Optimization of statistical methods impact on quantitative proteomics data

Pursiheimo, A.; Vehmas, A.P.; Afzal, S.; Suomi, T.; Chand, T.; Strauss, L.; Poutanen, M.; Rokka, A.; Corthals, G.L.; Elo, L.L.

DOI
10.1021/acs.jproteome.5b00183

Publication date
2015

Document Version
Final published version

Published in
Journal of Proteome Research

Journal of proteome research

Article

pubs.acs.org/jpr

Optimization of Statistical Methods Impact on Quantitative Proteomics Data

Anna Pursiheimo,[†,‡,#] Anni P. Vehmas,[†,#] Saira Afzal,[†,#] Tomi Suomi,[†,§] Thaman Chand,[†] Leena Strauss,[∥] Matti Poutanen,[∥] Anne Rokka,[†] Garry L. Corthals,[*,†,⊥] and Laura L. Elo[*,†,‡]

[†]Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, Tykistökatu 6, FI-20520 Turku, Finland

[‡]Department of Mathematics and Statistics, University of Turku, FI-20014 Turku, Finland

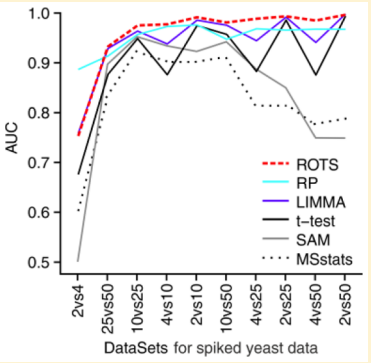
[§]Department of Information Technology, University of Turku, FI-20014 Turku, Finland

[∥]Department of Physiology and Turku Center for Disease Modeling, Institute of Biomedicine, University of Turku, Kiinamyllynkatu 10, FI-20520 Turku, Finland

[⊥]Van't Hoff Institute for Molecular Sciences, University of Amsterdam, P.O. Box 94157, 1090 GD Amsterdam, The Netherlands

Supporting Information

ABSTRACT: As tools for quantitative label-free mass spectrometry (MS) rapidly develop, a consensus about the best practices is not apparent. In the work described here we compared popular statistical methods for detecting differential protein expression from quantitative MS data using both controlled experiments with known quantitative differences for specific proteins used as standards as well as "real" experiments where differences in protein abundance are not known a priori. Our results suggest that data-driven reproducibility-optimization can consistently produce reliable differential expression rankings for label-free proteome tools and are straightforward in their application.



KEYWORDS: *proteomics, label-free mass spectrometry, quantitative analysis, statistical methods, reproducibility, ROTS*

INTRODUCTION

Quantitative label-free mass spectrometry (MS) is an increasingly used technique to determine the relative quantity of proteins in biological samples. A typical label-free quantitative experiment will contain the sequential steps starting with enzymatic digestion of a complex mixture of proteins, peptide separation by high-resolution chromatography, electrospray ionization, and finally mass analysis in a mass spectrometer. The most frequently used MS strategy for quantitative label-free proteomics typically starts with data-dependent MS analysis, where precursor ions (MS) are detected in a survey scan and selected automatically using a simple heuristic for subsequent fragmentation (MS/MS, tandem mass spectrometry) to reveal amino acid sequence information. The detected MS and MS/MS data are then compared against protein sequence databases to identify the sequence of each precursor ion[1,2] and subsequently quantify the identified proteins.

Currently there are two label-free quantification methods that use either MS or MS/MS level information for quantification. The MS methods quantify the precursor ion signals derived from peptides independent of MS/MS information, whereas the MS/MS methods measure the number of fragments ion spectra for each peptide, also termed spectral counting. Recently we evaluated the performance of MS/MS level information.[3] In this manuscript, we evaluate the performance of MS level information, which measures peptides directly and can therefore avoid the stochastic sampling of MS/MS methods.[4]

The analysis workflow for MS-level quantification consists of multiple steps, including an alignment of peptide feature maps of all analyzed samples in a study, finding the common features across the samples, data normalization, and finally statistical analysis of the data.[5,6] The peptide signals (features) used for quantification in the MS-level label-free quantification are defined by their m/z value, retention time, and intensity.[4] Quantification is based on the peak volumes, which is possible, as the intensity of electrospray ionization is linearly correlated with peptide ion abundance in the typical concentration range of peptides.[7,8] The identified peptide features with defined peak volumes are typically summarized into corresponding protein-level values before further analysis to draw biological conclusions from the results.[9,10]

Although label-free methods are scalable for large sample sets, flexible, easy to use, and inexpensive when compared with

Received: October 31, 2014
Published: August 31, 2015

DOI: 10.1021/acs.jproteome.5b00183
J. Proteome Res. 2015, 14, 4118−4126

stable isotope methods, they also present challenges to researchers due to data handling capacities and data analysis workflows. The number of features produced by a protein can vary substantially; the features from different proteins may overlap due to coelution during LC elution, and peptides can be incorrectly annotated or show poor quantitation values (poor or irreproducible signal-to-noise).[11] Although, there are many computational approaches designed for the discrete steps of the data analysis, such as alignment, feature finding, and normalization, they typically offer only limited possibilities for statistical evaluation of differential protein expression between samples. For example, popular packages such as MaxQuant[12] or Progenesis (Nonlinear Dynamics, U.K.) offer only basic statistical tests such as *t* test and ANOVA (Analysis of variance), whereas the OpenMS/TOPP[13] (the OpenMS proteomics pipeline) system does not provide any statistical tools. Because of the large demand, widespread use, but lack of statistical approaches, there is a clear need for flexible, reliable, and easy to use statistical workflow for the output tables of different MS-level quantification tools.

The analysis of large-scale quantitative label-free data is complicated because MS-based proteome data are statistically challenging due to the fact that they typically consist of data matrices with relatively low numbers of samples (few to tens typically) but high numbers of variables[14] (thousands to tens of thousands). In gene expression studies several efficient methods have been developed for analyzing such high-dimensional lopsided data;[15−19] however, only a limited selection of these tools is currently in use for proteome data mining quantitatively. For instance, the Student's *t* test is still a widely used approach in proteome studies to determine differences in protein abundances between sample groups, although its limitations are well known from gene expression studies,[15,16] especially with small sample sizes.

