Journal of proteome research

Article

pubs.acs.org/jpr

Summarization vs Peptide-Based Models in Label-Free Quantitative Proteomics: Performance, Pitfalls, and Data Analysis Guidelines

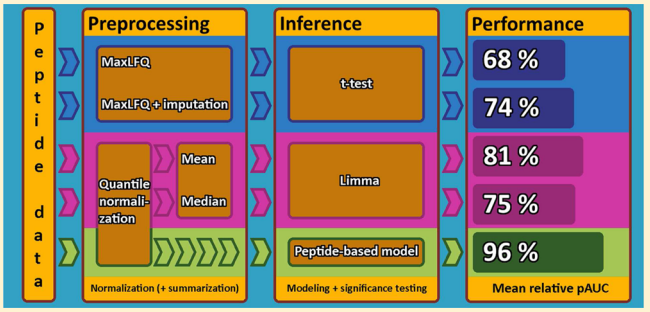
Ludger J. E. Goeminne,[†,‡,§,∥,⊥] Andrea Argentini,[‡,§,⊥] Lennart Martens,[‡,§] and Lieven Clement*[,†]

[†]Department of Applied Mathematics, Computer Science, and Statistics, [‡]Department of Medical Protein Research, VIB, and [§]Department of Biochemistry, Ghent University, 9000 Ghent, Belgium

[∥]Department of Plant Systems Biology, VIB, Ghent University, 9052 Ghent, Belgium

Supporting Information

ABSTRACT: Quantitative label-free mass spectrometry is increasingly used to analyze the proteomes of complex biological samples. However, the choice of appropriate data analysis methods remains a major challenge. We therefore provide a rigorous comparison between peptide-based models and peptide-summarization-based pipelines. We show that peptide-based models outperform summarization-based pipelines in terms of sensitivity, specificity, accuracy, and precision. We also demonstrate that the predefined FDR cutoffs for the detection of differentially regulated proteins can become problematic when differentially expressed (DE) proteins are highly abundant in one or more samples. Care should therefore be taken when data are interpreted from samples with spiked-in internal controls and from samples that contain a few very highly abundant proteins. We do, however, show that specific diagnostic plots can be used for assessing differentially expressed proteins and the overall quality of the obtained fold change estimates. Finally, our study also illustrates that imputation under the "missing by low abundance" assumption is beneficial for the detection of differential expression in proteins with low abundance, but it negatively affects moderately to highly abundant proteins. Hence, imputation strategies that are commonly implemented in standard proteomics software should be used with care.



KEYWORDS: *differential proteomics, data analysis, linear model*

1. INTRODUCTION

Current high throughput mass spectrometry (MS) experiments enable the simultaneous identification and quantification of thousands of peptides and proteins in biological samples under various experimental conditions. These methods allow us to extend our understanding of biological processes and are important for the identification of biomarkers for the early detection, diagnosis, and prognosis of disease. Quantitative proteomics workflows broadly fall into two categories: labeled approaches and label-free approaches.[1] Labeled workflows rely on the labeling of proteins or peptides with isobaric or isotopic mass tags and are currently more commonly used. Label-free proteomics workflows, however, do not require these additional labor intensive and expensive sample processing steps.[2] Label-free approaches can perform quantitative proteome comparisons among an unlimited number of samples and can also be applied retroactively to previously acquired data.[3] Although label-free quantifications tend to have slightly higher coefficients of variation compared to SILAC labeling, label-free quantifications are more reproducible and can identify up to 60% more proteins than labeled quantifications.[4,5]

A typical label-free shotgun MS-based proteomics workflow consists of (a) a protein extraction step followed by enzymatic digestion, (b) reverse phase high performance liquid chromatography (HPLC) separation, (c) mass spectrometry (MS), (d) a data analysis step involving the identification and quantification of peptides and proteins, and (e) a statistical analysis for assessing differential protein abundance.[6,7] In a typical data-dependent analysis, selected peptides are isolated and fragmented, generating a fragmentation spectrum that is then used for peptide identification.[8] Technological constraints, however, limit the number of peptides in each fraction that can be selected for fragmentation. As the selection criteria typically involve the MS peak intensities in a particular time window, the identifications in MS-based experiments are inherently associated with the abundance of ionized peptides. Moreover, the steric effects of digestion enzymes[9] and differences in ionization efficiency favor particular peptides. Coeluting peptides heavily influence the observed MS intensities.[10] Hence, proteomics data suffer from nonrandom missing values and a large variability, rendering the development of reliable data analysis pipelines for quantitative proteomics a challenging task.[11]

The current data analysis strategies for label-free quantitative proteomics are typically based on spectral counting or peak

Received: December 9, 2014
Published: April 1, 2015

DOI: 10.1021/pr501223t
J. Proteome Res. 2015, 14, 2457−2465

intensities.[1] In the former approach, the number of peptide-to-spectrum matches (PSMs) for a given peptide are counted, and these are then accumulated over all peptides from a given protein.[12] Even though these methods are very intuitive and easy to apply, they remain controversial.[13,14] Moreover, these methods necessarily ignore a large part of the information available in high precision mass spectra and are not very efficient in detecting low fold changes.[15] Peak-intensity-based methods, however, use the maximum intensity or the area under the peak as a proxy for peptide abundance and tend to produce more precise protein abundance estimates.[15] We therefore focus on these latter approaches.

Many peak-based data analysis methods for the preprocessing and differential analysis of quantitative label-free proteomics data have been described in the literature. Modular approaches consisting of a separate normalization, summarization, and data analysis step are commonly used.[16,17] Peptides originating from the same protein can indeed be considered technical replicates and theoretically should lead to similar abundance estimates. However, the summarization of the peptide intensities into protein expression values is cumbersome, and most summarization-based methods do not correct for differences in peptide characteristics or for the between-sample differences in the number of peptides that are identified per protein. This might introduce bias and differences in uncertainty between the aggregated protein expression values, which are typically ignored in downstream data analysis steps. The aforementioned nonrandom character of missing peptides further exacerbates these issues.

In response, linear regression approaches have been developed that immediately estimate the differential abundance between the proteins from observed peptide intensities, and their authors have made bold claims on their performance.[18,19] Objective comparisons and general guidelines for the practitioner are, however, still lacking, which impedes the dissemination of more efficient data analysis pipelines into the proteomics community.

