HHS Public Access

Author manuscript

J Proteomics. Author manuscript; available in PMC 2016 November 03.

Published in final edited form as:

J Proteomics. 2015 November 3; 129: 121–126. doi:10.1016/j.jprot.2015.07.036.

QPROT: statistical method for testing differential expression using protein-level intensity data in label-free quantitative proteomics

Hyungwon Choi[#†],
Saw Swee Hock School of Public Health, National University of Singapore

Sinae Kim[#],
Department of Biostatistics, Rutgers University

Damian Fermin[#],
Department of Pathology, Yale University

Chih-Chiang Tsou, and
Department of Pathology, University of Michigan Medical School

Alexey I. Nesvizhskii[†]
Departments of Pathology and Computational Medicine and Bioinformatics, University of Michigan Medical School

[#] These authors contributed equally to this work.

Abstract

We introduce QPROT, a statistical framework and computational tool for differential protein expression analysis using protein intensity data. QPROT is an extension of the QSPEC suite, originally developed for spectral count data, adapted for statistical significance analysis using continuously measured protein-level intensity data. QPROT offers a new intensity normalization procedure and model-based differential expression analysis, both of which account for missing data. Determination of differential expression of each protein is based on the standardized *Z*-statistic based on the posterior distribution of the log fold change parameter, guided by the false discovery rate estimated by a well-known Empirical Bayes method. We evaluated the classification performance of QPROT using the quantification calibration data from the clinical proteomic technology assessment for cancer (CPTAC) study and a recently published *E. coli* benchmark dataset, with evaluation of FDR accuracy in the latter.

Keywords

Differential expression; intensity; continuously normalized spectral counts; missing data

[†]To whom all correspondence should be addressed. hyung_won_choi@nuhs.edu.sg, nesvi@med.umich.edu..

Availability: The software suite is freely available on the Sourceforge website http://sourceforge.net/p/qprot/.

Introduction

Mass spectrometry-based proteomics is an essential tool for profiling post-translational expression level of genes. With the evolving technology and data analysis pipeline, tandem mass spectrometry (MS/MS) has been increasingly applied to proteome-wide investigations, not only for protein identification but also quantification [1]. Early quantitative proteomics relied on relative quantification based on labelling proteins with stable isotope containing compounds [2] or isobaric chemicals [3]. More recently, label-free quantification has become a widely used method to generate semi- or fully quantitative measurements.

Spectral counting has been a popular label-free quantification method due to the ease with which the data can be obtained by counting the number of peptide-spectrum matches for each protein from MS/MS search results. Spectral counting has been shown to correlate well with known protein concentrations [4] and regarded as a robust measure of protein abundance in the analysis of cell lines as well as tissue samples. However, count data suffers from poor resolution in the low abundance range (e.g. proteins detected with a single peptide-spectrum match), and there are ambiguities in counting peptides which are shared among homologous proteins, requiring careful handling of counts in complex organisms such as human [5]. With rapid advances in MS instrumentation and informatics tools for data extraction such as MaxQuant [6] and OpenMS suite [7], peptide ion intensities can be easily extracted from high resolution MS datasets, providing a measurement accuracy superior to spectral counts. More recent development includes data independent mode of acquisition of MS/MS spectra, providing a large amount of fragment-level intensities to derive peptide/protein intensities with improved accuracy and coverage of proteome [8, 9, 10, 11], broadening the range of intensity-based quantification.

With the availability of high resolution and high mass accuracy intensity data, it is of interest for many experimentalists to have access to an appropriate statistical method to detect differentially expressed proteins based on intensity data. For spectral count data, a handful of statistical methods for differential expression (DE) analysis had been developed, including standard hypothesis testing procedures [12, 13], power law global error model [14], QSPEC [15], and Bayesian mixture model-based method [16]. On the other hand, it has been generally perceived that DE analysis of intensity data can be performed using the methods developed for analyzing gene expression microarray data because the data are in a continuous scale, such as simple hypothesis tests or LIMMA [17]. More recently developed methods enabled a regression model-based analysis framework that performs peptide-protein roll-up within the statistical models, such as DANTE [18] and MSstats [19, 20]. These methods estimate the fold change parameters from peptide or transition intensity data and compute statistical significance scores (*p*-values) for individual proteins.

While these tools cater to many applications, not all of them are applicable to simple comparisons based on protein intensity data from a small number of samples. For instance, proteins with poor sequence coverage in the MS/MS data will have few peptides. In addition, not all peptides necessarily have high quality peak clusters with clear isotopic patterns, hampering consistent quantification across samples (across multiple MS runs).

Hence the model-based approaches built for peptide-level intensity data will not be able to perform significance analysis unless those values are imputed. The model-based methods [18] may offer a treatment of missing data at the peptide-level, but their likelihood-based imputation scheme requires more than several samples per comparison group. Hence the procedure requires removal of peptides with to few intensity values across different samples, which may occur frequently for a large number of low abundance proteins in complex samples.