As the proteomics field rapidly evolves, consensus about the best practices is not apparent, which makes the choice of "the best practice" difficult. In the present study we compare statistical methods to find out if there are differences between the methods in their ability to detect differentially expressed proteins and optimize statistical testing for reliable analysis of label-free proteome MS data. In particular, we discuss the systematic evaluation of sensitivity and specificity of commonly used statistical methods on a diverse range of data, including both controlled experiments where the quantitative differences for specific proteins were known beforehand and "real" experiments where differences in protein abundance were not known a priori.

MATERIALS AND METHODS

UPS1 (Universal Proteomics Standard) in Yeast Whole Cell Lysate

Different concentrations of the UPS1 (equimolar amounts of 48 human proteins, Sigma-Aldrich) were mixed with trypsin-digested soluble yeast proteins. UPS1 final concentrations of 2, 4, 10, 25, and 50 fmol/μL were made. The amount of total yeast peptides per injection was 100 ng. Three replicates per (UPS1) concentration were analyzed by MS, as described in our previous work on spectral counting.[3] In brief, the UPS1 peptide mixture was dissolved in 6.0 M urea/25 mM ammonium bicarbonate buffer, reduced with 200 mM DTT, alkylated with 200 mM iodoacetamide, and digested with trypsin overnight at 37 °C. After digestion, the peptide mixture was desalted using C18 pipet tips (OMIX, Agilent Technologies) according to manufacturer's instructions, evaporated to dryness, and resuspended in 0.1% formic acid. Digested UPS1 mixture was spiked into a yeast proteome digest, provided by VTT Technical Research Centre of Finland.

MS/MS was performed on peptides eluting from a nanoflow HPLC (high-performance liquid chromatography) system (Easy-nLCII, Thermo Fisher Scientific) coupled to the LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) equipped with a nanoelectrospray ionization source. For this process, 100 ng of peptides was loaded on an in-house packed C18 trap column (2.5 cm long, 75 μm inner diameter, Magic AQ C18 resin -5 μm/200 Å, Bruker-Michrom, Billerica, MA), and separated by a 15 cm long analytical column packed with the same C18 particles (75 μm inner diameter). A 110 min long linear gradient from 98% solvent A (98% H_2O, 2% ACN and 0.2% HCOOH) to 35% solvent B (95% ACN, 5% H_2O and 0.2% HCOOH) with a flow rate of 300 nL/min was used for peptide elution. The 20 most intense doubly or triply charged precursor ions were selected for fragmentation (MS/MS) by collision-induced dissociation (CID). Three runs per UPS1 concentration were analyzed.

The collective database search for the 15 spectrum files was performed in Proteome Discoverer (version 1.3.0.339, Thermo Fisher Scientific) using the default peak-picking scheme. The samples were analyzed using Mascot and Sequest algorithms against UniProtKB/Swiss-Prot yeast database (accessed 110615), with UPS1 and cRAP (the common Repository of Adventitious Proteins, accessed 110403) protein sequences appended. We searched for peptides formed by trypsin digestion, where two missed cleavage sites were allowed, methionine oxidation was selected as a dynamic, cysteine carbamidomethylation was selected as a fixed modification, accepted precursor mass tolerance was set to 5 ppm, fragment mass tolerance was set to 0.5 Da, and false discovery rate of the Percolator decoy database search[20,21] was set to 1%. Protein hits with at least two peptide spectral matches and at least one unique peptide were exported to Progenesis 4.0 software (Nonlinear Dynamics, Newcastle upon Tyne, U.K.).

Spectral data from the total of 15 MS runs were imported to Progenesis for quantification. The analysis area of the gradient was set to 0−100 min. The feature detection in the aggregate LC−MS map of the 15 runs was performed by automatic peak picking in default sensitivity mode, maximum charge of precursor was set to 3+, and the retention time window limit to 12 s. After the removal of peptides originating from the cRAP contaminants database, 1442 proteins were identified and quantified in Progenesis with at least one peptide identified in one MS analysis. For the comparison of the statistical methods, we considered those 944 proteins that were quantified using at least two peptides, including 46 of the 48 spiked UPS1 proteins. One UPS1 protein did not fulfill the stringent criteria used for the identification in the database search, and one was quantified using only one peptide. The data were normalized using median scaling.[22] Supplementary Figure 1 shows the profiles of the 46 spiked UPS1 protein intensities for the different concentrations.

Mouse Data

The mouse data contained measurements of liver samples excised from 3-month-old transgenic male mice overexpressing human P450 aromatase (AROM+, n = 5) and their age-matched wild type controls (WT, n = 7). The aromatase

DOI: 10.1021/acs.jproteome.5b00183
J. Proteome Res. 2015, 14, 4118−4126

enzyme converts androgens (male sex hormones) to estrogens (female sex hormones). AROM+ male mice exhibit various phenotypic alterations, whereas females do not present with any obvious defects.[23−25] The mice were housed according to international guidelines on the care and use of laboratory animals at Central Animal Laboratory, University of Turku. All animal handling fulfilled the requirements of Finnish Animal Ethics Committee, the Institutional animal care policies of the University of Turku (Turku, Finland), and NIH Guide on animal experimentation.

A 0.25 × 0.25 cm piece of liver was homogenized into 100 mM K-phosphate/150 mM KCL buffer (pH 7.4) with protease inhibitors (Complete Mini EDTA-free Protease Inhibitor Cocktail, Roche) using TissueLyser (Qiagen, Austin, TX). The supernatant was subjected to acetone precipitation and kept −20 °C overnight. The pellet was dried at room temperature and reconstituted in 6.0 M urea/25 mM ammonium bicarbonate buffer, after which the proteins were reduced by 200 mM DTT. Alkylation was accomplished by 200 mM iodoacetamide. The digestion was completed overnight at 37 °C by the addition of trypsin. Sample cleanup was performed after digestion by Empore C18-SD extraction disk cartridges (3M) according to manufacturer's instructions.

The liver peptides were analyzed similarly to the yeast samples with few exceptions: The amount of liver peptides user per injection was 200 ng on LTQ Orbitrap Velos Pro mass spectrometer (Thermo Fisher Scientific, Waltham, MA) coupled to EASY-nLC nanoflow liquid chromatography system (Thermo Fisher Scientific). Peptides were separated on an in-house built C18 analytical column by a 95 min LC gradient as previously described but were subjected to data-dependent analysis with the top 15 ions selected for fragmentation by CID. Each liver peptide sample was analyzed once, and a mixed sample consisting of all liver tissue samples was injected at regular intervals for standardization.

