A Bayesian Mixture Model for Comparative Spectral Count Data in Shotgun Proteomics

James G. Booth*§, Kirsten E. Eilertson#, Paul Dominic B. Olinares¶$, and Haiyuan Yu*

*Department of Biological Statistics and Computational Biology, Cornell University, Comstock Hall, Ithaca, NY 14853

#Department of Statistical Science, Cornell University, Malott Hall, Ithaca, NY 14853

¶Department of Chemistry and Chemical Biology, Baker Laboratory, Ithaca, NY, 14853

$Department of Plant Biology, Cornell University, Emerson Hall, Ithaca NY, 14853

§Corresponding author: 1178 Comstock Hall, Ithaca, NY 14853. Phone: 607-254-6505. Fax: 607-255-4698. Email: jim.booth@cornell.edu

Running title: Bayesian Model for Spectral Counts

Abbreviations:

SPCs: Spectral counts

LR: likelihood ratio

ML: maximum likelihood

ROC: receiver operating characteristic

CTAP: Clinical Proteomic Technology Assessment for Cancer

Sigma UPS1: Universal protein standard 1

MCMC: Markov Chain Monte Carlo

Summary

Recent developments in mass-spectrometry-based shotgun proteomics, especially methods using spectral counting, have enabled large-scale identification and differential profiling of complex proteomes. Most such proteomic studies are interested in identifying proteins, the abundance of which is different under various conditions. Several quantitative methods have recently been proposed and implemented for this purpose. Building on some techniques that are now widely accepted in the microarray literature, we developed and implemented a new method using a Bayesian model to calculate posterior probabilities of differential abundance for thousands of proteins in a given experiment simultaneously. Our Bayesian model is shown to deliver uniformly superior performance when compared with several existing methods.

Introduction

Mass-spectrometry-based shotgun proteomics has enabled large-scale identification and differential profiling of complex proteomes yielding significant insights into relevant biological systems (*1*). This approach typically involves liquid chromatography tandem mass spectrometry (LC-MS/MS) analysis and employs hybrid mass spectrometers with high data acquisition efficiency for intensity-based sampling of peptide ions (*1, 2*). The current quantification strategies for differential proteome analyses include the use of stable isotope-labeled reagents for chemical derivitization or metabolic labeling of protein samples (*3*). More recently, label-free techniques, such as peak intensity measurements and spectral counting, have emerged (*3*).

Spectral counting involves measuring the abundance of a given protein based on the number of tandem mass spectral observations for all its constituent peptides. Spectral counts (SPCs) have been shown to correlate well with the abundance of the corresponding protein extending over a linear dynamic range of at least two orders of magnitude for complex protein mixtures (*4-7*). SPCs can be readily extracted from the result files of all database search engines that are used for protein identification in shotgun proteomics analyses. As such, spectral counting is a flexible and straightforward technique. It thus offers a practical alternative to label-based quantification methods, which can be limited by high cost of reagents or incompatibilities with label incorporation. It is also a good substitute option for other label-free qualification methods such as peak intensity measurements, which relies heavily on computational efforts for chromatogram alignment and peak processing (*3*).

Maximizing the potential of spectral counting as a quantitative method has involved optimizations throughout the typical shotgun analysis workflow including sample preparation and fractionation,

instrument setup, data processing and statistical analysis. Intensity-based peptide sampling in shotgun LC-MS/MS is semi-random and depends largely on sample complexity, chromatographic separation and MS instrument parameters (*4*). Considerations on the impact of several of these factors to increase sampling depth have been studied (*6, 8*). Various schemes for counting matched spectra from database search results (*7, 9, 10*) as well as incorporation of additional information from including fragment ion MS/MS intensity and peptide count (*11*) and LC-MS peak area (*12*) have been explored. To more reliably reflect proteome abundances, appropriate transformations of raw SPCs have accounted for peptide length and total SPC within the sample (*13*) or probability of peptide detection (*14*). Statistical programs for significance analysis of spectral counting studies have also emerged and are based mainly on modeling the behavior of SPC datasets (*8, 15-19*).

More importantly, most proteomics studies are interested in finding proteins, the abundance of which changes in different cellular states, under different conditions, or with respect to different treatments. To this end, simple statistical methods have been employed to perform one protein at a time analysis using, for example, Wald or likelihood-ratio statistics. More recently, Choi et al. (*15*) implemented a Bayesian model (with an associated software, QSpec) in which all proteins are analyzed simultaneously with differential abundance for individual proteins identified using pseudo Bayes factors.

