 0.5, 0.2, 0.09 and 0.04 fmol μl^{-1} of each UPS1 peptide. This serial dilution was performed
166 in three replicates from aliquots of the same yeast culture, thus producing 33 samples.

167

168 **LS-MS/MS analyses**

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

180 Peptide ions were analyzed using Xcalibur 2.1 (Thermo Electron) with the
181 following data-dependent acquisition steps: (1) MS scan (mass-to-charge ratio (m/z) 300 to 1
182 400, 70 000 resolution, profile mode), (2) MS/MS (17 500 resolution, normalized collision
183 energy of 30, profile mode). Step 2 was repeated for the eight major ions detected in step (1).
184 Dynamic exclusion was set to 30 seconds. Xcalibur raw datafiles were transformed to
185 mzXML open source format using msconvert software in the ProteoWizard 3.0.3706 package
186 . During conversion, MS and MS/MS data were centroided. The raw MS output files and
187 protein abundances were deposited on-line using PROTeICdb database at the following URL:
188 http://moulon.inra.fr/protic/filtering_quanti_methods. They are currently available with the
189 following username: filtering and password: review. The mass spectrometry proteomics data
190 have also been deposited with the ProteomeXchange Consortium via the PRIDE partner
191 repository with the dataset identifier PXD009740. They are currently available with the
192 following username: reviewer32109@ebi.ac.uk and password: JH5JcHXE. They will be made
193 freely available after publication.

194

195 **Protein identification**

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

201 containing the sequences of standard contaminants was also interrogated. The decoy database
202 comprised the reverse sequences of yeast and UPS1 proteins. Database search was performed
203 with X!Tandem (version 2015.04.01.1;) using the following settings.
204 Carboxyamidomethylation of cysteine residues was set to static modification. Oxidation of
205 methionine residues, N-terminal acetylation with or without excision of the N-terminal
206 methionine, deamination of N-terminal glutamine and of carbamidomethylated cysteines and
207 loss of H₂O from N-terminal asparagins were set to possible modifications. In the refine
208 mode, excision of signal peptides was searched on the 50 first N-terminal amino acids
209 possibly acetylated. Precursor mass precision was set to 10 ppm. Fragment mass tolerance
210 was 0.02 Thomson (Th, unit of m/z). Only peptides with an E-value smaller than 0.05 were
211 reported.

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

217

218 **Peptide ion quantification and intensity data filtering**

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

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

233 were removed (*shared peptide filter*). The third dataset (Dataset 3) comprised two filters.
234 Peptides with a standard deviation of retention time higher than 30 seconds were first
235 removed (*RT filter*). Since these peptides were considered as dubious, this filter was applied
236 before normalization of yeast peptide intensities. Then, shared peptides were removed. The
237 fourth dataset (Dataset 4) comprised three filters. It was obtained by applying an *occurrence*
238 *filter* to Dataset 3, which resulted in the selection of peptide ions quantified in at least 28
239 samples, with no more than one missing value per UPS1 concentration. Thus, a maximum of
240 15.15% of missing values per peptide ion was tolerated and the selected peptide ions were
241 quantified in at least two replicates for each UPS1 concentration. To ensure the quality of
242 normalization, which depends on the number of peptide ions quantified both in a sample
243 chosen as reference and in a sample to be normalized, we decided to apply this filter after
244 normalization. Several peptide ions removed by the *occurrence filter* are good quality
245 peptides whose intensities may fall below the detection threshold because their ionization
246 efficiency is low. The fifth dataset (Dataset 5) comprised four filters and was obtained by
247 applying an *outliers filter* to Dataset 4. To this end, Pearson correlations between log10-
248 transformed intensities were computed for each pair of peptide ions belonging to the same
249 protein. To avoid bias induced by outlier values due to individual technical variations, the
250 correlations were computed on mean values of peptide ion intensities per concentration. The
251 peptide ion with the highest number of coefficients of correlation greater than or equal to the
252 mean of the positive coefficients of correlation was chosen as a reference for the protein. The
253 peptide ions showing a non-significant correlation to the reference (p-value ≥ 0.01) or whose
254 coefficients of correlation to the reference were lower than 0.8 were considered as outliers and
255 were removed (*outliers filter*). Proteins quantified by fewer than two peptide ions were
256 removed from all the datasets. Missing intensity values were not imputed. Consequently, the
257 number of peptide ions used to compute protein abundances could vary from one sample to
258 another.