The 12 mouse liver peptide samples and the first injection of the mixed sample were analyzed in DDA (data-dependent acquisition) mode, but to minimize the redundancy in the precursor ion identifications for the second, third, and fourth injection of the mixed sample, we chose a directed proteomics approach. In this procedure, only those precursor ions that were not identified in previous injections of the sample are selected for analysis using an inclusion list.[4,26] The inclusion lists based on unidentified ions in the mixed sample were made as follows: The spectrum files from injections 1−3 were imported to Progenesis 4.0 (Nonlinear Dynamics), where the areas of the precursor ions were determined by an automatic feature detection algorithm with the strictest preset peak picking threshold. Double and triple charged precursor ions that were eluted over a period of at least 6 s and were not identified in Mascot by two or more peptides and mass error lower than 5 ppm were exported to an inclusion list. Additionally, when transferred to the list, the retention time window for each precursor was expanded by 1 min to compensate for possible variations in retention time.

The collective database search for the mouse liver spectrum files was performed in Proteome Discoverer against the UniProtKB/Swiss-Prot mouse database (16 686 sequences, accessed 130215) using the same parameters as with the spiked yeast data (see previous section). For the construction of inclusion lists for the mixed peptide samples 1−4, only Mascot (Matrix Science, London, U.K.) was used. Spectral data from the 16 MS runs were imported to Progenesis 4.0 and quantified similarly as the spiked yeast data. After removing the peptide-feature matches that had fewer than two matches or had precursor mass tolerance higher than 5 ppm and the peptides originating from the contaminants database, a total of 1499 proteins were identified and quantified with at least one peptide. For the comparison of the statistical methods, we considered those 974 proteins that were quantified using at least two peptides. The data were normalized using median scaling.[22]

Data Availability

The MS proteomics data are deposited into the ProteomeXchange Consortium[27] via the PRIDE partner repository with the data set identifiers PXD002099 (yeast data) and PXD002025 (mouse data).

The R-scripts, used with the different statistical methods and the intermediate results from the yeast data are available at our Web site (http://www.btk.fi/research/research-groups/elo).

Statistical Methods

The statistical methods evaluated in this study included the ordinary *t* test, Significance Analysis of Microarrays, SAM,[15] Linear Models for Microarray Data, LIMMA,[16] Rank Product, RP,[17] and Reproducibility-Optimized Test Statistic, ROTS.[18,19] The null hypothesis for the statistical tests was that there is no difference between the two sample group means of a protein, and the alternative hypothesis was that the sample group means of a protein differ from each other. SAM, LIMMA, and ROTS are modified versions of the ordinary *t* test, whereas RP is based on ranking the items to be analyzed. The analyses were performed using the R programming language (www.r-project.org), version 3.0.1. The methods were applied using the default parameter settings. Proteins with false discovery rate (FDR) below 0.05 were considered as significant.

The ordinary *t* test compares the means of two groups of samples. With an assumption of equal group variances, the *t*-statistic is calculated as

$$t(i) = \frac{\bar{x}_1(i) - \bar{x}_2(i)}{s(i)}$$

where $\bar{x}_j(i)$ is the average abundance level of protein i in sample group j and $s(i)$ is the pooled standard error for the expression of the protein i estimated as

$$s(i) = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{(n_1 - 1)s_1^2(i) + (n_2 - 1)s_2^2(i)}{n_1 + n_2 - 2}}$$

Here $s_j^2(i)$ is the variance of the abundance level of protein i in sample group j and n_j denotes the number of samples in group j. We used the Benjamini−Hochberg method to determine FDR.[28]

The SAM statistic,[15] like the ordinary *t* test, gives a score to each protein, based on the difference in its average abundance between the sample groups, relative to the standard error. The SAM statistic is calculated as

$$d(i) = \frac{\bar{x}_1(i) - \bar{x}_2(i)}{s(i) + s_0}$$

where the average abundance level $\bar{x}_j(i)$ and the standard error $s(i)$ are defined as in ordinary *t* test. To make the coefficient of variation of $d(i)$ across all proteins approximately constant, a small positive value s_0 (a percentile of the standard error values)

DOI: 10.1021/acs.jproteome.5b00183
J. Proteome Res. 2015, 14, 4118−4126

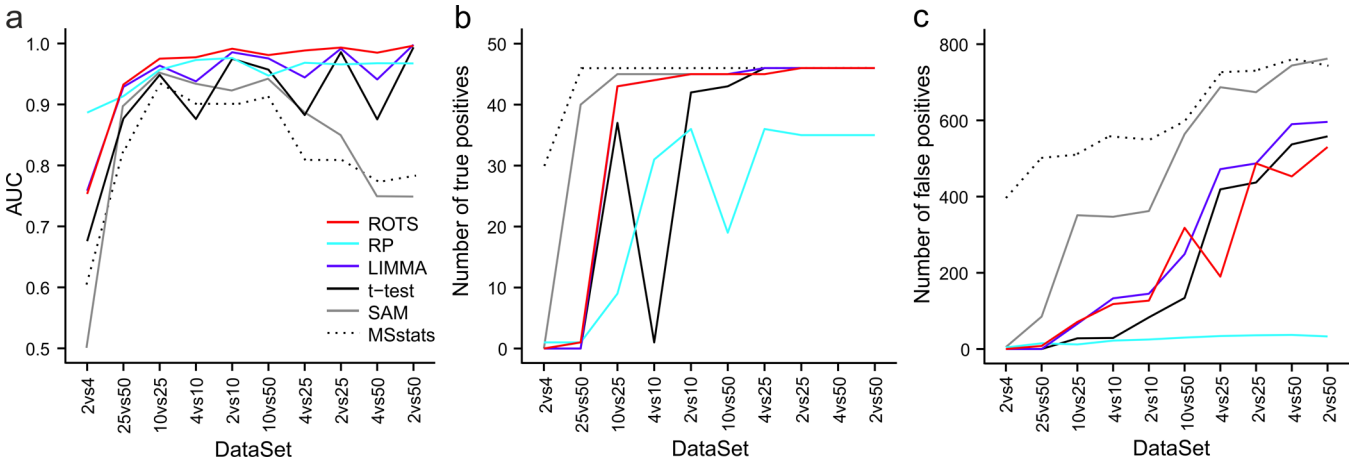


Figure 1. Performance of the statistical methods in the spiked yeast data. (a) Area under the receiver operating characteristic (ROC) curves (AUC) for the different statistical methods (the different colors) across all of the possible pairwise comparisons (x axis). (b) Number of UPS1 proteins detected (true positives) at false discovery rate FDR ≤ 0.05. (c) Number of yeast proteins detected (false positives) at false discovery rate FDR ≤ 0.05. The data sets on the x axis were ordered by the average number of detections over all the methods (both true and false positives included). All x-axis values are femtomolar concentrations of spiked UPS1 standard; for example, 2 vs 4 fmol spiked UPS1.

is added to the denominator. For the FDR estimation, a permutation procedure is used.