In this paper we propose an alternative Bayesian model for comparing spectral counts under two treatments/conditions. The model allows for simultaneous testing of several thousand proteins through the calculation of posterior probabilities of their null and non-null status, with proteins in the non-null group being those affected by the treatment. This two-group classification approach is analogous to widely accepted statistical methods for analyzing microarray data *(20, 21)*. The necessary computations are easily implemented via Markov chain Monte Carlo methods using the OpenBUGS software package *(22*). Furthermore, we show (see Results) that classification based on the Bayesian approach of Choi et

al. *(15)* is similar to the one-protein-at-a-time likelihood ratio test and substantially inferior in performance to posterior classification using our Bayesian model.

Experimental Procedures

Synthetic Yeast Proteome Dataset

We used the F2 synthetic dataset generated by Choi and colleagues *(15)* from a yeast shotgun proteomics analysis *(8)*. The yeast dataset consisted of proteins extracted from *Saccharomyces cerevisiae* strain BY4741 grown at middle log phase in media enriched in ^{14}N- or ^{15}N-labeled amino acids. Four independent cultures were grown in each medium type. 500 ug total protein from each growth condition were mixed in a 1:1 ratio resulting in four biological replicates. The resulting mixtures of ^{14}N- and ^{15}N-labeled proteins were then TCA-precipitated, urea-denatured, reduced, alkylated, and digested with Lys-C and then with trypsin. The extracted peptides were fractionated using a 12-step multidimensional protein identification technology (MUDPIT) setup and analyzed in an LTQ linear ion trap mass spectrometer (ThermoFinnigan) equipped with a nano-LC electrospray ionization source. Data-dependent acquisition settings include a full MS scan followed by CID fragmentation and MS/MS analysis of the five most abundant peptide ions with the following dynamic exclusion parameters: repeat count, 1; repeat duration, 30 s; exclusion duration, 300 s. Peak lists were obtained from RAW files using the extract_ms.exe program and were then searched using SEQUEST *(23)* with the appropriate mass modifications for ^{15}N-labeled peptides against a yeast protein sequence database appended with decoy sequences. DTASelect *(24)* was used to generate protein inventories with SEQUEST score filtering that yielded a false protein identification error rate of less than 1% (calculated based on decoy hits). 1307 proteins were identified at least once in the four biological replicates and the SPCs for these proteins were obtained from the DTASelect-filtered SEQUEST search results. To generate the F2 synthetic dataset *(15)*, the protein list in the original yeast dataset was randomized and the

abundance of the first 200 proteins was modified to reflect 2-fold changes between ^{14}N- and ^{15}N-labeled proteins. The 2-fold change was multiplied to the four replicates of ^{14}N-labeled proteins if the corresponding mean SPC was greater than the mean SPC for the four replicates of ^{15}N-labeled proteins and vice versa. For proteins having zero SPC in replicates belonging to the group with the smaller mean SPC, a randomly generated Poisson count was used with the resulting mean SPC being equal to the 2–fold change.

Human proteins spiked in yeast proteome background

The human-yeast proteome dataset was obtained from the analysis by Li and colleagues *(19)* of the dataset obtained from the Clinical Proteomic Technology Assessment for Cancer (CPTAC) Study 6 *(25)*. In this CPTAC study, a lyophilized yeast lysate (60ng/uL) was reconstituted with or without the addition of 48 human proteins (Sigma UPS1) that were spiked in varying amounts (0.25, 0.74, 2.2, 6.7 and 20 fmol/µL). We only used the datasets comparing the yeast reference proteome spiked with 6.7 and 2.2 fmol/ µL UPS1, which yielded a 3-fold difference in abundance. The resulting mixtures were reduced, alkylated, and digested with trypsin. Preparation and processing of these samples was performed centrally at the National Institute for Standards and Technology (NIST) and were distributed in various groups for MS analyses using various instruments as described in *(25)*. The dataset used here was derived from samples fractionated by reverse phase LC−MS/MS and analyzed in triplicate in one LTQ instrument (ThermoFinnigan) and on two LTQ-Orbitrap instruments (ThermoFinnigan) at Vanderbilt University. Data-dependent acquisition settings include a full MS scan in the LTQ for the standalone LTQ study or in the Orbitrap for the LTQ-Orbitrap instruments followed by CID fragmentation and MS/MS analysis of the eight most abundant peptide ions in LTQ in both instrument types. The following dynamic exclusion parameters were used: repeat count, 1 and exclusion duration, 60 s. For data processing and filtering *(19)*, the resulting Thermo RAW files were converted to the mzML format by the ProteoWizard