In this paper, we therefore present a rigorous comparison among modular and peptide-based regression methods for analyzing label-free quantitative proteomics data. We exploit the availability of the benchmark data sets to provide insight into the performance differences and technological artifacts that often arise in label-free proteomics experiments. It should also be noted that the benchmark data used here present a range of concentration differences, which enables us to analyze the suitability of different methods for different situations (e.g., small abundance differences versus large abundance differences or few missing peptides versus many missing peptides across analyses). In section 2 we present the benchmark data, the different data analysis methods, and the performance criteria that will be used in our comparison. The results are presented and discussed in sections 3 and 4.

2. MATERIALS AND METHODS

We used the publicly available data set from Study 6 of the Clinical Proteomic Technology Assessment for Cancer (CPTAC) Network[20] for assessing the performance of different data analytic workflows for quantitative label-free proteomics. In the CPTAC study, a mixture of 48 human proteins from the Sigma-Aldrich Universal Proteomics Standard 1 (UPS) was spiked into a 60 ng of protein/μL resuspended yeast lysate of *Saccharomyces cerevisiae* strain BY4741 (*MATa, leu2Δ0, met15Δ0, ura3Δ0,* and *his3Δ1*). Spike-ins were performed at five different concentrations: 0.25 fmol of UPS protein/μL (A), 0.74 fmol of UPS protein/μL (B), 2.22 fmol of UPS protein/μL (C), 6.67 fmol of UPS protein/μL (D), and 20 fmol of UPS protein/μL (E). The prepared samples were then sent to five different laboratories and analyzed on four different mass spectrometry platforms.

We identified peptides by searching the data using MaxQuant v1.5 against the yeast UniprotKB/Swiss-Prot protein database (v 15.14) to which the 48 UPS protein sequences were added. Detailed search settings can be found in the Supporting Information. A general overview of the number of identified peptides and proteins in our search can be found in Table S1, Supporting Information. Statistical analyses were implemented in RStudio version 0.98.978 (RStudio, Boston, MA) interfacing R 3.1.0 ("Spring Dance"). Standard Perseus analysis workflows were executed in Perseus version 1.5. We introduce two Perseus workflows in section 2.1, two different modular pipelines that aggregate peptide intensities into protein expression values in section 2.2, and three different peptide-based regression methods in section 2.3. The performance criteria used to compare the different methods can be found in section 2.4.

2.1. Perseus-Based Workflows

We used a typical workflow implemented in the software package Perseus. The analysis starts from (a) the LFQ intensities given in the MaxQuant's proteinGroups.txt file, which consist of normalized and summarized intensities at protein level. The MaxLFQ procedure then proceeds as follows: for all pairwise comparisons of a protein between samples, the median ratio for the common peptides in both samples is calculated. Next, the abundance protein profile that optimally satisfies these protein ratios is reconstructed with a least-squares regression model. The whole profile is then rescaled to the cumulative intensity across the samples with preservation of the total summed intensity for a protein across the samples. As the resulting LFQ intensities are already normalized by the MaxLFQ procedure,[21] no additional normalization step is required. In (b), the LFQ protein intensities are read into Perseus. The proteins that are only identified by a modification site, the contaminants, and the reversed sequences are removed from the data set, and the remaining intensities are $\log_2$-transformed. Next, (c) involves the imputation of missing values using Perseus' standard settings.[22] Finally, (d) consists of inference by pairwise two-sample t tests. The multiple testing problem is addressed using the Benjamini–Hochberg False Discovery Rate (FDR) procedure.[23] The (a)–(d) pipeline is referred to as perseusImp. We also consider a second variant, perseusNoImp, in which the imputation step (c) is omitted.

2.2. Summarization-Based Workflows

A typical modular workflow for quantitative proteomics consists of a normalization, summarization, and statistical analysis step.[6,7] In our contribution, we assess two customized pipelines that build upon popular mean and median summarization strategies for the summarization of peptide intensities into protein expression values. The following steps are considered in the analysis pipelines: (a) the intensities from MaxQuant's peptides.txt output file are $\log_2$-transformed and normalized using quantile normalization (with the peptides mapping to reversed sequences or mapping to multiple proteins being removed from the data), (b) peptide intensities are aggregated into protein expression values using mean or

DOI: 10.1021/pr501223t
J. Proteome Res. 2015, 14, 2457–2465

median summarization, and (c) summarized protein expression values are further analyzed using empirical Bayes moderated *t* tests implemented in the R/Bioconductor package "limma".[24] The Benjamini–Hochberg FDR procedure is used to correct for multiple testing. The two resulting methods are referred to as limmaMean and limmaMedian.

In the limma analysis, the following model is considered for each protein i:

$$y_{ikl} = treat_{ik} + exp_{il} + \varepsilon_{ikl} \quad (1)$$

with y_{ikl} being the aggregated protein intensity for the k-th treatment (*treat*) and the l-th experiment (*exp*) correcting for (lab × instrument × repeat) batch effects. ε_{ikl} is a random error term that is assumed to be normally distributed with mean 0 and variance σ_i^2. Note that the *treat* effect is the effect of interest. Contrasts between *treat* parameters can be interpreted as $\log_2$ fold changes for protein i. For instance, k = A indicates condition A (spike-in concentration of 0.25 fmol of UPS protein/μL) and k = E indicates condition E (spike-in concentration of 20 fmol of UPS protein/μL). If so, then $treat_{iE} - treat_{iA}$ indicates the expected $\log_2$ difference in concentration for protein i between group E and group A. The statistical significance of the contrasts can be addressed by using *t* tests. The limma analysis exploits the massively parallel nature of quantitative proteomics experiments and allows for the borrowing of strength across proteins to estimate the error variance, i.e., makes use of a moderated empirical Bayes variance estimator $\tilde{s}_i^2$:

$$\tilde{s}_i^2 = \frac{d_0 {s_0}^2 + d_i {s_i}^2}{d_0 + d_i}$$

with s_i and d_i being the standard deviation and the residual degrees of freedom for protein i, respectively, s_0 the estimated prior standard deviation, and d_0 the prior degrees of freedom. Both the prior standard deviation and the prior degrees of freedom are estimated using empirical Bayes by pooling information across all proteins. Hence, the protein-based variance s_i^2 is shrunk toward a common variance s_0^2, leading to more stable variance estimates ($\tilde{s}_i^2$). Note that the degrees of freedom from the moderated *t* test also increase to $d_0 + d_i$. Detailed information can be found in the work of Smyth.[24]

2.3. Peptide-Based Models

Peptide-based models use the MaxQuant peptides.txt file as input. In (a), the extracted peptide intensities are $\log_2$-transformed and quantile normalized (Figures S8 and S9, Supporting Information), and peptides mapping to reversed sequences or mapping to multiple proteins are removed from the data. In (b), the peptide data are modeled with three different candidate models. In (c), inference is done by pairwise contrast testing. Multiple testing is addressed using the Benjamini–Hochberg FDR.