259

260 **Protein quantification**

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

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

$$269 \quad I_{ijk} = \mu + A_i + R_j + P_k + \theta_{ij} + \epsilon_{ijk}$$

$$270 \quad \text{where} \quad \theta_{ijk} \sim N(0, \sigma_{\theta}^2)$$

$$271 \quad \epsilon_{ijk} \sim N(0, \sigma_{\epsilon}^2)$$

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

285

286 **RESULTS AND DISCUSSION**

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

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

306

307 **1. The amplitude of peptide filtering affects protein data composition**

308 Yeast and UPS1 datasets were differently affected by the filters. The proportion
309 of shared peptides removed was much higher for yeast than for the UPS1 standard (-4.2% vs -
310 0.8%, respectively). Although the UPS1 standard was designed to contain few proteins with
311 similar sequences, yeast is a living organism that contains many duplicated genes resulting
312 from whole genome duplication and other small-scale duplications .

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

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

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

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

361

362 **2. Quantification methods do not respond equally to peptide filters**

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

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

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

393 of abundances of the proteins they belong to , shared peptides are globally more intense than
394 proteotypic peptides. When taken into account, they can therefore lead to strongly
395 overestimating protein abundances, especially when computed by *iBAQ* and *Average* (Figure
396 4A). These results indicate that in the case of these two quantification methods, it is important
397 to filter not only shared peptides but also all types of dubious peptide ions of high intensity
398 (see for example Figure 4B). By contrast, *Average-Log* and *Model* were only slightly
399 improved by the *shared peptide filter*: both methods are based on log-transformed intensities,
400 where the difference between peptide ions of high and low intensity is reduced. In addition,
401 the *Model* discards the peptide ion effect, which results in a similar weight of all peptides in
402 the computation of protein abundance.

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

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

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

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

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

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

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

477

478 **3. The performances of one quantification method over another depend on how the data** 479 **were filtered**

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

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

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

499

500 CONCLUSION

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

523

524

525 **ACKNOWLEDGEMENTS**

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

529

530

531 **FIGURE LEGENDS**

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

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

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

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

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

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

575

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

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

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

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

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

593 **Figure S6** Accuracy of UPS1 protein abundance according to number of peptides. For the
594 whole experimental setup and the five methods of quantification, 35 UPS1 proteins (detected
595 in the five datasets) were split into two groups -lower (orange boxplots) and higher (green
596 boxplots)- according to median of peptide number determined in each dataset (Dataset 1 and
597 2: 18 peptides, Dataset 3: 17 peptides, Dataset 4 and 5: 6 peptides). Accuracy was computed
598 at each UPS1 concentration as inverse of coefficient of variation (CV) between average
599 abundances of UPS1 proteins (n = 3 replicates).

600

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

608

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-Pomare
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.