MSConvert tool *(26)* and searched using the Myrimatch *(27)* search algorithm against a yeast protein database with the 48 human protein and contaminant sequences as well as the corresponding reverse sequences. IDPicker *(28)* was employed to filter peptide matches to a 2% false discovery rate (FDR). All data from the three instruments were assembled into a single protein list requiring a minimum of 2 distinct peptides per protein. Only 46 out of the 48 human proteins were identified in the assembled dataset. Furthermore, the integration of the protein lists resulted in an increase in decoy hits (22% protein FDR) and an additional filter of five total SPC per protein was thereby imposed yielding a 6.8% protein FDR. The final dataset consisted of 46 human and 1342 yeast proteins (total of 1488 proteins).

Statistical Methods

Consider a data set consisting of spectral counts for p proteins in n replicates. Suppose that the replicates are either controls (e.g. wildtype) or from a treatment group. Let Y_{ij} denote the spectral counts for protein i in replicate j, and let T_j be a binary indicator for treatment. The objective of our analysis is to classify each protein as null or non-null with respect to the treatment.

A naive approach is to simply conduct one-at-a-time statistical tests on each protein. Since the responses are counts, a natural starting point for analysis is the log-linear model,

$$\log \mu_{ij} = \beta_{0i} + \beta_{1i} T_j + \log L_i + \log N_j, \qquad \text{(Eq.1)}$$

where μ_{ij} denotes the expected count for protein i in replicate j, and the offsets $\log L_i$ and $\log N_j$ respectively account for the length of the protein and the replicate effect. The hypothesis, $H_0: \beta_{1i} = 0$, represents the no treatment effect for protein i. Under the assumption that the counts are independent Poisson variables, this hypothesis can be assessed one protein at a time using a Wald or likelihood ratio (LR) test statistics,

$$W_i = \left|\frac{\hat{\beta}_{1i}}{\hat{\sigma}(\hat{\beta}_{1i})}\right|^2 \quad \text{and} \quad \lambda_i = -2\ln\frac{f(y_i;\hat{\mu}_i^{(0)})}{f(y_i;\hat{\mu}_i^{(1)})}, \qquad \text{(Eq.2)}$$

where $f(y_i;\mu_i)$ is the Poisson likelihood for the counts for protein i with fitted means $\hat{\mu}_i^{(k)}$, for $k = 0,1$, in the null and non-null cases respectively. Both Wald and LR require calculation of the ML estimates for the non-null model which can be obtained very quickly and efficiently, for example, using the **glm** function in R *(29)*, but involve an iterative fitting algorithm. In contrast, the score statistic *(30)* only involves the ML estimate under the null model which is available in closed form. In fact, it can be shown (see Supplemental Materials) that the score statistic for testing $H_0: \beta_{1i} = 0$ is given by

$$S_i = \frac{n\left[\sum_{j=1}^{n}\left(y_{ij} - \frac{N_j}{N}\bar{y}_i\right)T_j\right]^2}{\bar{y}_i\left(\sum_{j=1}^{n}\frac{N_j}{N}T_j\right)\left(\sum_{j=1}^{n}\frac{N_j}{N}(1-T_j)\right)}. \qquad \text{(Eq.3)}$$

These statistics, W_i, λ_i and S_i, are typically compared to a chi-squared distribution with 1 degree of freedom to determine significance, although with small sample sizes the chi-squared reference distribution might not be appropriate. Alternatively, to account for possible overdispersion with respect to Possion variation, one could conduct these tests under the assumption the counts are independent negative binomial variables with means given by the model given in equation 1 or use a quasilikelihood-based test *(19)*.