Linear Model without Sample Effect. For each protein i, the following model is proposed:

$$y_{ijklm} = pep_{ij} + treat_{ik} + exp_{il} + \varepsilon_{ijklm} \quad (2)$$

with y_{ijklm} being the $\log_2$-transformed intensity for the j-th peptide sequence pep_{ij} of the k-th treatment $treat_{ik}$ and the l-th experiment exp_{il}. ε_{ijklm} is a normally distributed error term with mean 0 and variance σ_i^2. The index m refers to multiple spectra that are identified for the same peptide in the same experiment and the same treatment. Contrasts in $treat_{ik}$ parameters can again be interpreted as $\log_2$ fold changes for protein i. The model also incorporates a pep_{ij} effect to account for peptide-specific fluctuations around the mean protein intensity, which originate from differences in digestion and ionization efficiency, among others.[9]

Linear Model with Sample Effect. Model 2 is extended by incorporating an additional sample effect, $sample_{ikl}$, to capture deviations specific to each MS run (lab × instrument × treatment × repeat):

$$y_{ijklm} = pep_{ij} + treat_{ik} + exp_{il} + sample_{ikl} + \varepsilon_{ijklm} \quad (3)$$

Note that all remaining effects are similar to those of model 2.

Mixed Model with Random Sample Effect. The mixed model extends the linear model 3 by putting a normal prior on the sample effect, $sample_{ikl} \sim N(0,\sigma_{sample,i}^2)$. This model accounts for the correlation within samples and incorporates both within- and between-sample variability when inference is performed on contrasts in the $treat_{ik}$ effects. The degrees of freedom of the *t* tests are approximated using the Satterthwaite approximation,[25] and the Benjamini–Hochberg FDR procedure is used to account for multiple testing.[23]

2.4. Performance

For each method, *p* values are converted to *q* values using the Benjamini–Hochberg FDR procedure,[23] and a cutoff is set at 5% FDR. At this level, the number of false positives (FP), true positives (TP), false negatives (FN), and true negatives (TN) are recorded, and the nominal FDR level is compared to the observed false discovery rate $\overline{FDR} = FP/(FP + TP)$. Note that the observed FDR equals 1 minus the positive predictive value, $PPV = TP/(FP + TP)$.

The ROC curves are constructed on the basis of the ordering of the *p* values. Bias (Figures S4 and S5, Supporting Information), standard deviation (sd), median absolute deviation (mad), and root mean squared error (RMSE) (Figures S6 and S7, Supporting Information) are calculated both for yeast and for UPS proteins. We also calculated the F1 score, which is defined as the harmonic mean of the PPV and the sensitivity. Higher F1 scores indicate that a method provides a good balance between the PPV and the recall (Figures S1 and S2, Supporting Information).

3. RESULTS

We investigated the sensitivity, specificity, and F1 score of the test procedure as well as the accuracy and the precision of the fold change (FC) estimates for three peptide-based methods and four summarization-based data analysis pipelines using the CPTAC Study 6 data set.[20] This data set consists of samples with a uniform yeast proteome background in which human UPS peptides are spiked at five different concentrations (0.25, 0.74, 2.22, 6.67, and 20 fmol/μL). All 10 pairwise comparisons are assessed in each analysis. The following peptide-based methods are considered: a linear model without sample effect (lmNoSamp), a linear model with sample effect (lmSamp), and a mixed model with a random sample effect (mixedSamp). The summarization-based approaches consist of mean and median summarizations of peptides into protein expression values followed by limma analyses (limmaMean and limmaMedian) as well as the more advanced MaxLFQ summarization[21] followed by a standard Perseus workflow with and without imputation (perseusImp and perseusNoImp). All peptide identifications and intensities were based on MaxQuant so as to avoid biases due to the search engine or peak intensity calculation algorithm.