LIMMA[16] fits a linear model to determine differential expression. A linear model $E[x(i)] = X\beta(i)$, where $x(i)$ is the expression data of protein i, is fitted to obtain the estimators for the coefficient $\tilde{\beta}(i)$ and for the residual variance $s^2(i)$ of the linear model. A moderated t test is used to find the differentially expressed proteins and is computed as

$$\tilde{t}(i) = \frac{\tilde{\beta}(i)}{u(i)\tilde{s}(i)}$$

where $u(i)$ is the unscaled standard deviation and $\tilde{s}^2(i)$ is the posterior residual variance of the linear model, defined as

$$\tilde{s}^2(i) = \frac{f_0 s_0^2 + f_i s_i^2}{f_0 + f_i}$$

Here f_0, and f_i are the prior and the residual degrees of freedom, respectively, and s_0^2 and s_i^2 are the prior and residual variance for the linear model, respectively. For the FDR estimation the Benjamini−Hochberg method[28] is used.

The rank product[17] method ranks the proteins according to their fold changes. The combined ranking score for protein i across replicates is then calculated as a rank product, separately for both directions of change, up and down

$$\mathrm{RP}^{\mathrm{up}}(i) = \prod_{l=1}^{n} \left(\frac{r_l^{\mathrm{up}}(i)}{k_l} \right)$$

$$\mathrm{RP}^{\mathrm{down}}(i) = \prod_{l=1}^{n} \left(\frac{r_l^{\mathrm{down}}(i)}{k_l} \right)$$

Here $r_l^{\mathrm{up}}(i)$ and $r_l^{\mathrm{down}}(i)$ are the ranks of protein i in the ordered protein lists of length k_l sorted in decreasing or increasing order for replicate l (out of the total of n replicates). For the most strongly upregulated protein $r_l^{\mathrm{up}}(i) = 1$ and, similarly, for the most strongly downregulated protein $r_l^{\mathrm{down}}(i) = 1$. FDR is estimated using a permutation-based approach.

ROTS[18,19] is a reproducibility-optimization procedure that selects a protein ranking statistic for a given data by maximizing the reproducibility of the detections among a family of modified t-statistics

$$d_\alpha(i) = \frac{\bar{x}_1(i) - \bar{x}_2(i)}{\alpha_1 s(i) + \alpha_0}, \quad \alpha_0 \geq 0, \alpha_1 \in \{0, 1\}$$

Similarly as above, $\bar{x}_j(i)$ and $s(i)$ are the average abundance level of protein i in sample group j and the pooled standard error for the expression of the protein i, respectively. The parameters α_0 and α_1 specify the statistic, and the optimal combination of these parameters is determined by reproducibility maximization. Reproducibility R is defined as the average overlap of the top-ranked proteins across bootstrap data sets (sampling with replacements), relative to the null reproducibility R^0 in randomized data sets (random permutation of the samples).[18] More specifically, we maximize the reproducibility Z-score $Z_{k,\alpha} = (R_{k,\alpha} - \mathrm{R}^0{}_{k,\alpha})/s_{k,\alpha}$ over varying top list sizes k and parameter combinations $\alpha = (\alpha_0, \alpha_1)$. Here $R_{k,\alpha}$ and $R^0{}_{k,\alpha}$ are the observed and random reproducibility at top list size k and parameter α and $s_{k,\alpha}$ is the standard deviation of the bootstrap distribution. A dense grid of values is browsed through to find the optimal parameter combination. The final ROTS output is then calculated from the original data using the optimized parameters that maximize the reproducibility Z score. The special cases of ROTS include the ordinary t-statistic ($\alpha_0 = 0$, $\alpha_1 = 1$), the signal log-ratio ($\alpha_0 = 1$, $\alpha_1 = 0$), as well as the SAM statistic ($\alpha_1 = 1$, α_0 is a percentile of the standard error values). Importantly, the ROTS statistic does not rely on any parametric assumptions. The FDR is estimated by randomly permuting the sample labels.

For the statistical approaches to be directly comparable, we wanted to focus on methods that use exactly the same input data, which is in our case protein level data. Additionally, however, we also tested the widely used MSstats tool, which uses peptide level data as input data.[9,29] MSstats is based on a flexible family of linear mixed models and relies on the functionalities of fixed effects model and mixed effects model.

DOI: 10.1021/acs.jproteome.5b00183
J. Proteome Res. 2015, 14, 4118−4126

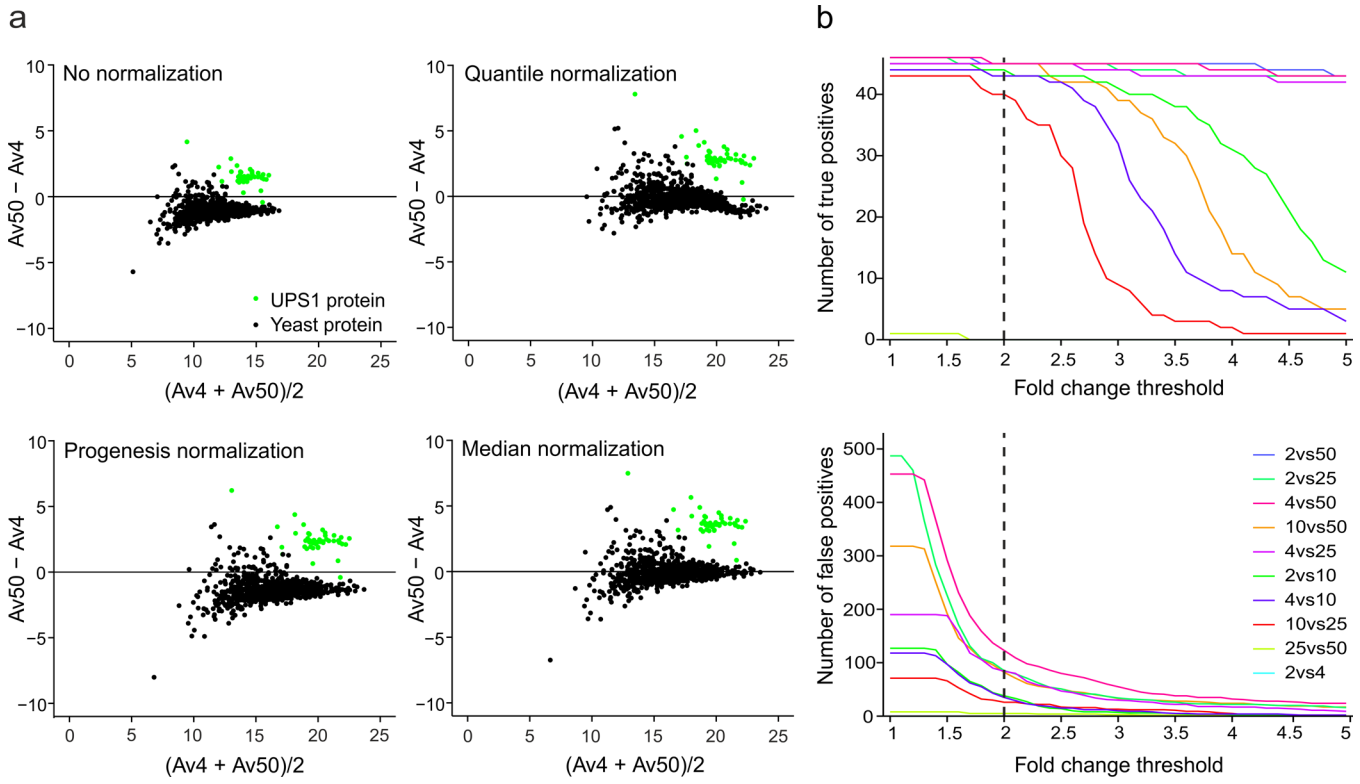


Figure 2. Effect of normalization and fold change filtering on the detected differentially expressed proteins in the UPS1 spiked yeast data. (a) Logarithmic fold change (y axis) between the two groups under comparison versus the average abundance across the groups using four different approaches to normalization: no normalization, quantile normalization, Progenesis normalization, and median scaling normalization. The data set 4vs50 is shown as a representative example. The other data sets showed similar patterns. (b) Number of true and false positives (y axis in the upper and lower panels, respectively) as a function of fold change threshold (x axis) when using ROTS at false discovery rate FDR $\leq$ 0.05.

RESULTS

Combined UPS1 and Yeast Data

We first evaluated the performance of the statistical methods in controlled LC−MS/MS experiments where 48 human proteins (UPS1) were added with different concentrations (2, 4, 10, 25, and 50 fmol/μL) into a whole yeast cell protein digest.[3] Three samples were analyzed at each concentration, and differential protein expression was determined in each possible pairwise comparison between two different concentrations (denoted by 2vs4, 2vs10, etc.). The performance of the methods was assessed in terms of sensitivity and specificity of the detections. The detected UPS1 proteins were considered as true positives, whereas all of the other proteins detected were considered as false positives. In total, 46 UPS1 proteins passed the stringent identification and quantification criteria in all of the samples analyzed and were included in the evaluations.

Figure 1a summarizes the performance of the different statistical methods in terms of the areas under the receiver operating characteristic (ROC) curves (AUC). The AUC values with ROTS, RP, and LIMMA were steadily between 0.9 and 1, with the exception of the data set 2vs4 that gave overall very low AUC values with the different methods analyzed (Supplementary Table 1). The AUC values with the ordinary t test varied markedly between the different data sets; half of the values were below 0.9 and half were above. SAM exhibited considerably lower AUC values than the other methods, with only three of the values above 0.9. The original ROC curves are provided as Supplementary Figure 3. Interestingly, the observed reproducibility Z-scores from the ROTS procedure showed a high correlation with the observed AUC values (Spearman correlation 0.89, $p < 0.01$; Supplementary Table 2). In particular, the ROTS reproducibility Z-score of 1.66 in the 2vs4 data set suggested compromised reliability of the corresponding results.

The FDR level of 0.05 is a commonly used significance threshold applied in proteome studies. By using this threshold, the sensitivity, that is, the proportion of true positive spiked UPS1 proteins that are correctly identified as differentially expressed (also called true positive rate), of most of the statistical methods was high (Figure 1b). The median sensitivity was 97.8% with the modified t-statistics ROTS, LIMMA, and SAM, and 91.3% with the ordinary t test, whereas RP provided the poorest median sensitivity of 71.7% (Supplementary Table 1). In each of the data sets, RP missed at least 10 of the spiked UPS1 proteins, whereas the other methods typically detected nearly all of them. In two of the data sets (2vs4 and 25vs50), where the UPS1 protein fold-change was lowest, however, all of the methods showed low sensitivity.

Although the ROC analysis supported the ability of most of the methods to produce reliable rankings of the differentially expressed proteins, the use of the FDR threshold of 0.05 revealed relatively low specificity (the proportion of true negatives among the proteins regarded as significant, also called true negative rate) with relatively many false positives (Figure 1c and Supplementary Table 1). The only exception was RP, which, on the contrary, showed the worst sensitivity. In particular, SAM produced large numbers of false positives. The two data sets that gave the lowest numbers of true positives