A Bayesian Model

The fact that there is only a small amount of data per protein suggests that power can be gained by borrowing strength across (the large number of) proteins. A general modeling strategy for can be achieved by formulating the problem in a Bayesian framework. Choi et al. *(15)* proposed an approach based on involving two Bayesian model fits, both requiring MCMC simulation, and implemented in a package they called QSpec. The first (full) model assumes the counts are conditionally independent Poisson variables with means given by the loglinear model:

$$\log \mu_{ij} = a_0 + b_{0i} + b_{1i}T_j + \log L_i + \log N_j,$$

with prior specification, $a_0 \sim N(0, \sigma_a^2)$, $b_{0i} \sim N(0, \sigma_0^2)$ and $b_{1i} \sim N(0, \sigma_1^2)$ independently, and hyperpriors $\sigma_0^{-2} \sim \text{gamma}(0.1,0.1)$ and $\sigma_1^{-2} \sim \text{gamma}(0.1,0.1)$. The second (restricted) model has the same form but omits the treatment effect term, $b_{1i}T_j$. Thus, the full model allows for a treatment effect for *all* proteins simultaneously, while the restricted model does not permit a treatment effect for any protein. Proteins are then classified as null or non-null on the basis of a *pseudo*-Bayes factor of the form

$$BF_i = \frac{f(y_i;\tilde{\mu}_i^{(1)})}{f(y_i;\tilde{\mu}_i^{(0)})}, \qquad \text{(Eq.4)}$$

where $\tilde{\mu}_i^{(k)}$ is the vector of means for protein i evaluated at the estimated posterior means of the regression parameters obtained from the full ($k = 1$) and restricted ($k = 0$) model fits. That is the BF-statistic is a function of both model fits and we note its similarity to the likelihood-ratio statistic in equation 2. We will refer to this as the pseudo-Bayes method in what follows.

A Bayesian Mixture Model

We now propose an alternative approach in which we formulate the problem as a Bayesian classification method. Specifically, we define I_i to be an indicator for non-null status of the ith protein and suppose that the indicators are independent Bernoulli(π_1) variables. We then propose to classify proteins as null or non-null according to the posterior odds

$$O_i = \frac{P(I_i=1|\text{data})}{P(I_i=0|\text{data})} \qquad \text{(Eq.5)}$$

for $i = 1, \ldots p$, with protein i classified as non-null if $O_i > c$ for a suitably large positive c. This "two-groups" mixture model approach is widely used and accepted in the microarray literature *(20,21,31,32)*, with the key difference being that the responses in the microarray context are continuous and often

modeled as (log) normal random variables. More generally, the inclusion of latent group indicators in the statistical model is a core component of Bayesian classification methods *(33)*.

The choice of the threshold c may be somewhat arbitrary. The modern statistical approach is to attempt to control the false discovery rate (FDR) *(34)*; i.e. the proportion of proteins classified as non-null for which there is in fact no treatment effect. In a recent paper *(21)* it is argued that FDR control can be achieved approximately using a posterior probability threshold and a value of 0.8 (or equivalently a posterior odds threshold of 4) is suggested for general use. However, in practice the choice of the threshold may be influenced by time and financial constraints on the number of follow-up experiments that are feasible.

In order to compute the posterior odds we consider the following modified version of model 1:

$$\log \mu_{ij} = \beta_0 + \beta_1 T_j + b_{0i} + b_{1i} I_i T_j + \log L_i + \log N_j. \qquad \text{(Eq.6)}$$

The linear predictor in equation 6 consists of β_0 and β_1, an overall mean for the control replicates and an overall treatment effect; b_{0i} and b_{1i}, the corresponding protein specific effects; and offsets $\log L_i$ and $\log N_j$.

Suppose that conditional on the means, μ_{ij}, the counts, Y_{ij}, are independent Poisson variables. Then the Bayesian model specification is completed by placing prior distributions the model parameters. Since π_1, β_0, and β_1 are global parameters we expect their posterior distributions to be relatively insensitive to the choice of prior. Hence, we use a uniform (Laplace) prior for the Bernoulli probability, π_1, and diffuse independent normal priors, $\beta_0 \sim N(0, 10^2)$ and $\beta_1 \sim N(0, 10^2)$, for the global regression coefficients. We consider three choices of prior distributions for the protein specific coefficients:

1. $(b_{0i}, b_{1i}) \sim N_2(0, \Sigma)$ independently for $i = 1, \ldots, p$, with $\Sigma^{-1} \sim \text{Wishart}(I, \upsilon)$, where I is the identity matrix and $\upsilon = 10$;

2. $b_{0i} \sim N(0, \sigma_0^2)$ and $b_{1i} \sim N(0, \sigma_1^2)$ independently for $i = 1, \ldots, p$, with $\sigma_0^{-2} \sim \text{gamma}(0.1, 0.1)$ and $\sigma_1^{-2} \sim \text{gamma}(0.1, 0.1)$ independently; and
3. $b_{0i} \sim N(0, \sigma_0^2)$ and $b_{1i} \sim N(\delta, \sigma_1^2)$ independently for $i = 1, \ldots, p$, with $\sigma_0^{-2} \sim \text{gamma}(0.1, 0.1)$ and $\sigma_1^{-2} \sim \text{gamma}(0.1, 0.1)$ independently, and $\delta \sim N(0, 10^2)$.