DOI: 10.1021/pr501223t
J. Proteome Res. 2015, 14, 2457–2465

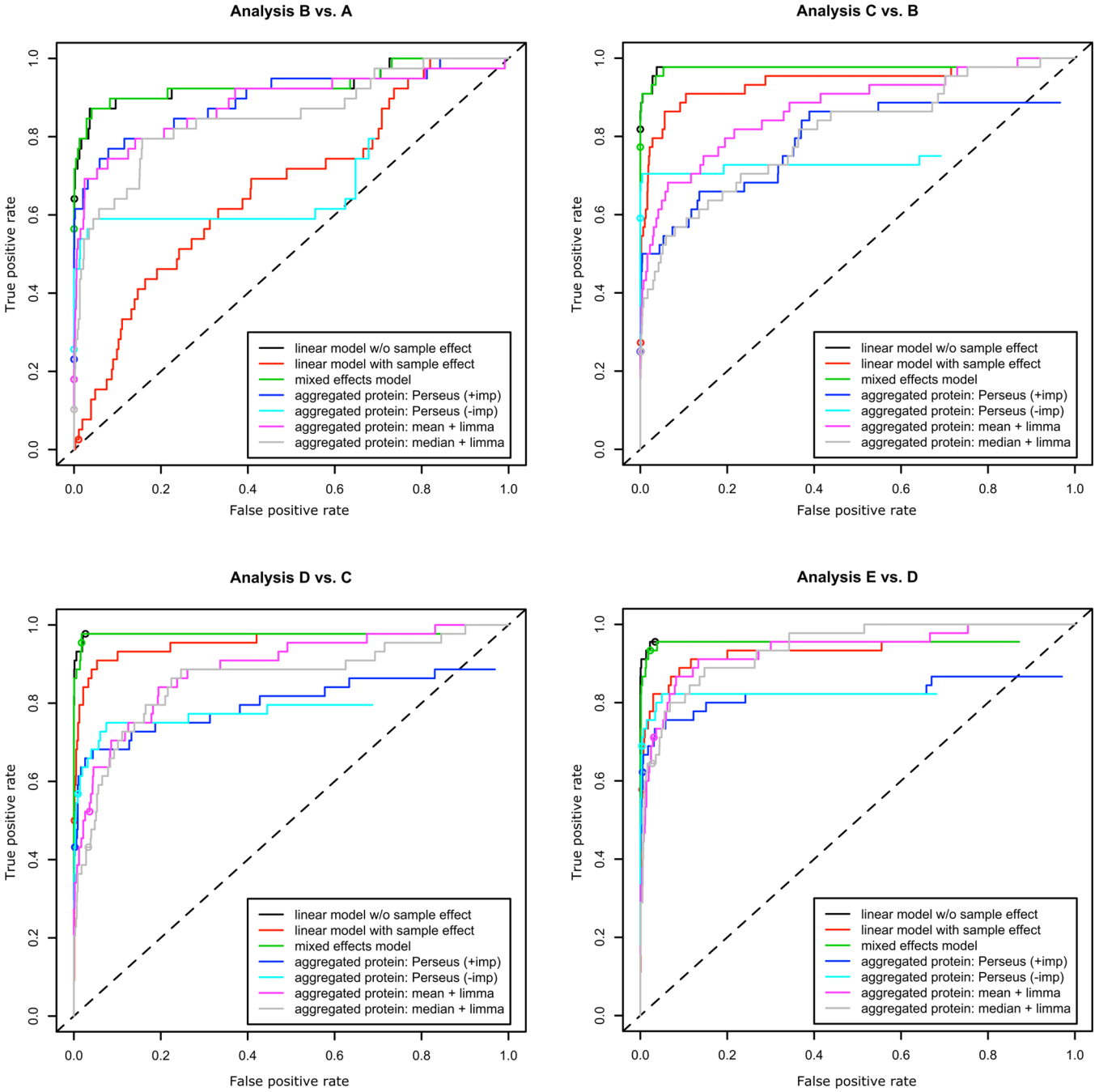


Figure 1. Receiver operating characteristic (ROC) curves for the seven analysis methods in comparisons B−A, C−B, D−C, and E−D. The UPS proteins in these comparisons were spiked in at a ratio close to 3:1. Dots denote the estimated cutoff for each method at 5% FDR. The termination of the curve before the point (1, 1) indicates either that proteins are prematurely removed from the analysis (e.g., for the Perseus workflows) or that there is an inability of the models to fit a protein with too few observations (e.g., for peptide-based models).

Receiver operating characteristic (ROC) curves for the four comparisons with the smallest differences in spiked-in protein abundance (B−A, C−B, D−C, and E−D) are shown in Figure 1. Detecting the differential abundance of UPS proteins is most challenging in these comparisons as they only involve fold changes (FCs) very close to 3. ROC curves for the six remaining comparisons can be found in Figure S1 in the Supporting Information. Figure 1 shows that the lmNoSamp and mixedSamp models clearly outperform the other methods. The lmNoSamp, mixedSamp, and perseusImp workflows do control the FDR at 5% for comparisons B−A, C−A, and C−B, but perseusNoImp could only control the FDR for comparisons B−A and C−B, and both limmaMean and limmaMedian could only control the FDR for comparison B−A (see Tables S2−S8 and S12 in the Supporting Information). The lmSamp method is unable to control the FDR. When differences in spiked-in concentrations increase, however, none of the methods are able to control the FDR correctly (Tables S2−S8 and S12 in the Supporting Information). lmSamp is more conservative but cannot control its FDR at 5%, either. The ROC curves also show that the mean summarization outperforms the more robust but less efficient median summarization.

DOI: 10.1021/pr501223t
J. Proteome Res. 2015, 14, 2457−2465

Table 1. Relative Partial Area under the Curve (rpAUC) for FPR <0.1 for All Seven Models for Each of the Ten Comparisons[a]

proteome comparison	lmNoSamp	lmSamp	mixedSamp	perseusImp	perseusNoImp	limmaMean	limmaMedian
B−A	83.01%	12.41%	83.93%	69.65%	55.70%	68.14%	56.21%
C−A	98.33%	49.31%	98.60%	85.93%	59.76%	87.22%	77.26%
D−A	99.20%	72.65%	99.26%	86.31%	63.42%	96.79%	95.18%
E−A	99.72%	89.06%	99.72%	89.62%	63.79%	97.92%	95.93%
C−B	95.10%	77.81%	94.60%	52.45%	70.08%	61.79%	50.54%
D−B	97.05%	89.60%	96.72%	71.00%	72.06%	89.21%	83.54%
E−B	96.98%	93.50%	95.90%	77.68%	72.41%	90.94%	87.69%
D−C	96.51%	84.85%	96.01%	64.48%	67.09%	59.77%	54.25%
E−C	97.34%	94.48%	96.21%	72.92%	74.85%	81.94%	79.23%
E−D	94.06%	79.45%	92.76%	71.13%	78.02%	74.16%	71.21%
mean	95.73%	74.31%	95.37%	74.12%	67.72%	80.79%	75.11%

[a]The UPS proteins were spiked in at concentrations ranging from 0.25−20 fmol/μL (conditions A−E).

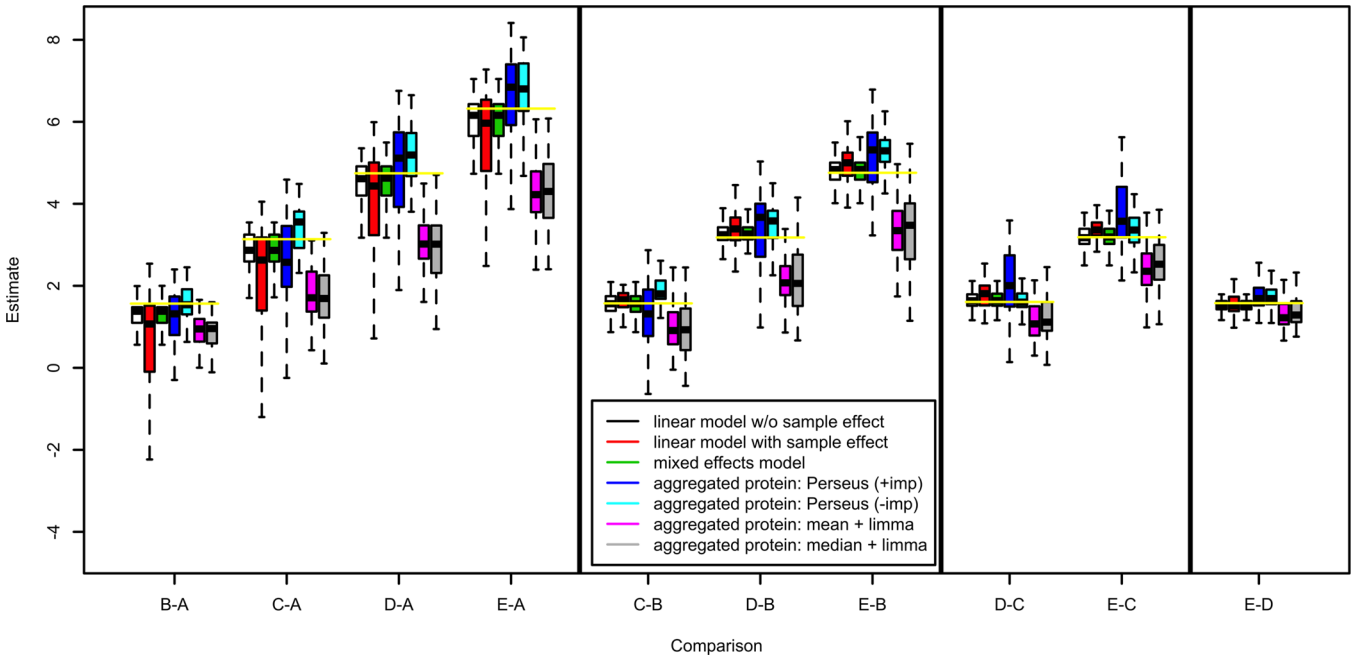


Figure 2. Boxplots showing the distributions of the DE estimates of the UPS proteins for each of the seven methods in each of the 10 comparisons. Outliers are not shown. The actual fold changes of the spikes are indicated with the yellow horizontal lines.