- 639 [10] X. Lai, L. Wang, H. Tang, F.A. Witzmann, A Novel Alignment Method and Multiple
640 Filters for Exclusion of Unqualified Peptides To Enhance Label-Free Quantification
641 Using Peptide Intensity in LC—MS/MS, *J. Proteome Res.* 10 (2011) 759–785.
642 doi:10.1146/annurev-cellbio-092910-154240.Sensory.
- 643 [11] B.J.M. Webb-Robertson, L.A. McCue, K.M. Waters, M.M. Matzke, J.M. Jacobs, T.O.
644 Metz, S.M. Varnum, J.G. Pounds, Combined statistical analyses of peptide intensities
645 and peptide occurrences improves identification of significant peptides from MS-based
646 proteomics data, *J. Proteome Res.* 9 (2010) 5748–5756. doi:10.1021/pr1005247.
- 647 [12] A.D. Polpitiya, W.-J. Qian, N. Jaitly, V.A. Petyuk, J.N. Adkins, D.G. Camp, G.A.
648 Anderson, R.D. Smith, DAnTE: a statistical tool for quantitative analysis of -omics
649 data, *Bioinformatics.* 24 (2008) 1556–1558. doi:10.1093/bioinformatics/btn217.
- 650 [13] J. Forshed, H.J. Johansson, M. Pernemalm, R.M.M. Branca, A. Sandberg, J. Lehtiö,
651 Enhanced Information Output From Shotgun Proteomics Data by Protein
652 Quantification and Peptide Quality Control (PQPQ), *Mol. Cell. Proteomics.* 10 (2011)
653 M111.010264. doi:10.1074/mcp.M111.010264.
- 654 [14] B. Zhang, M. Pirmoradian, R. Zubarev, L. Käll, Covariation of Peptide Abundances
655 Accurately Reflects Protein Concentration Differences, *Mol. Cell. Proteomics.* (2017)
656 1–42.
- 657 [15] M. Blein-Nicolas, M. Zivy, Thousand and one ways to quantify and compare protein
658 abundances in label-free bottom-up proteomics, *Biochim. Biophys. Acta - Proteins
659 Proteomics.* 1864 (2016) 883–895. doi:10.1016/j.bbapap.2016.02.019.
- 660 [16] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, M.
661 Selbach, Global quantification of mammalian gene expression control, *Nature.* 473
662 (2011) 337–342. doi:10.1038/nature10098.
- 663 [17] J.C. Silva, M. V. Gorenstein, G.-Z. Li, J.P.C. Vissers, S.J. Geromanos, Absolute
664 Quantification of Proteins by LCMS^E, *Mol. Cell. Proteomics.* 5 (2006) 144–156.
665 doi:10.1074/mcp.M500230-MCP200.
- 666 [18] R.E. Higgs, M.D. Knierman, V. Gelfanova, J.P. Butler, J.E. Hale, Comprehensive label-
667 free method for the relative quantification of proteins from biological samples, *J.
668 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. Lubner, 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. Duanes, 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

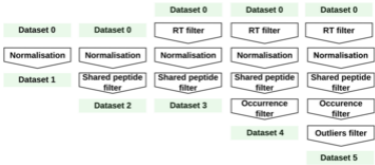
728 **Table 1** Effect of filters on number of peptide ions and proteins. Numbers in parenthesis
729 indicate percentage of data removed by filter from previous dataset.

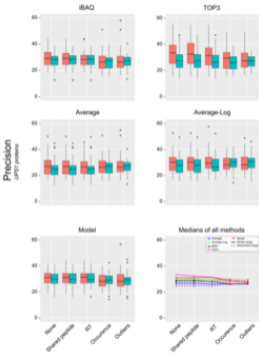
		No filter	<i>Shared peptide filter</i>	<i>RT filter</i>	<i>Occurrence filter</i>	<i>Outliers filter</i>
Peptide ions	Total	22 950	22 044 (-3.9%)	21 857 (-0.8%)	13 561 (-38.0%)	4 882 (-64.0%)
	Yeast	21 820	20 915 (-4.2%)	20 778 (-0.7%)	13 314 (-35.9%)	4 666 (-65.0%)
	UPS1	1 138	1 129 (-0.8%)	1 079 (-4.4%)	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%)

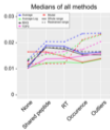
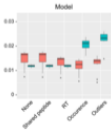
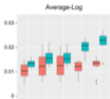
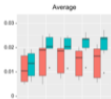
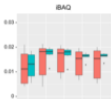
730

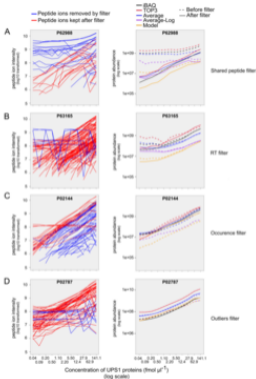
731

732

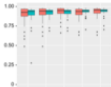








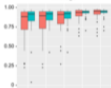
IBAQ



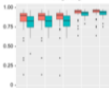
TOP3



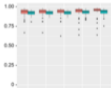
Average



Average-Log



Model



Medians of all methods

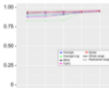
Linearity
LysEY proteinsNone
Missing proteins
RT
Outliers
OutliersNone
Missing proteins
RT
Outliers
Outliers

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	<i>Shared peptide filter</i>	<i>RT filter</i>	<i>Occurrence filter</i>	<i>Outliers filter</i>
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 ●

Occurrence

Dataset 4 ◆

Outliers

Dataset 5 ★

Protein quantification methods

iBAQ

TOP3

Average

Average-Log

Model

Methods Performances

Accuracy

Linearity