Model 1 allows for potential correlation between the protein specific coefficients, whereas models 2 and 3 assume they are independent. Model 3 allows the posterior mean of the protein specific treatment effects to be different in the null and non-null groups. This final modification is important (see Results) if the non-null proteins are predominantly more abundant in one of the treatment groups.

The most straightforward method of computing posterior probabilities of null and non-null status, and hence the posterior odds given in equation 5, is to simulate a Markov chain with a limiting distribution equal to the posterior distribution of the parameters and latent factors given the data. Specifically, after a suitable "burn-in" period, each successive iteration of the Markov chain can be regarded as a draw from the posterior distribution, and therefore posterior means (or probabilities, as in equation 5) can be computed as Monte Carlo averages. See *(35)* for a more detailed description of the theory behind MCMC methods. OpenBUGS *(22)* is an open source statistical package that implements MCMC methods for a large class of hierarchical Bayesian models that can be represented as directed acyclic graphs. The Bayesian models discussed in this paper are all of this type and therefore all the necessary computations can be carried out without the development of new, model-specific software.

Results

Figure 1 contrasts the performances of one-protein-at-a-time tests and the Bayesian methods discussed in the previous section based on their receiver operating characteristic (ROC) curves for the two publicly

available datasets described earlier. Figure 1A shows ROC curves for the synthetic dataset generated by Choi et al. *(15)* based on the yeast shotgun proteomics analyses performed by Pavelka et al. *(8)*.

Figure 1 here

One key finding is that the pseudo-Bayes method *(15)*, which identifies proteins that are differentially abundant in the two treatments using the BF-statistics given in equation 4, has similar performance to the one-protein-at-a-time score and likelihood-ratio tests. The poor performance of the Wald test with the synthetic 2-fold spiked dataset is not surprising because many of the proteins had very low SPC values and the standard error for the estimated coefficient $\hat{\beta}_{1i}$ is extremely unstable such cases. Our Bayesian model 3 uniformly dominates the one-at-a-time methods (and pseudo-Bayes) in both datasets. However, models 1 and 2, while essentially identical to model 3 in classifying the spiked proteins in the synthetic data from *(15)*, perform similarly to the one-at-a-time methods (and pseudo-Bayes) in the CPTAC human-yeast dataset. An explanation for the different performance of models 1 and 2 in the two data sets is that the 2-fold spiking of the SPCs in the synthetic data was done in approximately the same number of mutant samples as wild type (see Figure 2). For this reason, the posterior mean of the treatment effect is close to zero for both null and non-null proteins. In contrast the human proteins in the CPTAC dataset are all spiked higher in the D-samples. Thus, the posterior mean in the non-null group is positive, a possibility not allowed for in models 1 and 2.

Figure 2 here

Discussion

Strictly speaking Bayes factors are ratios comparing the marginal probability of the data under one model specification to another (*35*). In the context of shotgun proteomic studies with two conditions (e.g. wildtype and mutant) there are 2^p possible models, where p is the number of proteins, since each

protein can have either equal or differential abundance under the two conditions. The approach of Choi et al. *(15)* only considers two of these models, one in which differential abundance (non-null status) is allowed for *every* protein, and one in which there is no difference between the conditions for any protein. Thus, their protein-specific pseudo-Bayes factors cannot be interpreted in terms of marginalizing over all other proteins. In contrast, our Bayesian model essentially considers all 2^p possibilities simultaneously through the inclusion of latent indicators of null and non-null status for each protein. For this reason, we believe that our Bayesian mixture model, which leads to a simple classification scheme based on posterior probabilities or odds, is much more statistically coherent and defensible than an approach based on Bayes factors. As we noted in the Introduction, similar models are now widely accepted for the analysis of microarray data *(20,21)*. Finally, our approach is straightforward to implement using a widely-used (open source) software package, OpenBUGS *(22)*.