As only a part of the ROC curve is relevant in practice, i.e., that experimenters typically want to restrict the number of candidate proteins for validation in follow-up experiments, we also compared the relative partial areas under the curve (rpAUC) for FPR <0.1. Relative pAUCs (Table 1) are obtained by dividing pAUC values (Table S13, Supporting Information) by the maximum pAUC value of 0.1. Table 1 also demonstrates that the lmNoSamp and mixedSamp models are superior to the competing pipelines in terms of pAUC. Their power is higher in spite of the fact that no information is borrowed across proteins for estimating the variance (as compared to the limma workflows) and that there is an absence of imputation (as compared to the standard Perseus method). perseusImp outperforms limmaMean and limmaMedian in terms of pAUC when differential expression in very-low-abundance proteins needs to be detected (e.g., comparison B−A). In these situations, imputation under the assumption of low abundance strongly boosts the performance of the method. The perseusImp workflow outperforms the perseusNoImp workflow for all comparisons involving A, i.e., when very-low-abundance differentially expressed (DE) proteins are involved in the comparison. But in comparisons with more abundant UPS spikes, the opposite is observed, and perseusImp shows a suboptimal performance compared to that of perseusNoImp.

When the F1 score is examined, the lmNoSamp and mixedSamp models show very comparable patterns (Figure S2, Supporting Information). In comparisons E−A, E−B, E−C, and D−B, the lmSamp is superior to the other peptide-based models. This is most likely due to its more conservative nature. lmNoSamp and mixedSamp suffer from many false positives for these comparisons. For the summarization-based models (Figure S3, Supporting Information), we notice that perseusNoImp outperforms the other summarization-based methods for most comparisons. In comparisons D−A, D−B, E−A, E−B, and E−C, perseusImp shows a higher F1 score than Perseus without imputation. Again, the mean summarization method almost consistently outperforms the median summarization in terms of F1 score, although the differences are generally not very large.

The accuracy and precision of the pipelines are assessed by comparing the differential expression estimates to the true $\log_2$ fold changes of the spiked UPS peptides [$\log_2$ FC $\approx \log_2(3)$,

DOI: 10.1021/pr501223t
J. Proteome Res. 2015, 14, 2457−2465

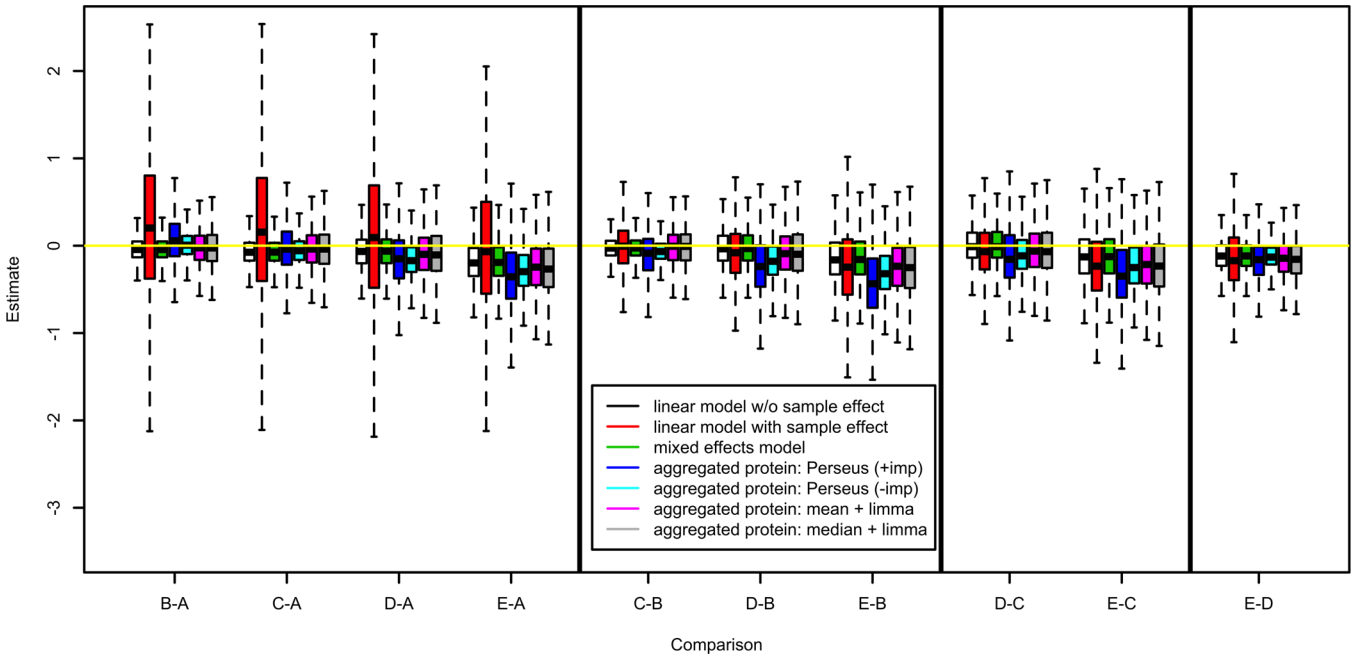


Figure 3. Boxplots showing the distributions of the DE estimates of the yeast proteins for each of the seven methods in each of the 10 comparisons. Outliers are not shown. All samples consisted of the same yeast background. Hence, no differential expression should occur for these proteins.

$\log_2(9)$, $\log_2(27)$, and $\log_2(80)$] and the yeast peptides ($\log_2$ FC = 0). Figures 2 and 3 show boxplots of the different DE estimates of the different methods for UPS and yeast proteins, respectively. The actual $\log_2$ FC is also indicated in the plot.