Reflecting this, label-free intensity data summarized up to the protein level are increasingly used in proteome-level quantification studies, exemplified by popular methods such as iBAQ [21, 22] and "top 3 peptides" approaches [23, 24, 25]. When protein intensity data (summed over peptides) are analyzed by standard or advanced statistical tests (e.g. LIMMA), the missing data problem arise again for the proteins not quantified in certain samples, requiring imputation of those missing data. In this scenario, a reasonable missing data treatment is one-time imputation of intensity values, typically chosen as a value smaller than the smallest intensity in the sample (e.g. half). However, such an imputation procedure does not always represent underlying missing mechanisms and the choice of plug-in value for imputation can change the outcome of the statistical tests.

In this work, we developed QPROT, a model-based DE analysis for protein intensity data. The new implementation, packaged with the QSPEC method for spectral count data, features a new intensity normalization procedure, a hierarchical model for differential expression analysis using protein intensity data, and a revised procedure for estimating the false discovery rates (FDR) and thus guiding the selection of differentially expressed proteins. This extension for intensity data will also have added utility for the analysis of continuously re-scaled spectral counts such as exemplified by NSAF [26], APEX [27], and SINQ [28], and hence it will be a useful tool for first-pass statistical analysis using protein-level intensity data.

Materials and Methods

Statistical model

Suppose that we have a matrix data containing log-transformed protein intensities $Y = \{y_{ij}\}$, where y_{ij} is the measurement of protein i in sample j for $i = 1, 2, \ldots, P$ and $j = 1, 2, \ldots, N$. Further, let Y_j denote the quantitative data for all P proteins in sample j. We define T_j, a binary indicator taking on 0 or 1, as the comparison groups such as control (0) and treatment (1). We assume that the data $\{y_{ij}\}$ were log-transformed, and they follow normal distribution

$$y_{ij} \sim \mathrm{N}\left(\mu_i + d_i T_j, \sigma_i^2\right) \quad (1)$$

where d_i is the magnitude of differential expression in log scale.

Here, note that we model protein level expression data directly, not peptide level data. While assuming peptide-level data is an ideal format for intensity analysis in theory, many studies rely on protein intensity computed by summing the intensities of all or selected peptides for each protein. The peptide to protein roll-up is performed because, in practice, it is not

possible to obtain consistent measurement across multiple samples for many peptides. For example, a protein can be consistently identified and quantified by multiple peptides in all experiments that are being compared, but not necessarily by the same peptides. In these cases, analysis based on the average intensity of top 3 most intense peptides per protein has been shown to be more effective than methods based on using all peptides in certain applications [29].

QPROT also offers an optional intensity normalization step in which all percentile points of the *observed* quantitative values are equalized across the samples. The procedure is equivalent to the quantile normalization in microarray data, with the difference that the percentiles are normalized accounting for the fact that some proteins have no measurement in certain samples / experiments. First, all missing observations are removed to form a trimmed set Y_j' for each j. Then the 0% to 100% percentile points are computed for each sample separately and these points are set to the median values across samples at each percentile, and the data are finally normalized by interpolation between two nearest percentile points for each protein in each sample.

Furthermore, since not all y_{ij} are observed, we treat them as missing data and consider the following truncation rules for integrating likelihood over the low abundance range. First, we integrate the density over $(-\infty, \varphi_{i,T_j})$, where φ_{i,T_j} denotes the truncation point in the low abundance area for protein i in group T_j. φ_{i,T_j} is set to be the smallest value for protein i if it is observed at least in one sample of group T_j. If there are missing values in both groups of comparison, we set the truncation points at the minimum of φ_{i0} and φ_{i1}. If values are all missing in group 0 and fully observed (i.e. observed across all samples) in group 1, φ_{i0} is set to the 10 percentile point of all observed values in group T_j, and vice versa. Once the truncation point is determined, the likelihood for protein i is defined as follows. Let us introduce the following notation for the observed data and missing data:

$$y_{ij} = \begin{cases} y_{ij}^{(o)} & \text{if} \quad y_{ij} \geq \phi_{i,T_j} \\ y_{ij}^{(m)} & \text{if} \quad y_{ij} < \phi_{i,T_j} \end{cases} \quad (2)$$

where $y^{(o)}$ and $y^{(m)}$ are the observed and missing values, respectively. For protein i, the likelihood function is the probability of the observed values, i.e.