DOI: 10.1021/acs.jproteome.5b00183
J. Proteome Res. 2015, 14, 4118−4126

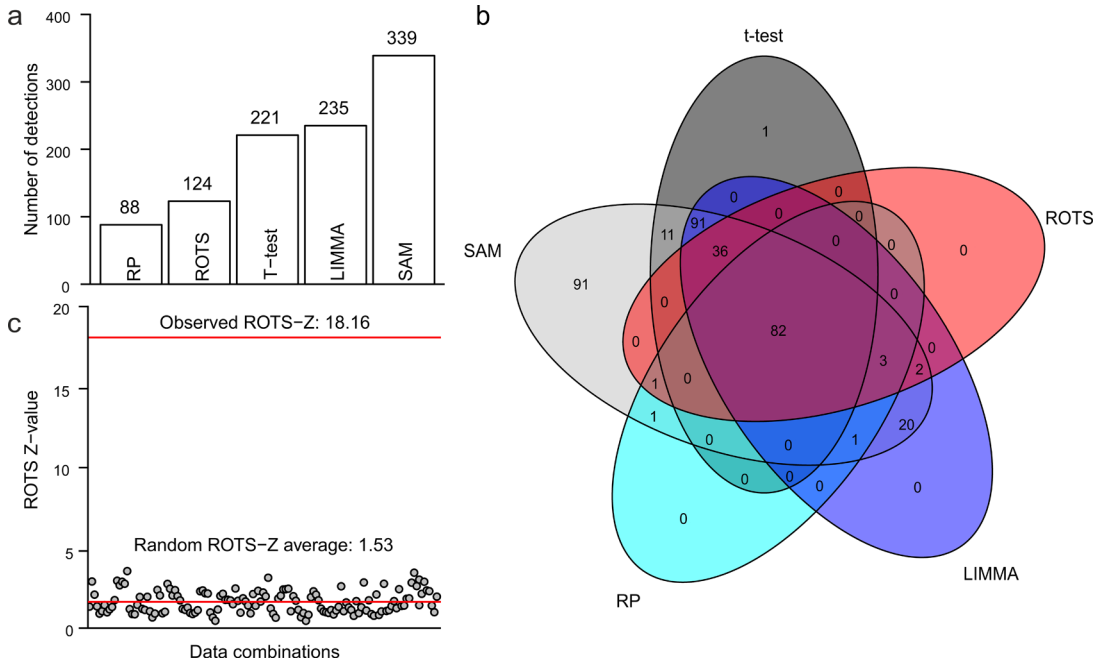


Figure 3. Performance of the statistical methods in the mouse data. (a) Number of detected differentially expressed proteins with each statistical method at false discovery rate FDR ≤ 0.05. (b) Venn diagram of the detected differentially expressed proteins illustrating the overlaps between the statistical methods. (c) Comparison of the ROTS reproducibility *Z* score observed in the actual comparison with those observed in the 140 random comparisons within the WT mouse group where no differentially expressed proteins were expected.

(2vs4 and 25vs50) produced also the lowest numbers of false positives.

In addition to the statistical methods that used protein level data as input, we also tested the performance of the MSstats package, which uses peptide level data as input data. MSstats performed surprisingly poorly in our spiked data. With an exception of the data set 2vs4, MSstats detected all of the UPS1 proteins (Figure 1b and Supplementary Table 1), but the number of false positives was very large (Figure 1c and Supplementary Table 1) compared with the other methods. Also, the AUC values were in almost every data set the poorest ones (Figure 1a and Supplementary Table 1).

In general, several background yeast proteins were observed to display differential expression. This may be due to one or more of a number of factors inherent to the LC−MS or ESI (electrospray ionization) process for the UPS1 proteins/yeast sample. The false positives may also be quantitatively unstable peptides containing missed cleavages or modifications representing 10−21% of the identified peptides. The percentage of these peptides was rising steadily with increasing amounts of UPS being ∼10% in 2 fmol and ∼20% in 50 fmol. Furthermore, it was observed that as the UPS1 concentration increased, fewer yeast proteins were detected and the algorithms interpreted the results as significant differences, even though these were not de facto differences (Supplementary Figure 2). Such systematic differences were particularly observed when we analyzed raw, un-normalized data or the data normalized with the method available in the Progenesis software; however, the median scaling reduced the number of false positives (Figure 2a, Supplementary Table 3). We also note that the widely used quantile normalization produced some spurious tailing in our data (Figure 2a).

We also tested the common practice of setting a threshold for the minimum fold-change required when determining the lists of differentially expressed proteins to reduce the number of false positives. In the yeast data, the investigation of the number of true and false positives as a function of the fold-change threshold suggested that setting the threshold to two significantly reduced the number of false positives without markedly affecting the number of true positives (Figure 2b).

Finally, we assessed how the number of peptides used for quantification affected the performance of the methods in the UPS1/yeast data. When the minimum number of quantified peptides per protein was two or three, the number of detected UPS1 proteins was 46, while the number decreased to 44 when four or five peptides were required per protein. Notably, the AUC values and the number of true positives maintained approximately at the same level with all peptide-based filtering criteria used (Supplementary Figures 4 and 5, respectively). In contrast, the number of detected false positives typically decreased with all of the methods when more stringent peptide-based filters were applied (Supplementary Figure 6).

Mouse Data

In the mouse data the number of true positive proteins was unknown, which is typical for experimental studies. Investigating the detection rates revealed that ROTS and RP detected the lowest numbers of differentially expressed proteins at FDR ≤ 0.05, whereas SAM also here (as with UPS1/yeast) gave a much higher number of detections than the other methods (Figure 3a). In total, the methods detected 82 common proteins to be differentially expressed (Figure 3b). Notably, nearly all of the proteins detected by ROTS (120 out of 124) were detected also by at least three other methods, supporting their potential relevance. SAM, on the contrary, detected 91 additional proteins that were not detected with any of the other methods tested.

To further evaluate the reliability of the differential expression detections, we used the seven replicates of the wild-type mouse samples. More specifically, we ran each statistical method on all possible 3 + 3 combinations of the

DOI: 10.1021/acs.jproteome.5b00183
J. Proteome Res. 2015, 14, 4118−4126

samples within the WT mouse group (140 different data combinations). Because no differences were expected to be detected in these analyses, the approach provides an estimate for the number of false positives in the absence of ground truth. In these evaluations, the average number of differentially expressed proteins was below three, indicating a low number of false positives. The only exception was SAM, which incorrectly detected on average 123 proteins (median 9, minimum 0, maximum 537).

Overall, the results in the mouse data were in line with those in the UPS1/yeast data. RP tended to miss many true positives, whereas SAM reported the largest number of differentially expressed proteins (Figure 3a) but showed also the highest rate of false positives. The observed reproducibility *Z*-score of 18.16 from ROTS further supported the reliability of the ROTS detections. In the randomized comparisons, the value was over 10 times lower, being 1.53 on average (Figure 3c).

DISCUSSION

In this study we have performed a systematic comparison of commonly used statistical methods to determine differential protein expression from quantitative label-free MS data using both known change differences of UPS-1 in a yeast lysate as well as a complex sample (from mice), where the de facto abundance differences are not known. Our evaluations demonstrated the utility of the reproducibility-optimization procedure ROTS to reliably detect fold differences in complex proteome samples. In contrast, the ordinary *t* test often performed poorly with small sample sizes; SAM detected many false positives in our data; whereas RP failed to detect several true positives as significant.

Even though the work specifically aims to compare a selection of statistical analysis tools, there are a few aspects worth mentioning that can affect the reliability of the results but are integral to label-free quantification. First, even though the MS precursor ion-based label-free quantification method avoids problems related to stochastic sampling that occurs with tandem MS, during data-dependent analysis more abundant precursor ions are chosen for fragmentation. Even though the mass spectrometer can be guided toward the analysis of lower abundant peptides through, for example, inclusion lists, only identified features are quantified and therefore the lower abundant peptides are not efficiently quantified. Second, the analysis of complex mixtures by LC−MS/MS presents challenges in terms of separation capacity. As the quantification is performed at the precursor ion (MS1) level, isobaric ion species can elute and ionize simultaneously and result in an overlapping feature pattern in the LC−MS map of the quantification software. These features are challenging to the quantification software and are often excluded from the analysis producing false negative results.

In general, all of the methods were able to detect true positives reliably. The lowest sensitivity at the widely used FDR threshold of 0.05 was often obtained with RP. The ordinary *t* test provided both good and poor results. Originally, the *t* test was developed for larger sample sizes and it is no surprise that it does not guarantee reliable results with small sample sizes.[15] The SAM method performed worst with our data, as it was incapable of producing reliable rankings. To reduce the number of false positives in the SAM procedure, one could manually select the delta threshold; however, this causes problems in studies where the number of comparisons is large, as presented here.

In the UPS1/yeast data, all methods generated many false positives at the FDR level of 0.05, which is likely related to the spiked experimental setup (Supplementary Figure 2). In the mouse data, the estimated proportions of false positives were estimated to be low, except for SAM that showed a high rate of false positives; however, because the mouse data, like typical proteomics studies, measures signals from real experiments, it is not known whether the larger number of detections truly comes from false positives by SAM or false negatives by the other methods. The spike-in data from this report supports the first option. Overall, setting an additional cutoff to the fold-changes can help in reducing the number of false positives.[30] Furthermore, the common practice of requiring at least two peptides per protein for identification and more than two for quantification reduced the number of false positives in our comparisons.

The proteins in our data were quantified using the Progenesis software. Although Progenesis provides a normalization procedure (http://www.nonlinear.com/progenesis/qi-for-proteomics/v1.0/faq/), in the present study, it did not perform as well as the simple median scaling normalization. This was reflected as better sensitivity and specificity values observed in the median normalized UPS1/yeast data compared with the corresponding Progenesis normalized data. To our surprise, the widely used quantile normalization created very peculiar tails to the scatterplots of the spiked yeast data (example scatterplot from the comparison 4vs50 in Figure 2a), suggesting some artificial biases caused by the normalization.

Currently in proteome research the selection of a statistical method does not exclusively depend on the reliability of the results. Users without strong statistical, mathematical, or computational skills tend to prefer a variety of methods based on ease of use and the provision of clear examples and instructions in their software manuals. For instance, preprocessing of the data to a suitable format for the statistical analysis or postprocessing of the output results for interpretation can cause confusion. Among the methods tested in our study, SAM involves most tuning, whereas LIMMA requires the user to provide a model matrix specifying the linear model to be fitted. A benefit of ROTS is that all of the parameters are inferred directly from the data without any need of user intervention. The output parameters, on the contrary, can provide some useful information about the success of the analysis. In particular, low reproducibility *Z*-score in ROTS (for example below 2 to 3) often indicates that the data or the statistic family is not sufficient for reliable detection. For instance, in the yeast data, the most challenging comparison 2vs4 fmol UPS1, in which the performance of all the methods was poor, also yielded a relatively low reproducibility *Z*-score with ROTS, which could indicate the approximation of the lower quantification limit of the Progenesis algorithm.

To conclude, the constant growth in the volume of quantitative proteome data requires methods that are reliable, user-friendly, and straightforward to interpret. Even though there are a number of statistical methods that are applied to discover differential protein expression, the reliability of some methods is concerning. The extensive statistical analyses carried out in this study lead us to conclude that ROTS provides reliable rankings of differential expression in proteome data and that it is simple to use. ROTS-package for R is available at our Web site (http://www.btk.fi/research/research-groups/elo/) as well as in Chipster software (http://chipster.csc.fi/).

DOI: 10.1021/acs.jproteome.5b00183
J. Proteome Res. 2015, 14, 4118−4126

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.5b00183.

Supplementary Table 1. Sensitivity and specificity of the statistical methods in the yeast spiked data. Supplementary Table 2. Optimized ROTS parameters and the observed reproducibility values in the spiked yeast comparisons. Supplementary Table 3. Median sensitivity and specificity using different normalization approaches. Supplementary Figure 1. Profiles of the 46 UPS1 protein intensities for different concentrations. Supplementary Figure 2. Profiles of the intensities of the ROTS detections from different comparisons. Supplementary Figure 3. Receiver operating characteristic (ROC) curves in the spiked yeast data sets together with the area under the curve (AUC) for each method. Supplementary Figures 4−6. Effect of the number of peptides used for quantitation on the ROC performance, number of UPS1 proteins detected (true positives), and on number of yeast proteins detected (false positives) in the spiked yeast data. (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*L.L.E.: Tel: +358 2 333 8009. Fax: +358 2 251 8808. E-mail: laliel@utu.fi.

*G.L.C.: Tel: +31205255406. E-mail: corthals@uva.nl.

Author Contributions

#A.P., A.P.V., and S.A. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We gratefully acknowledge the help from the Turku Proteomics Facility of the Biocenter Finland Proteomics and Metabolomics infrastructure, the Finnish National Doctoral Programme in Informational and Structural Biology, P. Haapaniemi from the Turku Centre for Biotechnology for assistance, K. Hovirinta, H. Niittymäki, N. Messner, and J. Palmu from the Turku Center for Disease Modeling for their assistance, and the personnel at Turku Center for Disease Modeling for technical assistance with animal experimentation. This work was supported by JDRF [grant number 2-2013-32], the Päivikki and Sakari Sohlberg Foundation, Sigrid Jusélius foundation, and NordForsk (Nordic QP: Nordic Education Network for Quantitative Proteomics, grant number 070178).

■ REFERENCES

(1) Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M.-C.; Yates, J. R. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chem. Rev.* **2013**, *113* (4), 2343−2394.

(2) Sadygov, R. G.; Cociorva, D.; Yates, J. R. Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. *Nat. Methods* **2004**, *1* (3), 195−202.

(3) Kannaste, O.; Suomi, T.; Salmi, J.; Uusipaikka, E.; Nevalainen, O.; Corthals, G. L. Cross-Correlation of Spectral Count Ranking to Validate Quantitative Proteome Measurements. *J. Proteome Res.* **2014**, *13* (4), 1957−1968.

(4) Domon, B.; Aebersold, R. Options and considerations when selecting a quantitative proteomics strategy. *Nat. Biotechnol.* **2010**, *28* (7), 710−721.

(5) Nahnsen, S.; Bielow, C.; Reinert, K.; Kohlbacher, O. Tools for Label-free Peptide Quantification. *Mol. Cell. Proteomics* **2013**, *12* (3), 549−556.

(6) Tuli, L.; Tsai, T.-H.; Varghese, R. S.; Xiao, J. F.; Cheema, A.; Ressom, H. W. Using a spike-in experiment to evaluate analysis of LC-MS data. *Proteome Sci.* **2012**, *10*, 13.

(7) Bondarenko, P. V.; Chelius, D.; Shaler, T. A. Identification and Relative Quantitation of Protein Mixtures by Enzymatic Digestion Followed by Capillary Reversed-Phase Liquid Chromatography−Tandem Mass Spectrometry. *Anal. Chem.* **2002**, *74* (18), 4741−4749.

(8) Chelius, D.; Bondarenko, P. V. Quantitative Profiling of Proteins in Complex Mixtures Using Liquid Chromatography and Mass Spectrometry. *J. Proteome Res.* **2002**, *1* (4), 317−323.

(9) Clough, T.; Thaminy, S.; Ragg, S.; Aebersold, R.; Vitek, O. Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs. *BMC Bioinf.* **2012**, *13* (Suppl 16), S6.

(10) Matzke, M. M.; Brown, J. N.; Gritsenko, M. A.; Metz, T. O.; Pounds, J. G.; Rodland, K. D.; Shukla, A. K.; Smith, R. D.; Waters, K. M.; McDermott, J. E.; et al. A comparative analysis of computational approaches to relative protein quantification using peptide peak intensities in label-free LC-MS proteomics experiments. *Proteomics* **2013**, *13* (0), 493−503.

(11) Bantscheff, M.; Lemeer, S.; Savitski, M. M.; Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* **2012**, *404* (4), 939−965.

(12) 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.

(13) Sturm, M.; Bertsch, A.; Gropl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; et al. OpenMS - An open-source software framework for mass spectrometry. *BMC Bioinf.* **2008**, *9*, 163.

(14) Christin, C.; Hoefsloot, H. C. J.; Smilde, A. K.; Hoekman, B.; Suits, F.; Bischoff, R.; Horvatovich, P. A Critical Assessment of Feature Selection Methods for Biomarker Discovery in Clinical Proteomics. *Mol. Cell. Proteomics* **2013**, *12* (1), 263−276.

(15) Tusher, V. G.; Tibshirani, R.; Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98* (9), 5116−5121.

(16) Smyth, G. K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Stat. Appl. Genet. Mol. Biol.* **2004**, *3* (1), 1.

(17) Breitling, R.; Armengaud, P.; Amtmann, A.; Herzyk, P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.* **2004**, *573* (1−3), 83−92.

(18) Elo, L. L.; Filen, S.; Lahesmaa, R.; Aittokallio, T. Reproducibility-Optimized Test Statistic for Ranking Genes in Microarray Studies. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2008**, *5* (3), 423−431.

(19) Elo, L. L.; Hiissa, J.; Tuimala, J.; Kallio, A.; Korpelainen, E.; Aittokallio, T. Optimized detection of differential expression in global profiling experiments: case studies in clinical transcriptomic and quantitative proteomic datasets. *Briefings Bioinf.* **2009**, *10* (5), 547−555.

(20) Spivak, M.; Weston, J.; Bottou, L.; Kall, L.; Noble, W. S. Improvements to the Percolator algorithm for peptide identification from shotgun proteomics data sets. *J. Proteome Res.* **2009**, *8* (7), 3737−3745.

(21) Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J. Accurate and Sensitive Peptide Identification with Mascot Percolator. *J. Proteome Res.* **2009**, *8* (6), 3176−3181.

(22) Nezami Ranjbar, M. R.; Zhao, Y.; Tadesse, M. G.; Wang, Y.; Ressom, H. W. Gaussian process regression model for normalization

DOI: 10.1021/acs.jproteome.5b00183
J. Proteome Res. 2015, 14, 4118−4126

of LC-MS data using scan-level information. *Proteome Sci.* **2013**, *11* (Suppl 1), S13.
(23) Li, X.; Nokkala, E.; Yan, W.; Streng, T.; Saarinen, N.; Warri, A.; Huhtaniemi, I.; Santti, R.; Makela, S.; Poutanen, M. Altered structure and function of reproductive organs in transgenic male mice overexpressing human aromatase. *Endocrinology* **2001**, *142* (6), 2435−2442.
(24) Li, X.; Strauss, L.; Kaatrasalo, A.; Mayerhofer, A.; Huhtaniemi, I.; Santti, R.; Mäkelä, S.; Poutanen, M. Transgenic Mice Expressing P450 Aromatase as a Model for Male Infertility Associated with Chronic Inflammation in the Testis. *Endocrinology* **2006**, *147* (3), 1271−1277.
(25) Li, X.; Strauss, L.; Mäkelä, S.; Streng, T.; Huhtaniemi, I.; Santti, R.; Poutanen, M. Multiple Structural and Functional Abnormalities in the P450 Aromatase Expressing Transgenic Male Mice Are Ameliorated by a P450 Aromatase Inhibitor. *Am. J. Pathol.* **2004**, *164* (3), 1039−1048.
(26) Schmidt, A.; Gehlenborg, N.; Bodenmiller, B.; Mueller, L. N.; Campbell, D.; Mueller, M.; Aebersold, R.; Domon, B. An Integrated, Directed Mass Spectrometric Approach for In-depth Characterization of Complex Peptide Mixtures. *Mol. Cell. Proteomics* **2008**, *7* (11), 2138−2150.
(27) Vizcaíno, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Ríos, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **2014**, *32* (3), 223−226.
(28) Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc., Ser. B* **1995**, *57* (1), 289−300.
(29) Choi, M.; Chang, C.-Y.; Clough, T.; Broudy, D.; Killeen, T.; MacLean, B.; Vitek, O. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **2014**, *30*, 2524−2526.
(30) Gregori, J.; Villarreal, L.; Sánchez, A.; Baselga, J.; Villanueva, J. An effect size filter improves the reproducibility in spectral counting-based comparative proteomics. *J. Proteomics* **2013**, *95*, 55−65.

DOI: 10.1021/acs.jproteome.5b00183
J. Proteome Res. 2015, 14, 4118−4126