Figure 2 illustrates that the lmNoSamp and mixedSamp models are superior to the other methods in terms of both the accuracy and the precision of the FC estimates for the differentially abundant UPS proteins. The mean and median summarization methods systematically show a downward bias (Figure S4, Supporting Information). The bias is more pronounced in comparisons involving condition A. In condition A, the lowest concentration UPS (0.25 fmol/μL) is spiked in and, consequently, fewer UPS peptides are identified. Missingness, however, can be expected to involve peptides with a lower ionization efficiency, which typically display lower peak intensities than other peptides of the same protein. The simple mean and median summarization methods do not correct for differences in peptide characteristics, leading to an overestimation of the expression value for UPS proteins in condition A. This leads to moderation of the $\log_2$ fold change estimates involving condition A. Peptide-based pipelines correcting for peptide effects also suffer from a slight negative bias in comparisons that involve condition A. Note that imputation has a severe impact on the precision. Figure S7 in the Supporting Information also shows that summarization-based limmaMean and limmaMedian methods give the highest root mean squared error (RMSE = $[\text{bias}^2/\text{variance}]^{1/2}$) among all methods that were evaluated.

Figure 3 confirms that the lmNoSamp and the mixedSamp models are favorable in terms of accuracy and precision. For the yeast proteins (non-DE), the median and mean summarization methods show a bias similar to that of competing methods. An increasing downward bias of the $\log_2$ FC estimates can be observed for the null proteins (yeast) in comparisons involving increasing UPS concentrations. This becomes very apparent for comparisons that involve condition E. In this condition, a very high fraction of the total protein mass in the sample consists of UPS proteins. Hence, yeast peptides are likely to be masked by UPS, leading to an underestimation of the abundance of yeast peptides in the D and E mix. Most false positive yeast proteins had negative $\log_2$ FC estimates as opposed to the spiked UPS proteins, which show positive $\log_2$ FC estimates in each comparison. Therefore, issues involving the FDR are likely to be linked to the extreme sample composition under conditions D and E, which invokes an MS bias.[10] The F1 score masks this artifact, as it combines PPV and sensitivity. The same trend is also visible in MA plots for the linear model without sample effect (Figures S10 and S11, Supporting Information). In these graphs, the average FC is plotted in function of the average protein expression for a particular comparison. These graphs are therefore very helpful for screening for artifacts induced by the technical and data analysis workflows. Figure 3 also illustrates that the precision reduces with increasing FC, i.e., for comparisons involving conditions D and E. Finally, the lmSamp method shows a dramatic decrease in precision for comparison B−A. This is a data analysis artifact; the model is overidentified for many proteins, leading to the aliasing of sample and treatment effects. Due to the specific model parametrization, the overidentification has a larger impact on comparisons involving condition A.

We also investigated alternative data analysis strategies to alleviate this problem. For the peptide-based lmNoSamp method, we assessed the impact of testing against the median $\log_2$ FC of all proteins instead of testing against 0. This slightly improves the observed FDR except for comparisons C−A and D−B and improves the rpAUC except for comparisons D−A, D−B, D−C, and E−C (Table S14, Supporting Information). However, the method still returns too many false positives for the comparisons involving high concentrations (Table S9, Supporting Information). For the summarization-based methods, we assessed the impact of switching the order of the normalization and summarization steps. When the quantile normalization is performed after summarization, the observed FDR improved for comparisons involving D and E, but the

DOI: 10.1021/pr501223t
J. Proteome Res. 2015, 14, 2457−2465

performance decreased dramatically for comparisons B−A and C−A (Tables S10−S12. Supporting Information). The ROC curves also suggest that switching the order of the normalization and summarization steps deteriorates the performance of the limmaMean and limmaMedian workflows (Figure S12, Supporting Information).

4. DISCUSSION

Our analysis showed that peptide-based models perspicuously outperform summarization-based methods. Both the linear model without sample effect and the mixed model outperform the other methods in terms of accuracy, precision, sensitivity, and specificity. The ROC curves clearly indicate that these methods produce a more reliable ordering of DE proteins than the competing methods. The linear model with sample effect has a suboptimal performance but still outperforms the other methods in comparisons that do not involve A. Due to selective and periodic sampling in both MS stages, not all peptides are being observed or identified in all samples. Moreover, intensities from different peptides of the same protein vary considerably due to differences in cleavage and ionization efficiency among others.[26,27] Summarization thus typically involves different peptides and a different number of peptides in each sample. This leads to protein expression values with distinct characteristics, which induces bias and incorrect precision of the fold change estimates.

Peptide-based models are superior in correcting for individual peptide effects, which are typically quite strong[18,19] and accounting for the different number of peptides in each sample. Thus, bias is reduced and improved precision estimates are provided, leading to higher sensitivity and specificity. The mixed model can also account for the correlation that is present in peptides from the same protein within a sample. The peptide-based models with a fixed sample effect suffer from the unstable estimation of fold changes and variance components due to the overfitting of sparse proteins identified by a few peptides. Moreover, the inclusion of a fixed sample effect eliminates the between-sample variability from the analysis. Inference between the samples will be based on an underestimated variance, leading to a higher number of false positives in a top list. In the linear model without sample effect, fewer parameters have to be estimated, and the variances within and between samples are combined in the error term. Hence, the method is less prone to overfitting and incorporates both within- and between-sample variability in the test statistics, leading to a better control of the number of false positives. However, the method does not account for the correlation between peptides from a particular protein within a sample. The mixed modeling approach with a random sample effect does incorporate within- and between-sample variances as well as the within-sample correlation between peptides of the same protein. The mixed model and the linear model without sample effect are more or less on par in terms of all assessed performance criteria. Hence, the increased computational complexity of the mixed model cannot be justified for this particular application. However, in real experiments, more correlation can be expected due to the additional biological variation among samples.

We also showed that the use of FDR thresholds might be flawed under certain experimental conditions. This was observed for comparisons involving conditions D and E, i.e., the samples with the highest spiked-in UPS concentrations. Under these conditions, the UPS proteins correspond to a considerable fraction of the total protein mass in the sample. The ROC curves show that peptide-based methods still produce reliable top lists with a superb ordering, but the use of a 5% FDR threshold was too liberal. Hence, long protein lists are produced with many false positives. The majority of these false positives, however, had FC estimates in the opposite direction as those of spiked UPS proteins. This was due to a systematic downward bias in the FC estimates of non-differentially expressed yeast proteins. Competitive ionization makes the identification and quantification of yeast peptides cumbersome in samples with highly concentrated UPS spikes. Thus, the majority of false positives originate from technological artifacts rather than from flaws in the data analysis pipeline. We therefore recommend that researchers who are planning to use internal controls in their MS experiments avoid overspiking, as this can have detrimental effects on the quantification of the proteins of interest. Moreover, artifacts similar to those from spiked UPS proteins are bound to occur in certain experimental setups (e.g., undepleted blood plasma proteomic samples are known to be dominated by a few highly abundant proteins, and undepleted green tissue samples from plants will suffer from the omnipresence of RuBisCo). Our analysis showed that experimenters should interpret proteins further down the DE list with care. We therefore advise data analysts to use diagnostic plots based on all fold change estimates for assessing the quality of the FC estimates and for detecting potential artifacts. MA plots and boxplots were shown to be well suited for evaluating candidate DE proteins, to flag critical experimental conditions as well as flaws in the data analysis pipeline.