$$p\left(Y_i^{(o)} | \theta_i, \phi_{i0}, \phi_{i1}\right) \propto \int p\left(Y_i^{(o)}, Y_i^{(m)} | \theta_i, \phi_{i0}, \phi_{i1}\right) dY_i^{(m)} = \prod_{j: y_{ij} \geq \phi_{i,T_j}} \varphi\left(y_{ij} | \theta_i\right) \times \prod_{j: y_{ij} < \phi_{i,T_j}} \int_{-\infty}^{\phi_{i,T_j}} \varphi\left(y_{ij} | \theta_j\right) dy_{ij}, \quad (3)$$

where $\theta_i = \left(\mu_i, d_i, \sigma_i^2\right)$, and $Y_i^{(o)}$ is a vector of observed intensities for protein i and $\phi(x|a, b)$ indicates a normal density with mean a and variance b evaluated at x (the cumulative distribution is denoted by Φ hereafter). A more complete modeling would have been of a mixture form

$$p\left(y_{ij}\right)=\begin{cases}\varphi\left(y_{ij}|\mu_i+d_iT_j,\sigma_i^2\right) & \text{if} \quad y_{ij}\geq\phi_{i,T_j}\\ \delta_i+\left(1-\delta_i\right)\frac{\varphi\left(\phi_{i,T_j}|\mu_i+d_iT_j,\sigma_i^2\right)}{1-\Phi\left(\phi_{i,T_j}|\mu_i+d_iT_j,\sigma_i^2\right)} & \text{if} \quad y_{ij}<\phi_{i,T_j}\end{cases} \quad (4)$$

where $\delta_i = P(y_{ij} = -\infty)$ is the prior probability of true absence of protein *i*, and a truncated normal distribution is used to model the observed and missing intensity for truly present proteins. However, δ_i varies by samples and is usually not estimable in each data set.

Since our goal is to test differential expression, we are interested in the inference of the magnitude of differential expression d_i. In combination with the likelihood, we specify the priors as follows:

$$\left(\mu_i,d_i\right)\sim N\left(0,10^2\right)\times N\left(0,10^2\right) \quad (5)$$

$$\sigma_i^2\sim IG\left(1,0.1\right) \quad (6)$$

where $IG(\cdot, \cdot)$ stands for inverse gamma distribution with shape and scale parameters. For posterior inference, we use a standard Markov chain Monte Carlo sampler (Metropolis-Hastings algorithm [30]) to draw samples of the parameters from the appropriate posterior distributions. In this work 10,000 samples were drawn for the burn-in period and 100,000 samples of each model parameter were drawn for the main iterations. Particularly, the log fold change parameter d_i was recorded for every protein *i*, denoted by $\left(d_i^{(1)},\ldots,d_i^{(100,000)}\right)$. The resulting collection of 100,000 samples of d_i was used to construct the significance statistic Z_i later.

Significance statistics

Using the posterior samples of the log fold change parameter, the standardized significance statistic of DE for protein *i* is computed as

$$Z_i=\frac{\hat{d}_i}{\sqrt{\widehat{var}\left(\hat{d}_i\right)}} \quad (7)$$

where the numerator and denominator denote the mean and standard error computed from the posterior samples of d_i respectively. The test statistics are conceptually different from the statistics based on the "odds" of DE in the existing methods for spectral count data [15, 16]. Rather, *Z*-statistics are close to the most intuitive differential expression statistics such as *t*-statistic, where the mean differential d_i is normalized by the standard error. The difference between our *Z*-statistics and the existing tests is that the former was computed accounting for missing data. In our experience with QSPEC, the odds-based statistics such as Bayes factor tend to be considerably more sensitive for high abundance proteins due to sharp tails of Poisson distributions, whereas the drawback is relatively mitigated in *Z*-statistic. This change was motivated by the fact that biologically interesting proteins are

equally well populated in the low and intermediate abundance range, where fold change is accurately estimated but the odds of DE are often underestimated.

Multiple testing correction by FDR

After calculating (Z_1, Z_2, . . . , Z_P), we proceed to hypothesis testing with a multiple testing correction. To achieve this, we fit a semi-parametric mixture model to deconvolute the score distributions for the DE proteins and non-DE proteins as follows. In specific, we estimate

$$f(z_i) = \pi f_1(z_i) + (1-\pi) f_0(z_i) \quad (8)$$

$$= \pi f_1(z_i) + (1-\pi) \sum_{k=1}^{K} \gamma_k \varphi\left(z_i; \eta_k, \tau_k^2\right) \quad (9)$$

for all $i = 1, 2, \ldots, P$, where ϕ denotes the density of normal distribution, γ_k is the mixing proportion of k-th mixture component with mean and variance (η_k, τ_k^2), and π is the proportion of differentially expressed proteins. In addition, we estimate the overall distribution f by the non-parametric density estimation with Gaussian kernel [31]

$$\hat{f}_h(z) = \frac{1}{hP} \sum_{i=1}^{P} \varphi\left(\frac{z - z_i}{h}\right) \quad (10)$$

where $\phi()$ denotes the standard Gaussian density. We selected the bandwidth as $2.2\hat{\tau}P^{-1/5}$, twice the recommended size for extra smoothness of the curve, where $\hat{\tau}$ is the standard deviation of observed Z-statistics. Finally, the mixing proportion π by the methodology proposed in the Empirical Bayes method proposed in [32], i.e.

$$\pi = 1 - \min_z \{f(z)/f_0(z)\}. \quad (11)$$

Here K is the number of normal distributions consisting of the null distribution f_0, which was originally set as $K = 1$ in the first few releases of QPROT (up to version 1.3.0), under the normality assumption in the hypothesis testing framework. In our model, however, we allow for situations where the score statistic is not exactly normal but approximately bell-shaped. Hence we let the user choose the optimal number of components to capture the null component properly. The default value is $K = 3$ in the current release of the software (version 1.3.1), with the last argument to the command line call being the option to specify this number. One of the output files, with suffix "_density," provides the mixture model fit across the score range and allows the user to determine the optimal K. The model fit in this file can be visualized by an accompanying R script into a pdf file (drawMixFit.R script).

Following the model estimation, the FDR associated with a cuto z^* is computed as

$$FDR(z^*) = \frac{(1-\pi)\int_{z>z^*} f_0(z)\,dz}{\int_{z>z^*} f(z)\,dz}. \quad (12)$$

for the proteins over-expressed in group 1 and vice versa for the proteins over-expressed in group 0. We remark that the hypothesis testing framework in QSPEC [15], originally based on Bayes factors, has also been replaced by the testing framework based on *Z*-statistics and subsequent FDR estimation as proposed above.

Experimental Design: Independent sample comparison versus paired sample comparison

Although we do not fully demonstrate it here using real datasets, we note that the probability model for both QSPEC and QPROT now provide two experimental design options, namely independent sample comparisons versus paired sample comparisons. Most experiments are performed in the independent sample comparisons, where the two groups of samples are either biological replicates or biologically unrelated samples. In some experiments, however, the same samples are analyzed in different conditions or across time points. In this case, the basal protein concentrations will be more correlated within each biological sample, and therefore such correlation can be incorporated in the model, just as the statistical tests for paired samples do (e.g. paired *t*-test). The “paired design” option in QSPEC and QPROT accommodates this experimental designs with proper ordering of the paired samples in the input file (see the software manual).

Experimental Data sets

We obtained the processed quantification data for the Clinical Proteomic Technology Assessment for Cancer (CPTAC) data [33, 34] from the author of the latter publication. In this study, 48 proteins from an equimolar Universal Protein Standard (UPS1) sample were spiked into a 60ng yeast lysate background in five different concentrations (each differing by three fold) and run without pre-fractionation. Proteins were quantified using both spectral count data and intensity data processed through MaxQuant [6], as well as various forms of counting-based quantification methods such as NSAF [26], empai [35], and SINQ [28].

We also used another benchmark dataset from Shalit *et al* [36], who quantified *E. coli* digest spiked into a HeLa digest in four different concentrations (1.5 fold to 5 fold). Supplementary Table 3 of the paper contained the protein intensity data processed by Expressionist software (Expressionist), and all the *E. coli* proteins and human proteins were considered as differentially expressed and non-differentially expressed proteins respectively.

Results and Discussion

CPTAC data

We first applied QPROT to the analysis of intensity and continuously normalized spectral count data from the CPTAC study [33, 34]. In the CPTAC study, 48 UPS1 proteins were spiked into the background of yeast cell lysate (~1,000 proteins) with varying concentrations ranging from 0.74 to 20 fmol differing by 3 fold differences. Three biological replicates were generated for each concentration. Since there are four different concentration levels, one can perform 6 pairwise comparisons where 48 DE proteins are differentially expressed by 3, 9, or 27 fold, and therefore this provides an opportunity to assess the sensitivity and specificity of DE methods.

We evaluated the classification performance of QPROT in terms of the receiver-operating characteristic using MaxQuant intensity data and SINQ data, a normalized measure incorporating both MS and MS/MS data [34]. For reference purposes, we also compared the performance to QSPEC and a Bayesian mixture model approach of [16] using the spectral count data. As a more relevant alternative, we also compared QPROT to LIMMA [17] using the MaxQuant intensity and SINQ data, which implements an empirical Bayes method calculating moderated *t* statistics without special treatment of missing data. For LIMMA analysis, we imputed half the minimum observed intensity for missing observations in each sample, since the software package requires some positive number for log scaling. In group comparisons with 3, 9, and 27 fold differences, we recorded the number of UPS1 proteins at a fixed proportion of non-UPS1 background proteins to estimate the sensitivity at a fixed type I error rate.

Using the MaxQuant intensity data, QPROT tended to capture as many or more UPS1 proteins than LIMMA analysis at fixed false discovery proportion, i.e. fraction of non-UPS1 proteins across all comparisons (Figure 1). QPROT recovered ~ 40 UPS1 proteins at 5% background proportion (95% specificity) across all datasets with 3 fold difference, and selected 4 to 6 more UPS1 proteins (out of 48) at 93% to 97% specificity regions in 9 and 27 fold data, slightly improving LIMMA analysis. Meanwhile, QPROT analysis of SINQ data also showed comparable or better performance than LIMMA analysis of SINQ data with the exception of 0.74/6.7fmol comparison, and both analyses were also superior to all other analyses using spectral count data (QSPEC, BayesMix, and NSAF; data not shown as they were inferior or less comparable to SINQ). This result indicates that the intensity-based analysis with QPROT has the potential to improve upon the widely used LIMMA method in these data. Nevertheless, further investigation should be performed in datasets with more missing data, since there were too few missing observations in the CPTAC datasets overall. At most 95 observations were missing (1.3%) in the comparison between 0.74fmol versus 2.2fmol, in which the frequency of missing data was the highest.

E. coli spike-in dataset

Next we evaluated QPROT using the *E. coli* spike-in dataset, with 227 *E. coli* proteins and 1,824 human proteins representing differentially expressed and non-differentially expressed proteins, respectively [36]. Since the *E. coli* proteins were spiked in four different concentrations, we performed all 6 pairwise comparisons using both QPROT and LIMMA. We explored the analysis with and without normalization, but the overall performance remained about the same, and hence we present the analysis outcome using the version with the normalization step.

Figures 2A and 2B show that both QPROT and LIMMA yield good classification performance in this data, distinguishing differentially spiked *E. coli* proteins from the background HeLa proteome. The only analysis in which both methods performed poorly was the comparison between 10μg spike-in and 5μg spike-in, where the intensity values were indeed equivalent across many proteins. In the comparison between 5μg spike-in and 3μg spike-in, QPROT demonstrated superior sensitivity to LIMMA, and we thus investigated the proteins that were called significant by QPROT but not by LIMMA

(Supplementary Table 1). QPROT selected 40 *E. coli* proteins and 19 human proteins more than LIMMA at the same threshold. Many of the *E. coli* proteins uniquely captured by QPROT were those that had at least one missing value in either or both comparison groups (red color, Supplementary Table 1). Meanwhile, the human proteins captured by QPROT were mostly low abundance proteins but the estimated fold change was at least over 50% in half of them.

Related to this observation, Figures 2C and 2D illustrate that QPROT slightly overestimated the FDR in some comparisons (excluding 15μg vs 3μg) when benchmarked against the false discovery proportion based on the proportion of human proteins, whereas the *q*-values derived from the *p*-values reported by the empirical Bayes model in LIMMA showed a variable degree of accuracy against the benchmark depending on the fold difference and the data quality of each concentration. However, we noticed that most *E. coli* proteins with modestly large fold changes (2 fold or more) were mostly selected at score thresholds associated with very low FDR by both methods in this benchmark dataset, and thus it was difficult to conclude one method has better FDR accuracy than the other.

Conclusion

In this work, we presented a protein DE analysis software QPROT, implementation of flexible statistical models for the two major types of quantitative proteomics data and proper treatment of missing values for the low abundance proteins in the intensity and continuously normalized count data. QPROT also offers a flexible semi-parametric mixture model-based FDR estimation routine, which is powerful for differentiating the differential expression of even a minor effect size due to its adaptive property. Overall, our new development in this work should be immediately useful for many proteomics laboratories analyzing label-free quantitative datasets.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

The authors thank Dr. David Trudgian for sharing the label-free quantitative data for the CPTAC study. This work was supported in part by Ministry of Education Tier 2 grant R-608-000-088-012 (to HC); UMDNJ Foundation grant PC18-11 (to SK); and NIH grant R01-GM-094231 (to AIN).

References

1. Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat. Methods. 2007; 4(10):787–797. [PubMed: 17901868]
2. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen A, Pandey H, Mann M. Stable isotope labeling by amino acids in cell culture, silac, as a simple and accurate approach to expression proteomics. Mol. Cell. Proteomics. 2002; 1(5):376–386. [PubMed: 12118079]
3. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlet-Jones M, He A, Jacobson F, Pappin DJ. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol. Cell. Proteomics. 2004; 3(12):1154–1169. [PubMed: 15385600]

4. Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, Sevinsky JR, Resing KA, Ahn NG. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. Mol. Cell. Proteomics. 2005; 4:1487–1502. [PubMed: 15979981]

5. Fermin D, Basrur V, Yocum AK, Nesvizhskii AI. Abacus: a computational tool for extracting and pre-processing spectral count data for label-free quantitative proteomic analysis. Proteomics. 2011; 11(7):1340–1345. [PubMed: 21360675]

6. 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; 26:1367–1372. [PubMed: 19029910]

7. Sturm M, Bertsch A, Gropl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K, Kohlbacher O. OpenMS - an open-source software framework for mass spectrometry. BMC Bioinformatics. 2008; 9:163. [PubMed: 18366760]

8. Veneable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat. Methods. 2004; 1(1):39–45. [PubMed: 15782151]

9. Carvalho PC, Han X, Xu T, Cociorva D, Carvalho Mda G. Barbosa VC, Yates JR 3rd. XDIA: improving on the label-free data-independent analysis. Bioinformatics. 2010; 26(6):847–8. [PubMed: 20106817]

10. Gillet LC, Navarro P, Tate S, Röst HL, Selevsek N, Reiter L, Bonner R, Aeber-sold R. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analaysis. Mol. Cell. Proteomics. 2012; 11(6):O111.016717. [PubMed: 22261725]

11. Tsou C-C, Avtonomov D, Larsen B, Tucholska M, Choi H, Gingras A-C, Nesvizhskii AI. DIA-Umpire: comprehensive computational framework for data independent acquisition proteomics. Nat. Methods. 2015; 12(3):258–264. [PubMed: 25599550]

12. Heinecke NL, Pratt BS, Vaisar T, Becker L. PepC: proteomics software for identifying differentially expressed proteins based on spectral counting. Bioinformatics. 2010; 26(12):1574–1575. [PubMed: 20413636]

13. Pham TV, Piersma SR, Warmoes M, Jimenez CR. On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. Bioinformatics. 2010; 26(3):363–369. [PubMed: 20007255]

14. Pavelka N, Fournier ML, Swanson SK, Pelizzola M, Ricciardi-Castagnoli P, Florens L, Washburn MP. Statistical similarities between transcriptomics and quantitative shotgun proteomics data. Mol. Cell. Proteomics. 2008; 7(4):631–644. [PubMed: 18029349]

15. Choi H, Fermin D, Nesvizhskii AI. Significance analysis of spectral count data in label-free shotgun proteomics. Mol. Cell. Proteomics. 2008; 7(12):2373–2385. [PubMed: 18644780]

16. Booth JG, Eilertson KE, Olinares PD, Yu H. A Bayesian mixture model for comparative spectral count data in shotgun proteomics. Mol. Cell. Proteomics. 2011; 10:M110.007203. [PubMed: 21602509]

17. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statist. App. Gen. Mol. Biol. 2004; 3(1):3.

18. Karpievitch Y, Stanley J, Taverner T, Huang J, Adkins JN, Ansong C, Heffron F, Metz TO, Qian W-J, Yoon H, Smith RD, Dabney AR. A statistical framework for protein quantitation in bottom-up MS-based proteomics. Bioinformatics. 2009; 25(16):2028–2034. [PubMed: 19535538]

19. 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. [PubMed: 19891509]

20. Choi M, Chang CY, Clough T, Broudy D, Killeen T, MacLean BX, Vitek O. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics. 2014; 30(17):2524–2526. [PubMed: 24794931]

21. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. Global quantification of mammalian gene expression control. Nature. 2011; 473:337–342. [PubMed: 21593866]

22. Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, Paabo S, Mann M. Deep proteome and transcriptome mapping of a human cancer cell line. Mol. Sys. Biol. 2011; 7:548.

23. Ning K, Fermin D, Nesvizhskii AI. Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. J. Proteome Res. 2012; 11(4):2261–2271. [PubMed: 22329341]
24. Arike L, Valgepea K, Peil L, Nahku R, Adamberg K, Vilu R. Comparison and applications of label-free absolute proteome quantification methods on Escherichia coli. J. Proteomics. 2012; 75(17):5437–5448. [PubMed: 22771841]
25. Ahrne E, Molzhan L, Glatter T, Schmidt A. Critical assessment of proteome-wide label-free absolute abundance estimation strategies. Proteomics. 2013; 13(17):2567–2578. [PubMed: 23794183]
26. Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP. Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. J. Proteome Res. 2006; 5(9):2339–2347. [PubMed: 16944946]
27. Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat. Biotechnol. 2007; 25:117–124. [PubMed: 17187058]
28. Griffin NM, Yu J, Long F, Oh P, Shore S, Li Y, Koziol JA, Schnitzer JE. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nat. Biotechnol. 2010; 28:83–89. [PubMed: 20010810]
29. Silva JC, Gorenstein MV, Li G-Z, Vissers JPC, Geromanos SJ. Absolute quantification of proteins by LCMSE: a virtue of parallel ms acquisition. Mol. Cell. Proteomics. 2006; 5:144–156. [PubMed: 16219938]
30. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970; 57(1):97–109.
31. Silverman, BW. Density estimation for statistics and data analysis. Chapman & Hall/CRC; 1998.
32. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. J. Am. Statist. Assoc. 2001; 96:1151–1160.
33. Tabb DL, Vega-Montoto L, Rudnick PA, Variyath AM, Ham AJ, Bunk DM, Kil-patrick LE, Billheimer DD, Blackman RK, Cardasis HL, Carr SA, Clauser KR, Jaffe JD, Kowalski KA, Neubert TA, Regnier FE, Schilling B, Tegeler TJ, Wang M, Wang P, Whiteaker JR, Zimmerman LJ, Fisher SJ, Gibson BW, Kinsinger CR, Mesri M, Rodriguez H, Stein SE, Tempst P, Paulovich AG, Liebler DC, Spiegelman C. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. J. Proteome Res. 2010; 9:761–776. [PubMed: 19921851]
34. Trudgian DC, Ridlova G, Fischer R, Mackeen MM, Ternette N, Acuto O, Kessler BM, Thomas B. Comparative evaluation of label-free SINQ normalized spectral index quantitation in the central proteomics facilities pipeline. Proteomics. 2011; 26:2790–2797. [PubMed: 21656681]
35. Ishihama Y, Oda Y, Tabata T, Sato T, Takeshi N, Rappsilber J, Mann M. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol. Cell. Proteomics. 2005; 4(9):1265–1272. [PubMed: 15958392]
36. Shalit T, Elinger D, Savidor A, Gabashvili A, Levin Y. MS1-based label-free proteomics using a quadrupole orbitrap mass spectrometer. J. Proteome Res. 2015; 14(4):1979–1986. [PubMed: 25780947]