One minor drawback of the fully Bayesian mixture model analysis described in this paper is that it requires MCMC simulation for implementation and is therefore slower than the simple one-at-a-time methods (such as the score test which is virtually instantaneous). Even so, for a data set of the size described in the Results section ($n = 6$ or $8, p \sim 1000$), running three Markov chains of length 10,000 on a computer with an Intel Core 2 T9500 processor running at 2.60 MHz with 3.5 GB of RAM takes less than 20 minutes. This does not seem too big a price to pay given the far superior performance we have demonstrated.

References:

1. Domon, B., Aebersold, R. (2010) Options and considerations when selecting a quantitative proteomics strategy. *Nat Biotechnol* 28, 710-721
2. Domon, B., Aebersold, R. (2006) Mass spectrometry and protein analysis. *Science* 312, 212-217
3. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., Kuster, B. (2007) Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem* 389, 1017-1031
4. Liu, H., Sadygov, R. G., Yates, J. R., 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* 76, 4193-4201
5. Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G., Mendoza, A., Sevinsky, J. R., Resing, K. A., Ahn, N. G. (2005) Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics* 4, 1487-1502
6. Zhang, Y., Wen, Z., Washburn, M. P., Florens, L. (2009) Effect of dynamic exclusion duration on spectral count based quantitative proteomics. *Anal Chem* 81, 6317-6326
7. Cooper, B., Feng, J., Garrett, W. M. (2010) Relative, label-free protein quantitation: spectral counting error statistics from nine replicate MudPIT samples. *J Am Soc Mass Spectrom* 21, 1534-1546
8. Pavelka, N., Fournier, M. L., Swanson, S. K., Pelizzola, M., Ricciardi-Castagnoli, P., Florens, L., Washburn, M. P. (2007) Statistical similarities between transcriptomics and quantitative shotgun proteomics data. *Mol Cell Proteomics*,
9. Zhang, Y., Wen, Z., Washburn, M. P., Florens, L. (2010) Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. *Anal Chem* 82, 2272-2281
10. Zhou, J. Y., Schepmoes, A., Zhang, X., Moore, R. J., Monroe, M., Lee, J. H., Camp, D. G., Smith, R. D., Qian, W. J. (2010) Improved LC-MS/MS Spectral Counting Statistics by Recovering Low Scoring Spectra Matched to Confidently Identified Peptide Sequences. *J Proteome Res*,
11. Griffin, N. M., Yu, J., Long, F., Oh, P., Shore, S., Li, Y., Koziol, J. A., Schnitzer, J. E. (2010) Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat Biotechnol* 28, 83-89
12. Dicker, L., Lin, X., Ivanov, A. R. (2010) Increased power for the analysis of label-free LC-MS/MS proteomic data by combining spectral counts and peptide peak attributes. *Mol Cell Proteomics*,
13. Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K., Florens, L., Washburn, M. P. (2006) Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J Proteome Res* 5, 2339-2347
14. Lu, P., Vogel, C., Wang, R., Yao, X., Marcotte, E. M. (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* 25, 117-124
15. Choi, H., Fermin, D., Nesvizhskii, A. I. (2008) Significance analysis of spectral count data in label-free shotgun proteomics. *Mol Cell Proteomics* 7, 2373-2385
16. Carvalho, P. C., Fischer, J. S., Chen, E. I., Yates, J. R., 3rd, Barbosa, V. C. (2008) PatternLab for proteomics: a tool for differential shotgun proteomics. *BMC Bioinformatics* 9, 316
17. Heinecke, N. L., Pratt, B. S., Vaisar, T., Becker, L. (2010) PepC: proteomics software for identifying differentially expressed proteins based on spectral counting. *Bioinformatics* 26, 1574-1575
18. Pham, T. V., Piersma, S. R., Warmoes, M., Jimenez, C. R. (2010) On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. *Bioinformatics* 26, 363-369