The myriad missing values in quantitative proteomic experiments present severe challenges to the data analysis. The standard MaxQuant pipeline therefore utilizes the match-between runs option to boost the number of peptide intensities that different samples have in common. Moreover, Perseus also incorporates imputation-based routines to deal with missing protein expression values. We showed that imputation is beneficial for detecting differentially expressed proteins with low abundance but performs suboptimally for moderately to highly abundant proteins. Perseus' standard imputation algorithm assumes that missing values originate from lower intensity values. Hence, the imputation can lead to a downward bias for more abundant proteins. Moreover, experimenters should also be aware that imputation comes at the cost of a decreased precision for the FC estimates.

In general, the current peptide-based methods are prone to overfitting and rely on protein-by-protein variance estimates. Hence, the development of robust methods that can borrow information across peptides and proteins would enable proteomics researchers to further deploy label-free quantitative proteomics.

In summary, we have shown that issues inherent to the methodology create challenges in quantitative proteomics, even in highly controlled and standardized samples such as the CPTAC ones. We then go on to show that downstream statistical data analysis approaches differ in their ability to cope with these different issues, and that the importance of these issues depends on the characteristics of the sample under study (e.g., the dominance of a few highly abundant proteins or large protein concentration ratio differences between two samples). Crucial, perhaps, is the fact that although peptide-based approaches fare better than summarization methods, no single method currently exists that can easily tackle all possible issues

DOI: 10.1021/pr501223t
J. Proteome Res. 2015, 14, 2457−2465

in quantitative proteomics data. Hence, more sophisticated data processing approaches that recognize these various issues are needed and can compensate for such issues more successfully across the board.

5. CONCLUSION

In this paper, we compared the performance of peptide-based linear models, mean and median summarization followed by limma analysis, and the standard MaxQuant/Perseus workflow for assessing differential abundance in label-free quantitative proteomics experiments. The evaluation of the performance was assessed using the CPTAC benchmark data set. Peptide-based models outperformed the competing data analysis pipelines in terms of sensitivity, specificity, accuracy, and precision. Modeling quantitative proteomics data at the peptide level allows for the correction of strong peptide-specific effects, which avoids the bias associated with summarization-based methods that aggregate different types of peptide intensities into a single value. Moreover, peptide-based models also improve the precision estimates by accounting for the different numbers of peptides that are identified in a sample. We have also shown that the FDR cutoffs used to determine the length of lists with significant differentially expressed (DE) proteins could become problematic in experimental setups with samples that are dominated by a few very abundant proteins. Technological artifacts might induce bias in the non-DE proteins, which can inflate the number of false positives that are returned at a particular FDR level. However, the ordering of the top DE proteins in the lists was shown to remain valid. We therefore advise proteomics researchers to be careful when spiking internal controls, to deplete the highly abundant proteins, and to use diagnostic plots for assessing the candidate DE proteins as well as the overall quality of the obtained fold change estimates. Finally, standard proteomics software provides experimenters with the ability to impute missing values. Perseus' imputation strategy was shown to be beneficial for detecting DE proteins with low abundance but at the cost of reduced precision as well as a suboptimal performance for moderately to highly abundant DE proteins. Hence, we advise proteomics data analysts to use imputation strategies with care.

■ ASSOCIATED CONTENT

Supporting Information

Figures showing the receiver operating characteristic (ROC) curves for the seven analysis methods in comparisons, F1 scores for the studied models, comparison of the bias terms for yeast and UPS proteins, comparisons of the root mean squared error for yeast and UPS proteins, boxplots showing $\log_2$ peptide intensities, MA plots for linear models, and ROC curves for normalization on the peptide and protein level with mean and median aggregation. Tables showing a general overview per spike-in conditions for UPS and yeast proteins, characteristics for various models and workflows, an explanation of the outlined characteristics, partial areas under the curves for a false positive rate, and relative partial areas under the curves for a false positive rate. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/pr501223t.

■ AUTHOR INFORMATION

Corresponding Author

*L.C. E-mail: Lieven.Clement@UGent.be. Tel: +32 9 264 49 04. Fax: +32 9 264 49 95.

Author Contributions

⊥L.G. and A.A. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Part of this research was supported by IAP research network "StUDyS" grant no. P7/06 of the Belgian government (Belgian Science Policy) and the Multidisciplinary Research Partnership "Bioinformatics: From Nucleotides to Networks" of Ghent University. A.A. is supported by the IWT SBO grant "INSPECTOR" (120025). L.J.E.G. is supported by the IWT SBO grant "Differential Proteomics at Peptide, Protein, and Module Level" (141573). L.M. acknowledges the PRIME-XS project, grant agreement no. 262067, funded by the European Union Seventh Framework Program.

■ ABBREVIATIONS

FDR, false discovery rate; MS, mass spectrometry; SILAC, stable isotope labeling of amino acids in cell culture; HPLC, high performance liquid chromatography; PSM, peptide-to-spectrum match; CPTAC, Clinical Proteomic Technology Assessment for Cancer Network; UPS, Universal Proteomics Standard 1; LFQ, label-free quantitation; FP, false positives; FN, false negatives; TP, true positives; TN, true negatives; PPV, positive predictive value; FC, fold change; ROC, receiver operating curve; pAUC, partial area under the curve; rpAUC, relative partial area under the curve; DE, differential expression

■ REFERENCES

(1) Vaudel, M.; Sickmann, A.; Martens, L. Peptide and protein quantification: A map of the minefield. *Proteomics* **2010**, *10* (4), 650–670.

(2) Bluemlein, K.; Ralser, M. Monitoring protein expression in whole-cell extracts by targeted label- and standard-free LC-MS/MS. *Nat. Protoc.* **2011**, *6* (6), 859–869.

(3) Cox, J.; Hein, M. Y.; Luber, C. A.; Paron, I.; Nagaraj, N.; Mann, M. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **2014**, *13* (9), 2513–2526.