Significance

QPROT is a statistical framework with computational software tool for comparative quantitative proteomics analysis. It features various extensions of QSPEC method originally built for spectral count data analysis, including probabilistic treatment of missing values in protein intensity data. With the increasing popularity of label-free quantitative proteomics data, the proposed method and accompanying software suite will be immediately useful for many proteomics laboratories.

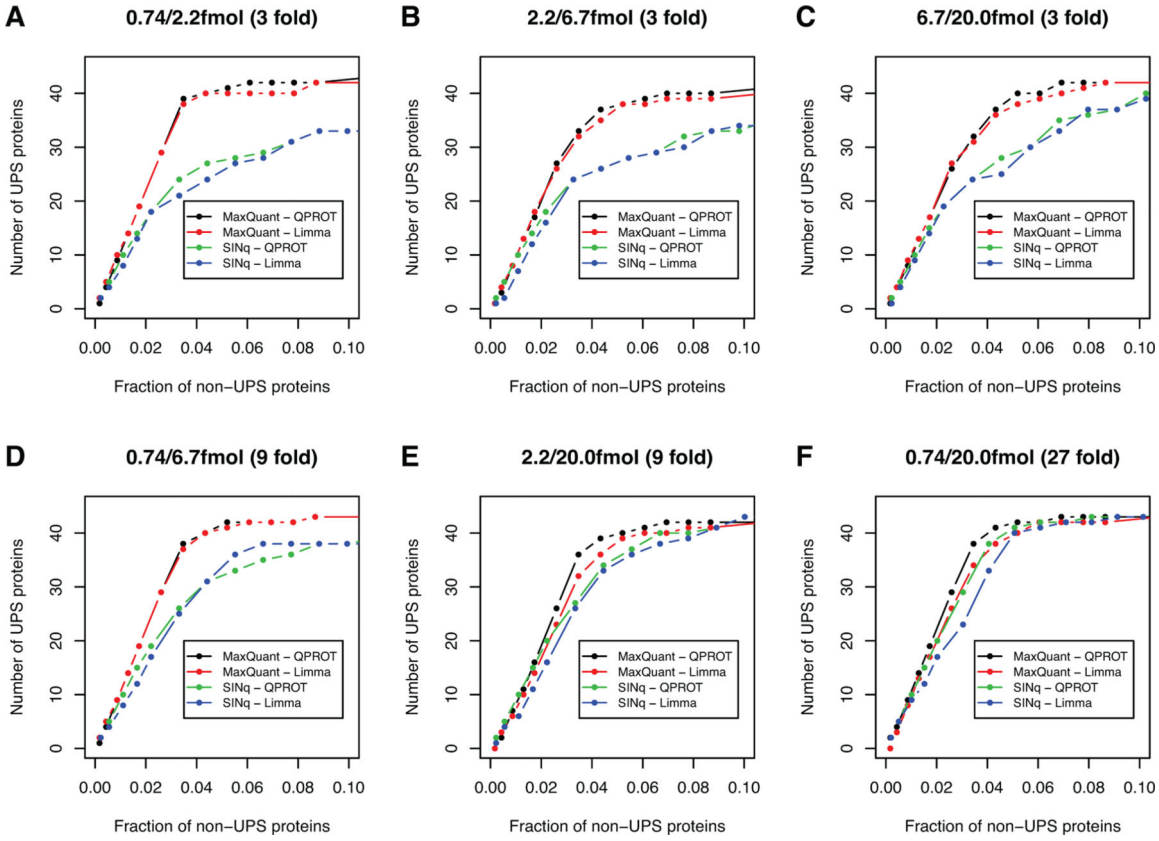

Figure 1.

Classification performance (receiver operating chracteristic) of QPROT and LIMMA applied to the MaxQuant intensity and SINQ data for the CPTAC dataset in all 5 comparisons between 4 different concentrations. The comparison between 10μg versus 5μg spike-in data was omitted due to lack of differences between the two.

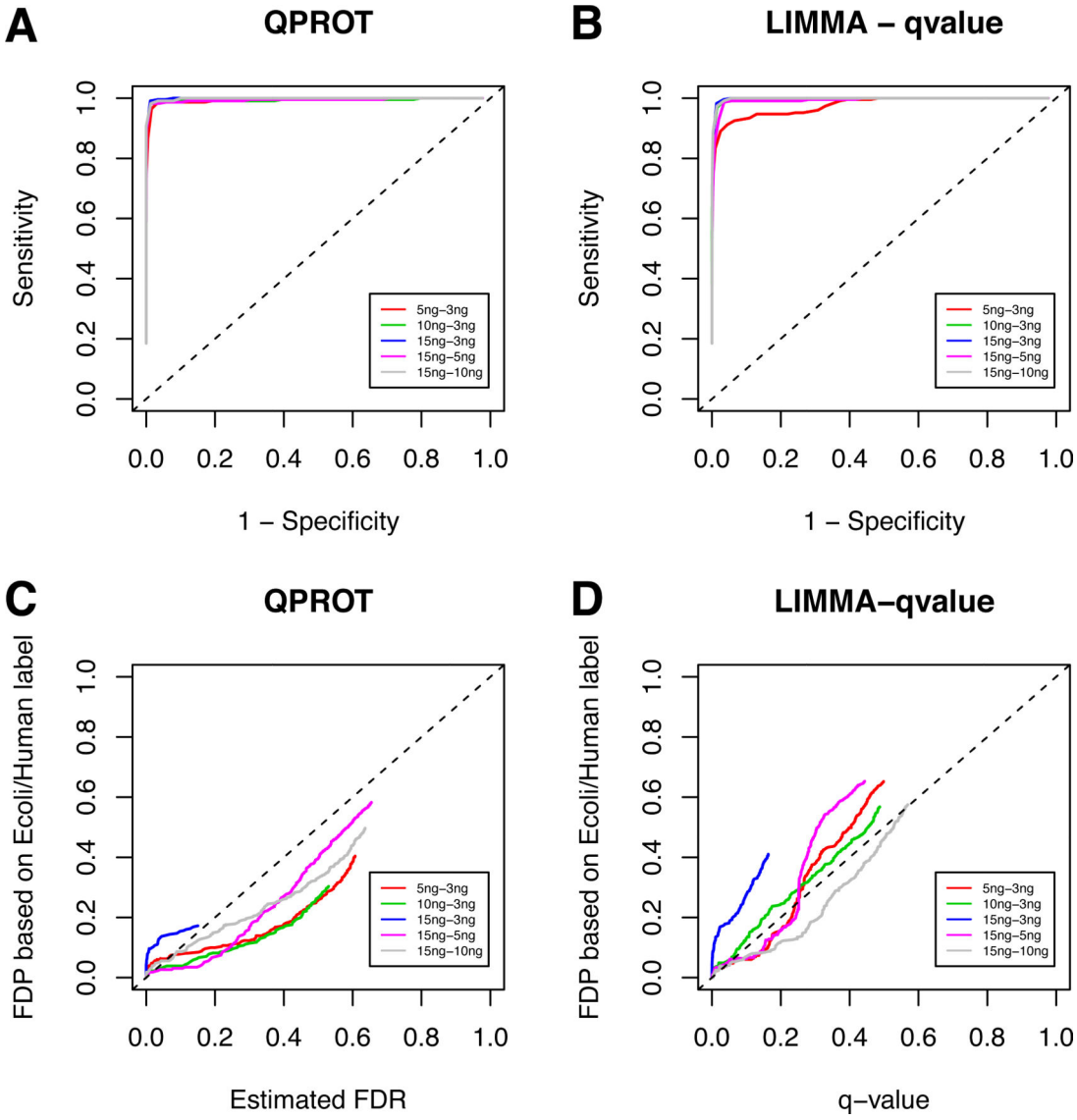

Figure 2.

Classification performance (receiver operating chracteristic) and FDR evaluation of QPROT and LIMMA applied to the Expressionist protein quantification data in the *E. coli* benchmark dataset. False discovery proportion in the *y*-axis of panels C and D indicates the proportion of human proteins among the selected proteins at each score threshold.