19. Li, M., Gray, W., Zhang, H., Chung, C. H., Billheimer, D., Yarbrough, W. G., Liebler, D. C., Shyr, Y., Slebos, R. J. (2010) Comparative shotgun proteomics using spectral count data and quasi-likelihood modeling. *J Proteome Res* 9, 4295-4305
20. Bar, H., Booth, J., Schifano, E., Wells, M. T. (2010) Laplace Approximated EM Microarray Analysis: An Empirical Bayes Approach for Comparative Microarray Experiments. *Stat Sci* 25, 388-407
21. Efron, B. (2008) Microarrays, empirical Bayes and the two-groups model. *Stat Sci* 23, 1-22
22. Lunn, D., Spiegelhalter, D., Thomas, A., Best, N. (2009) The BUGS project: Evolution, critique and future directions. *Stat Med* 28, 3049-3067
23. Eng, J. K., Fischer, B., Grossmann, J., Maccoss, M. J. (2008) A fast SEQUEST cross correlation algorithm. *J Proteome Res* 7, 4598-4602
24. Tabb, D. L., McDonald, W. H., Yates, J. R., 3rd (2002) DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* 1, 21-26
25. Paulovich, A. G., Billheimer, D., Ham, A. J., 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. (2010) Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol Cell Proteomics* 9, 242-254
26. Kessner, D., Chambers, M., Burke, R., Agus, D., Mallick, P. (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* 24, 2534-2536
27. Tabb, D. L., Fernando, C. G., Chambers, M. C. (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* 6, 654-661
28. Ma, Z. Q., Dasari, S., Chambers, M. C., Litton, M. D., Sobecki, S. M., Zimmerman, L. J., Halvey, P. J., Schilling, B., Drake, P. M., Gibson, B. W., Tabb, D. L. (2009) IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res* 8, 3872-3881
29. R Development Core Team (2010) *R: A language and environment for statistical computing*, (R Foundation for Statistical Computing, Vienna, Austria).
30. Cox, D. R., Hinkley, D. V. (1979) *Theoretical Statistics*, (Chapman & Hall/CRC), pp. 315.
31. Lonnstedt, I., Speed, T. (2002) Replicated microarray data. *Stat Sinica* 12, 31-46
32. Symth, G. (2004) Linear models for empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3, 2
33. Berger, J. O. (2006) *Statistical Decision Theory and Bayesian Analysis*, (Springer, ed. 2nd).
34. Benjamin, Y., Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* 57, 289-300
35. Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (2004) *Bayesian Data Analysis*, (Chapman & Hall/CRC), pp. 184.

Legends:

Figure 1: ROC plots for one protein at a time Wald, score and likelihood ratio tests, posterior odds derived from Bayesian models 1-3, and pBayes *(15)* for A) the 2-fold spiked synthetic dataset from Choi et al. *(15)* and B) the CPTAC Human-Yeast dataset from Paulovich *(24)*.

Figure 2: Abundance rates in the two treatment groups for the synthetic 2-fold spiked data *(15)* and the CPTAC human-yeast data *(19)*. The abundance rate is calculated as $\bar{Y}/(L\bar{N})$, where $\bar{Y}$ is the sample mean SPC, L is the protein length, and $\bar{N}$ is the mean SPC overall all samples in the treatment group.

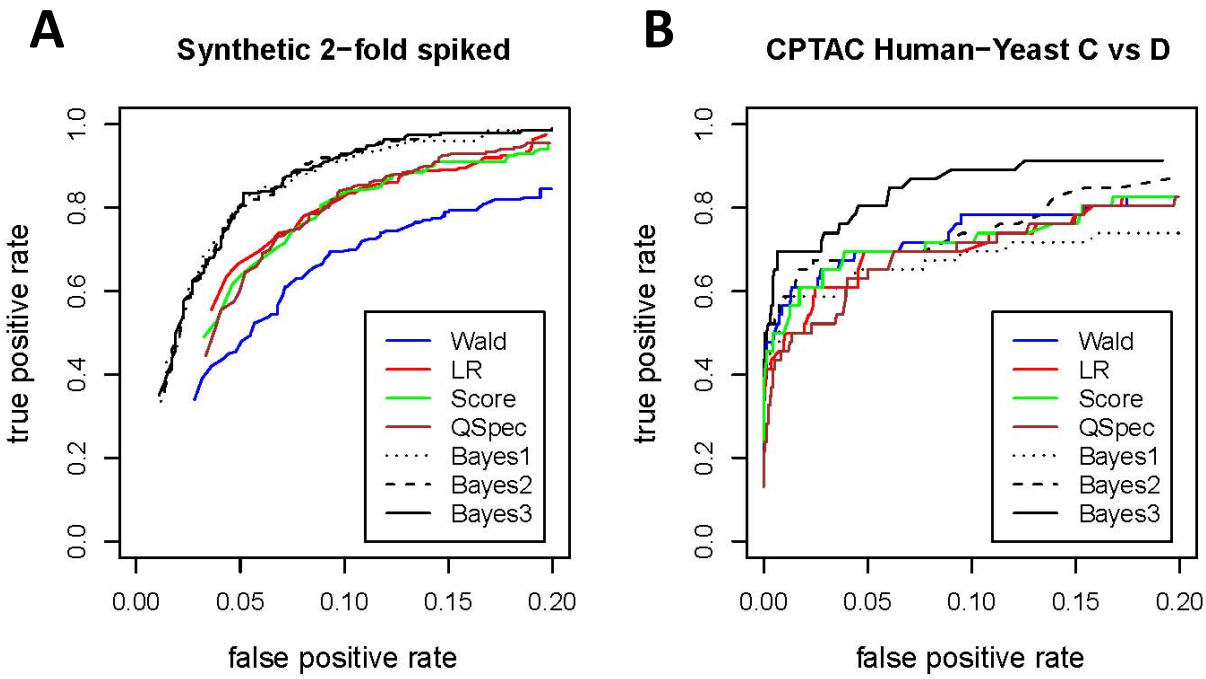
A
Synthetic 2−fold spiked
true positive rate
false positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
0.20
Wald
LR
Score
QSpec
Bayes1
Bayes2
Bayes3
B
CPTAC Human−Yeast C vs D
true positive rate
false positive rate
Wald
LR
Score
QSpec
Bayes1
Bayes2
Bayes3