(4) Liu, N. Q.; Dekker, L. J. M.; Stingl, C.; Güzel, C.; De Marchi, T.; Martens, J. W. M.; Foekens, J. A.; Luider, T. M.; Umar, A. Quantitative Proteomic Analysis of Microdissected Breast Cancer Tissues: Comparison of Label-Free and SILAC-based Quantification with Shotgun, Directed, and Targeted MS Approaches. *J. Proteome Res.* **2013**, *12* (10), 4627–4641.

(5) Mosley, A. L.; Sardiu, M. E.; Pattenden, S. G.; Workman, J. L.; Florens, L.; Washburn, M. P. Highly Reproducible Label Free Quantitative Proteomic Analysis of RNA Polymerase Complexes. *Mol. Cell. Proteomics* **2011**, DOI: 10.1074/mcp.M110.000687.

(6) Wang, G.; Wu, W. W.; Zeng, W.; Chou, C.-L.; Shen, R.-F. Label-Free Protein Quantification Using LC-Coupled Ion Trap or FT Mass Spectrometry: Reproducibility, Linearity, and Application with Complex Proteomes. *J. Proteome Res.* **2006**, *5* (5), 1214–1223.

(7) Silva, J. C.; Gorenstein, M. V.; Li, G.-Z.; Vissers, J. P. C.; Geromanos, S. J. Absolute Quantification of Proteins by LCMSE: A Virtue of Parallel ms Acquisition. *Mol. Cell. Proteomics* **2006**, *5* (1), 144–156.

DOI: 10.1021/pr501223t
J. Proteome Res. 2015, 14, 2457–2465

(8) Vaudel, M.; Sickmann, A.; Martens, L. Current methods for global proteome identification. *Expert Rev. Proteomics* **2012**, *9* (5), 519−532.
(9) Peng, M.; Taouatas, N.; Cappadona, S.; van Breukelen, B.; Mohammed, S.; Scholten, A.; Heck, A. J. R. Protease bias in absolute protein quantitation. *Nat. Methods* **2012**, *9* (6), 524−525.
(10) Schliekelman, P.; Liu, S. Quantifying the Effect of Competition for Detection between Coeluting Peptides on Detection Probabilities in Mass-Spectrometry-Based Proteomics. *J. Proteome Res.* **2013**, *13* (2), 348−361.
(11) Kumar, C.; Mann, M. Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Lett.* **2009**, *583* (11), 1703−1712.
(12) Liu, H.; Sadygov, R. G.; Yates, J. R. A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. *Anal. Chem.* **2004**, *76* (14), 4193−4201.
(13) Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* **2007**, *389* (4), 1017−31.
(14) Mueller, L. N.; Brusniak, M.-Y.; Mani, D. R.; Aebersold, R. An Assessment of Software Solutions for the Analysis of Mass Spectrometry Based Quantitative Proteomics Data. *J. Proteome Res.* **2008**, *7* (1), 51−61.
(15) Old, W. M.; Meyer-Arendt, K.; Aveline-Wolf, L.; Pierce, K. G.; Mendoza, A.; Sevinsky, J. R.; Resing, K. A.; Ahn, N. G. Comparison of Label-free Methods for Quantifying Human Proteins by Shotgun Proteomics. *Mol. Cell. Proteomics* **2005**, *4* (10), 1487−1502.
(16) Théron, L.; Gueugneau, M.; Coudy, C.; Viala, D.; Bijlsma, A.; Butler-Browne, G.; Maier, A.; Béchet, D.; Chambon, C. Label-free Quantitative Protein Profiling of vastus lateralis Muscle During Human Aging. *Mol. Cell. Proteomics* **2014**, *13* (1), 283−294.
(17) Hubner, N. C.; Bird, A. W.; Cox, J.; Splettstoesser, B.; Bandilla, P.; Poser, I.; Hyman, A.; Mann, M. Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *J. Cell Biol.* **2010**, *189* (4), 739−754.
(18) Clough, T.; Key, M.; Ott, I.; Ragg, S.; Schadow, G.; Vitek, O. Protein Quantification in Label-Free LC-MS Experiments. *J. Proteome Res.* **2009**, *8* (11), 5275−5284.
(19) Karpievitch, Y.; Stanley, J.; Taverner, T.; Huang, J.; Adkins, J. N.; Ansong, C.; Heffron, F.; Metz, T. O.; Qian, W.-J.; Yoon, H.; Smith, R. D.; Dabney, A. R. A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics* **2009**, *25* (16), 2028−2034.
(20) Paulovich, A. G.; Billheimer, D.; Ham, A.-J. L.; Vega-Montoto, L.; Rudnick, P. A.; Tabb, D. L.; Wang, P.; Blackman, R. K.; Bunk, D. M.; Cardasis, H. L.; Clauser, K. R.; Kinsinger, C. R.; Schilling, B.; Tegeler, T. J.; Variyath, A. M.; Wang, M.; Whiteaker, J. R.; Zimmerman, L. J.; Fenyo, D.; Carr, S. A.; Fisher, S. J.; Gibson, B. W.; Mesri, M.; Neubert, T. A.; Regnier, F. E.; Rodriguez, H.; Spiegelman, C.; Stein, S. E.; Tempst, P.; Liebler, D. C. Interlaboratory Study Characterizing a Yeast Performance Standard for Benchmarking LC-MS Platform Performance. *Mol. Cell. Proteomics* **2010**, *9* (2), 242−254.
(21) Cox, J.; Hein, M. Y.; Luber, C. A.; Paron, I.; Nagaraj, N.; Mann, M. Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Mol. Cell. Proteomics* **2014**, *13* (9), 2513−2526.
(22) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26* (12), 1367−1372.
(23) Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc.: Series B* **1995**, *57* (1), 289−300.
(24) Smyth, G. K., Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl. Genet. Mol. Biol.* **2004**, *3*, Article 3.
(25) Satterthwaite, F. E. An approximate distribution of estimates of variance components. *Biometrics* **1946**, *2* (6), 110−4.
(26) Rodriguez, J.; Gupta, N.; Smith, R. D.; Pevzner, P. A. Does Trypsin Cut Before Proline? *J. Proteome Res.* **2008**, *7* (1), 300−305.
(27) Abaye, D. A.; Pullen, F. S.; Nielsen, B. V. Peptide polarity and the position of arginine as sources of selectivity during positive electrospray ionisation mass spectrometry. *Rapid Commun. Mass Spectrom.* **2011**, *25* (23), 3597−3608.

DOI: 10.1021/pr501223t
J. Proteome Res. 2015, 14, 2457−2465