Figure 1:

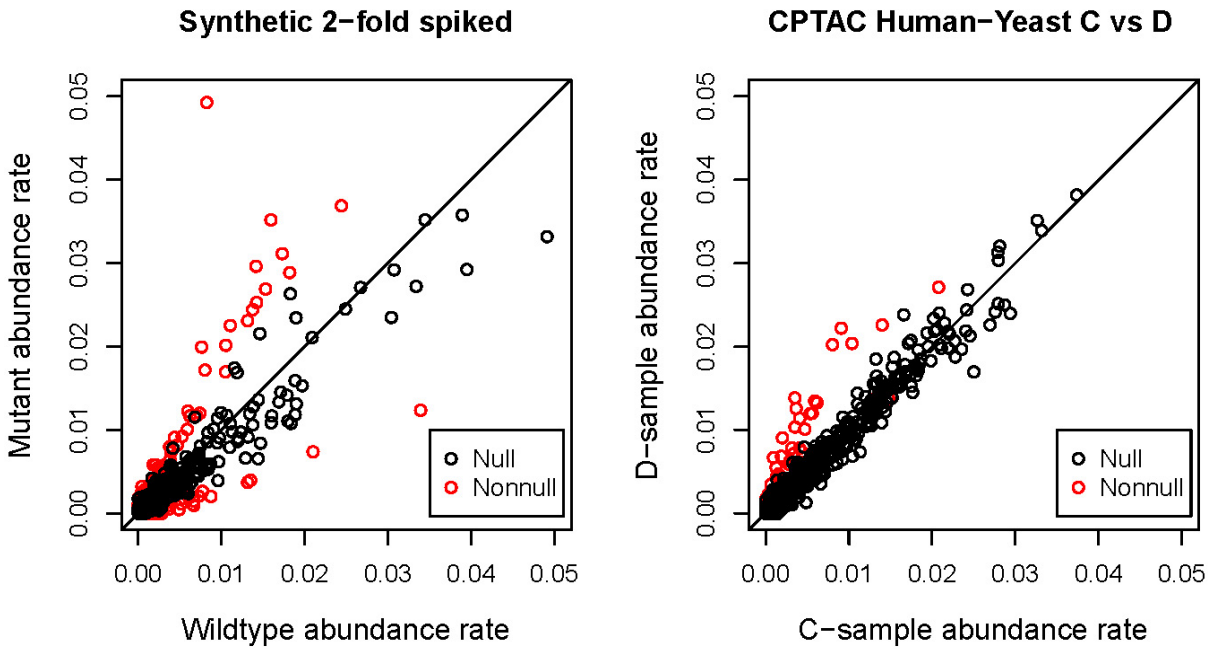
Synthetic 2−fold spiked
Mutant abundance rate
Wildtype abundance rate
0.00
0.01
0.02
0.03
0.04
0.05
Null
Nonnull
CPTAC Human−Yeast C vs D
D−sample abundance rate
C−sample abundance rate
0.00
0.01
0.02
0.03
0.04
0.05
Null
Nonnull

Figure 2: