Statistics and Its Interface Volume 5 (2012) 89–98

Generalized linear and mixed models for label-free shotgun proteomics

Matthew C. Leitch[*], Indranil Mitra[*] and Rovshan G. Sadygov[†]

Label-free shotgun proteomics holds great promise, and has already had some great successes in pinpointing which proteins are up or down regulated in certain disease states. However, there are still some pressing issues concerning the statistical analysis of label-free shotgun proteomics, and this field has not enjoyed as much dedication of statistical research towards it as microarray research has. Here we reapply previously used statistical methods, the QSpec and quasi-Poisson, as well as apply the negative binomial distribution to both a control data set and a data set with known differential expression to determine the successes and failure of each of the three methods.

Keywords and phrases: Count data, Statistical models, Spectral count, Label-free quantitative proteomics, p-values, FDR, Negative binomial model, Quasi-Poisson model, Mixture model, Generalized linear models.

1. INTRODUCTION

The on-line linking of liquid chromatography with tandem mass spectrometry enabled the development of shotgun proteomics (Link et al. 1999). It includes digesting a protein sample into peptides, and separating them with two liquid column chromatography steps. Stable-isotope labeling based methods of shotgun proteomics soon gained ground, such as ICAT (Gygi et al. 1999), SILAC (Ong et al. 2002) and iTRAQ (Ross et al. 2004). However, they require a higher amount of startup material, complex experimental protocols, and have higher reagent costs, as well as one must hope that the protein is not altered in any manner by the labeling.

The advent of the label-free methods have circumvented some of the issues of labeling methods (Bondarenko et al. 2002; Chelius, Bondarenko 2002). Spectral count methods became a popular and successful method for relative protein quantification using mass spectrometry data (Liu et al. 2004; Old et al. 2005). Determination of differential protein expressions in real biological samples is a very complex art, with many different complications that must be considered. Proteins from within the same system need to perform different tasks, leading to substantial differences in function, and therefore structure. The effects of digestion can be significantly different between proteins of differing structures. Not all proteins ionize in a similar manner, affecting their ability to be detected by the mass spectrometer (Vogel, Marcotte 2008). There is also the issue of size bias, since larger proteins are more likely to appear in spectral counts than smaller proteins due to the fact that they are on average creating a larger amount of peptides than their smaller counterparts (Paoletti et al. 2006). There are usually a small number of replicates, a problem also found within gene microarray data that is referred to as the "many genes, few replicates" problem (Pavelka et al. 2008).

Recently there have been several contributions to the development of statistical models specific to spectral count data. A global error model assumes that the amount of variability is a function of the measurement levels within all the measurements for a single experimental condition (Pavelka et al. 2004). The advantage of making this assumption is that the number of measurements used to estimate the global error is equal to the total number of genes or proteins (Pavelka et al. 2004). The power law global error model (PLGEM) is a global error model governed by a power law, where the frequency of an event varies as a power of some attribute of that event (Pavelka et al. 2004). This type of model has been successfully used with gene microarray data (Pavelka et al. 2004), with normalized spectral abundance factors to normalize for the size bias (Paoletti et al. 2006; Griffin et al. 2010; Zybailov et al. 2006; Zybailov et al. 2011). Pavelka and colleagues (Pavelka et al. 2008) showed that the coefficient of variation becomes smaller as the average abundance values increase, even though the absolute standard deviation still increases. They demonstrated that their parameters were still stable even under decreasing replicate numbers, which is an issue in shotgun proteomics. However, this model does not produce direct p-values to ease interpretations. Using Pavelka et al. 2008's data, Choi and colleagues created a Poisson model with a hierarchical Bayesian estimation referred to as QSpec (Choi et al. 2008). Spectral counts are considered random numbers amongst a large population of proteins, and therefore all replicates (and proteins) share model parameters. The per protein basis helps with the low replicate numbers (Choi et al. 2008). The Poisson model treats variance as equal to mean, which may be problematic as over-dispersion is often observed in spectral count data. Li and colleagues (Li et al. 2010) attempted to circumvent

[*] These authors have contributed equally to this work.
[†] Corresponding author.

this problem by using a quasi-likelihood generalized linear model. This quasi-Poisson model allows over-dispersion, using the F-test to calculate p-values and the false discovery rate (FDR) to rectify the effects of multiple hypothesis testing (Li et al. 2010). Asymptomatic properties of maximum likelihood models are minimized, dispersion is treated as a free parameter, regression coefficients are given even if the variance function is not specified, and a clear idea of the distribution is not needed beforehand to model. This was compared to the student's t-test and Fisher's exact test (Li et al. 2010).

Here we use existing quasi-Poisson and QSpec models, and test a new, negative binominal model to improve statistical understanding of label-free shotgun proteomics data. Over-dispersion is still an issue despite the successes of quasi-Poisson, and zero entries create problems for statistical models. A model that can handle small sample sizes and different variances is expected to be better suited for describing shotgun proteomics data.

2. DATASETS

In this work we used two freely available datasets from label-free shotgun proteomics studies. The first dataset is of the yeast *Saccharomyces cerevisiae* strain BY4741, which was obtained from the work of Pavelka and colleagues (Pavelka et al. 2008). This dataset constitutes a control or test dataset. The protein concentrations in the two samples are expected to be identical. The dataset is used to detect false positives produced by the statistical modeling. We call this dataset as a control dataset (isotope-labeled proteins are mixed at 1:1 ratio with unlabeled proteins). The isotopically labeled and unlabeled parts of the control dataset are called N15 and N14 datasets.

The second dataset was obtained from 20 head and neck squamous cell carcinomas (HNSCC) and 20 normal tonsillectomy tissues (Li et al. 2010). The proteins in the two states are expected to exhibit differential expressions and thereby it is possible to determine which proteins are up-regulated or down-regulated in the cancerous cells compared to the normal cells, hoping to point out which pathways are involved in head and neck squamous cell carcinomas.

The benchmarking of the statistical models on two different datasets, with distinct characteristics, is a good test platform for comparing quantification models of shotgun proteomics data. It is widely believed that the Poisson family of models provides a good prediction of the data. However, in scenarios with scarce data as is the case with spectral count data, the issue is far from clear. This is similar to the assumptions often made in applications such as the serial analysis of gene expression, SAGE (Huang et al. 2008). Negative binomial models have been particularly successful in explaining over-dispersed data. Ultimately choosing among these methods or if needed to devise a new strategy is a model selection problem (Neal, Simons 2007).

3. QUASI-POISSON, NEGATIVE BINOMIAL AND QSPEC MODELS

3.1 Problem statement

Spectral counts of proteins are indicative of protein abundance levels in a sample. Normally there are two groups, control and disease (or treatment), and the goal is to design an adequate model for assigning statistical significance levels to differences in observed protein spectral counts. The dependent variable is assumed to be the spectral count (Y) and the independent variable is the group (a categorical variable). The inference from the model will be obtained by comparing full (two groups) and reduced (a single group) models. Significance of individual protein differences is evaluated by the p-values from the inference. The p-values are adjusted for multiple testing, for example, by Benjamini-Hochberg procedure (Benjamini, Hochberg 1995; Hochberg, Benjamini 1990) which produces false discovery rates. Generalized linear models (GLM) with distributions describing count data are natural candidates for the outlined framework. A preferred distribution for spectral counts was Poisson distribution. However, equal mean variance assumption of the Poison model is not supported in the spectral count datasets (Pavelka et al. 2008). Therefore, more flexible models are used as discussed below.

The outcome of the analysis is a list of differentially expressed proteins between two samples. The list is used as an input into biological knowledge bases, such as FatiGo, Davids or IPA (Al-Shahrour et al. 2004; Shah et al. 2010; Schulz-Trieglaff et al. 2009). Information about affected signaling pathways, protein interaction networks, protein localizations and functionality are used in a systems biology view of a disease state, progression, efficacy of treatment, etc.

3.2 Maximum likelihood for quasi-Poisson and negative binomial models

The standard Poisson model has difficulty handling the over-dispersion of the variance found in shotgun proteomics data, and the quasi-likelihood model lets dispersion remain a free parameter. This quasi-Poisson likelihood model only claims to know the first and second moments, since trying to claim spectral counts fit a known distribution is unrealistic. The likelihood model offers more parameters and this extra flexibility is what gives the model much more strength compared to the standard Poisson model. The quasi-Poisson model has been previously used in SAGE data (Cai et al. 2004). In general, quasi-likelihood estimation is one way of allowing a greater variability in the data. However, a practical problem with Poisson regression is observed when the variance of the spectral counts is greater than the mean (over-dispersion). Though quasi-Poisson gives a better account of the variance compared to the original Poisson, still the quasi-Poisson is a part of the Poisson family. The negative binomial regression model is also a subclass of the exponential family of generalized linear models and in some cases it is a good predictor for over-dispersed data. The negative-binomial model deals with this problem by introducing noise into the linear predictor. We describe here briefly the two models from a unified perspective, both being a class of GLM's and amenable to maximum likelihood estimations.

The spectral counts, Y_i, of a protein are expressed as:

$$\ln(Y_i) = \ln(N_i) + \beta_{0i} + \beta_{1i} X_j + \varepsilon_i \tag{1}$$

where X_j denotes group, N_i is the total number of counts in replicate i and ε_i is the error. The group variable is a categorical variable with values 1 (for example control) and 0 (treatment). In addition we have the following constraints satisfied as

$$\begin{aligned} &E[Y_i] = \mu_i \quad \text{and} \quad Var(Y_i) = \varphi V(\mu_i), \\ &\quad V(\mu_i) = \mu_i \text{ with } Y_i \sim Poi(\mu_i, \varphi) \qquad \text{(quasi-Poisson)} \\ &E[Y_i] = \mu_i \quad \text{and} \quad Var(Y_i) = V(\mu_{ij}) = \mu_i + \alpha\mu_i^2, \\ &\quad Y_i \sim NB(\mu_i, \alpha), \qquad \text{(negative binomial)} \end{aligned} \tag{2}$$

where φ is the over-dispersion parameter and α is the clumping parameter in each case. The standard procedure in handling these exponential families is to define a score

$$U_i = \frac{Y_i - \mu_i}{Var(Y_i(\mu_i))} \tag{3}$$

The score satisfies the properties of a log likelihood function, namely, $E[U_i] = 0$ and $E[\frac{\partial U_i}{\partial \mu_i}] = V(U_i)$. So it is quite natural to construct the quantity

$$Q = \sum_{i=1}^{N} \int_{Y_i}^{\mu_i} \frac{Y_i - x}{Var(Y_i(x))} dx \tag{4}$$

which has equivalence to log likelihood which is then maximized to obtain the estimate of the regression coefficients (Li et al. 2010). Usually the procedure involves using iterative weighted least squares (IWLS), where

(5)

$$\begin{aligned} &\beta^{k+1} = (X' W_i^k X)^{-1} X' W^k \hat{Y}, \\ &\quad W_k = \operatorname{diag}\Big(\frac{\mu_1}{\theta}, \ldots, \frac{\mu_n}{\theta}\Big) \qquad \text{(quasi-Poisson)} \\ &\beta^{k+1} = (X' W_i^k X)^{-1} X' W^k \hat{Y}, \\ &\quad W_k = \operatorname{diag}\Big(\frac{\mu_1}{1+\alpha\mu_1}, \ldots, \frac{\mu_n}{1+\alpha\mu_n}\Big) \quad \text{(negative binomial)} \end{aligned}$$

It is clear that the variance of the outcome in the negative binomial model is greater than or equal to that for the Poisson regression model. The hypothesis of over-dispersion can be tested either by testing for evidence that $\alpha > 0$ against the null hypothesis that $\alpha = 0$, or by constructing a likelihood-ratio test comparing the Poisson and negative binomial regression models. In comparison to the quasi-Poisson model the over-dispersion in the negative binomial case is a multiplicative factor $1 + \alpha\mu$, showing that for the Poisson (in this case the quasi-Poisson model), the variance is linearly related to the mean, whereas for the negative binomial, the variance is quadratic in mean. Thus, in contrast to the quasi-Poisson where weights are directly proportional to the mean, for the negative binomial the weights have a concave relationship with the mean, which implies that very small means get little weight, while as the mean increases, weights level off to $1/\alpha$.

In general for models with a known dispersion the chi-squared test is most appropriate, and for those with dispersion estimated by moments and fits, the F test is most appropriate (Ver Hoef, Boveng 2007). However in our case the dispersion is not known a priori and is allowed to be a free parameter and ANOVA tests are used to generate the p-values for the case based on two fitted models, one with no distinction of the groups and in the other, where there are two explicit groups. However, in these cases we also have to compare many proteins for differences, for which we require multiple testing and thereby to control the FDR and adjust the obtained p-values.

3.3 QSpec statistical model

This is a mixed effects statistical model which uses hierarchical Bayes factor for taking into account the small number of replicates in the data, which is achieved by pooling the information on regression models across the proteins (Choi et al. 2008). The model equations get modified in this case to

$$\ln(Y_{ij}) = \ln(L_i) + \ln(N_j) + c_0 + \beta_{0i} + \beta_{1i} X + \varepsilon_i \tag{6}$$

where the changes have been to include the length L_i of the proteins and an introduction of a baseline abundance c_0. As a matter of fact the counts here too are assumed to have a Poisson distribution similar to the quasi-Poisson model, however the equation (6) gives a full model M_F as compared to the reduced model M_R when the treatment effect β_{1i} is not significant. The key difference in this case is the fact that the regression parameters are assumed to have prior normal distribution, along with an inverse γ distributed variance parameters, which gives rise to a mixed effects model. The Bayes factor and the FDR values are calculated as follows

(7)

$$B_i = \frac{p(X_i|M_F)}{p(X_i|M_R)}, \qquad \mathrm{FDR}(B_i) = \frac{\pi_0 p_0(B_i)}{\pi_0 p_0(B_i) + \pi_1 p_1(B_i)}$$

where p_0 and p_1 are the proteome wide distributions of the Bayes factor for proteins with trivial and significant differential expression respectively and π_0 and π_1 are the corresponding proportion of the proteins.

We have modified the Quasi-tel program (Li et al. 2010) in R to generate the quasi-Poisson and negative binomial models (attached in the supplementary infor-

Table 1. The top five proteins by FDR adjusted p-value are listed for each of the three methods for the control dataset

Method	Protein	Length	N14	N15	Quasi-Poisson	Neg. binomial	QSpec
QSpec	gi|6323768	592	$(98, 66, 73, 75)$	$(59, 17, 59, 54)$	0.74	0.019	0
	gi|6324534	106	$(12, 6, 9, 5)$	$(17, 7, 67, 36)$	0.784	0.035	0
	gi|6324027	144	$(36, 16, 6, 19)$	$(38, 30, 34, 49)$	0.74	0.02	0
	Cont_gi|7463016	269	$(56, 52, 83, 26)$	$(0, 0, 0, 0)$	0.74	0.999	0
	gi|6323104	221	$(15, 18, 4, 4)$	$(12, 12, 3, 3)$	0.993	0.333	0.114
negative binomial	gi|6323470	424	$(99, 64, 32, 5)$	$(64, 30, 41, 2)$	0.993	0.004	0.266
	gi|6323768	592	$(98, 66, 73, 75)$	$(59, 17, 59, 54)$	0.74	0.019	0
	gi|6324027	144	$(36, 16, 6, 19)$	$(38, 30, 34, 49)$	0.74	0.02	0
	gi|6322075	385	$(21, 1, 0, 0)$	$(4, 0, 0, 2)$	0.993	0.021	1
	gi|6324534	106	$(12, 6, 9, 5)$	$(17, 7, 67, 36)$	0.784	0.035	0
quasi-Poisson	gi|6323768	592	$(98, 66, 73, 75)$	$(59, 17, 59, 54)$	0.74	0.019	0
	gi|6324027	144	$(36, 16, 6, 19)$	$(38, 30, 34, 49)$	0.74	0.02	0
	gi|37362683	377	$(0, 0, 0, 0)$	$(0, 3, 1, 1)$	0.74	0.999	0.285
	gi|6320353	2748	$(18, 4, 0, 0)$	$(0, 0, 0, 0)$	0.74	0.997	0.326
	gi|6319520	334	$(2, 0, 0, 0)$	$(6, 0, 3, 3)$	0.74	0.08	0.421

mation http://www.intlpress.com/SII/p/2012/5-1/SII-5-1-leitch-supplement.pdf). The QSpec method was run on a dedicated server at the University of Michigan. Finally we generate all the plots and tables in R and python matplot.

4.1 Applications of the three methods to the control dataset

We have applied all methods to a control dataset from the Pavelka et al. 2008 paper. The dataset contains 1,314 proteins, reduced to 1,270 when proteins flagged by the QSpec software were removed. This dataset makes an excellent control since it should have no differentially expressed proteins and therefore none of our three methods should show any significant results.

It is established from Table 1 and fully demonstrated in the full data (available within the supplementary materials http://www.intlpress.com/SII/p/2012/5-1/SII-5-1-leitch-supplement.pdf) that there is no differentially expressed proteins in the control dataset once p-values are adjusted by the false discovery rate. This is expected since there is no true difference between the datasets, and any difference in expression should be due to chance or contamination and/or error. But even here we can begin to see in the quasi-Poisson, there is already a favoring of proteins that have zero variance among their spectral count replicates. The QSpec method's high number of zeros as p-values indicates that there could be some false positives in there, and it is interesting that a possible contaminant was ranked so highly.

Figure 1 demonstrates the relationship between the means of the N14 and N15 labeled samples. An interesting observation made from Figure 1 is that there is not an even spread of data points between the N14 and N15 datasets, with a large bias towards N15. This implies there is a systematic bias within the system. This bias is also very obvious from Table 1, as normal datasets have higher spectral count numbers than cancerous datasets. This will be shown later to exist as well within the HNSCC dataset. One advantage that the negative binomial, quasi-Poisson and QSpec methods have over a simple student's t-test is that they compensate for biases like this by taking the logarithm of the total number of proteins as in the case of the Poisson based models or adjust for the offset as in the case for negative binomial.

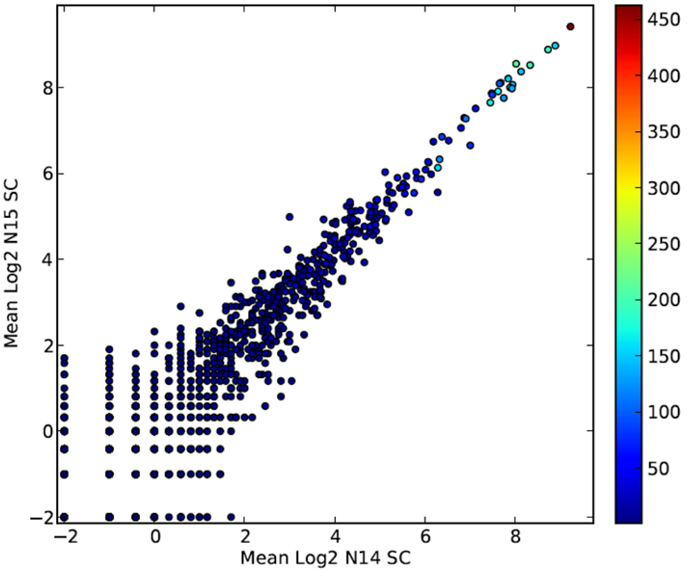

Figure 1. Graph for the log base2 of the mean values for the N14 dataset plotted versus the log base2 of the mean values for the N15 dataset. The variance is demonstrated by the coloration of the data points, with red data points having greater variance than blue data points.

In Figure 2 we have shown how the mean behaves with the variance in a control dataset for both controls and it is clear that the behavior is as expected for a non-differentially expressed dataset.

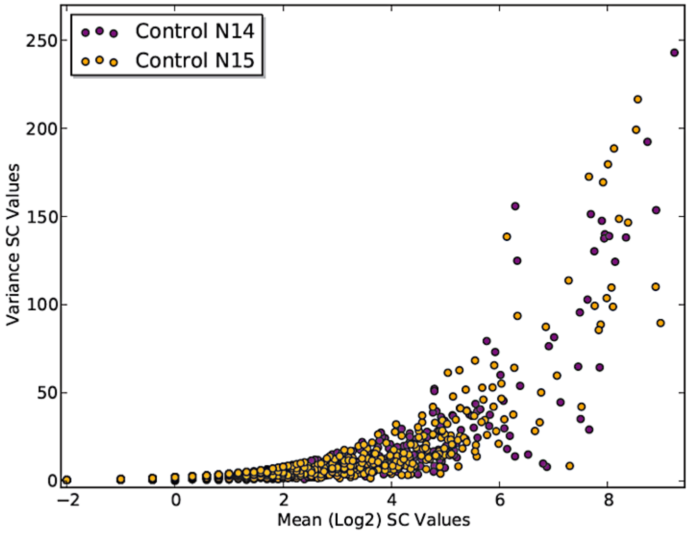

Figure 2. The mean spectral count against the variance for the control dataset. Yellow is N15 dataset, for which there is a clear bias towards the mean compared to the N14 (purple) dataset.

4.2 Comparison of the three methods on the HNSCC dataset

The main focus of our study was the application of the three statistical methods to the head and neck squamous cell carcinoma dataset obtained from the Li et al. 2010 paper. This dataset has been previously shown to have differentially expressed proteins (Li et al. 2010), and shall serve as the testing grounds for each of the three methods. The dataset had a total of 1,713 proteins, which was reduced to 1,617 once proteins that are flagged by the QSpec method were removed.

There are differentially expressed proteins in this dataset. There is some overlap, generally different proteins made it into the top 5 for each of the methods, with quasi-Poisson sharing none with the other methods, and only two of the top five are shared between negative binomial and QSpec. We can see that quasi-Poisson labels the greatest number of proteins as differentially expressed, while QSpec labels less than half that many as differentially expressed, and the negative binomial even fewer than that. This also shows that quasi-Poisson is only one of the three methods to recognize a great number of its top 100, while most of QSpec and the negative binomial's top 100 proteins are shared with each other.

Figure 3 is the Venn diagram of the top 150 proteins of each method (based on FDR adjusted p-values). FDR adjusted quasi-Poisson recognizes 158 proteins out of 1,617 as differentially expressed at a level of 0.01 and 167 at level of 0.1. QSpec recognizes 130 proteins out of 1,617 as differentially expressed at a level of 0.01 and 167 at a level of 0.1. FDR adjusted negative binomial recognizes none of the proteins out of 1,617 as differentially expressed at a level of 0.01, and 6 at a level of 0.1.

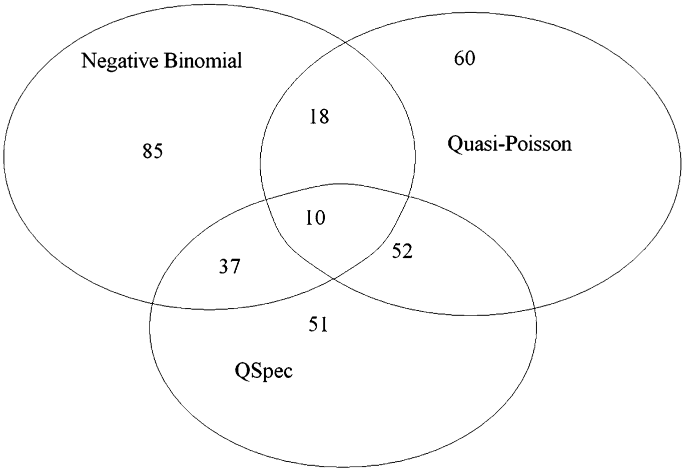

Figure 3. Venn diagram which shows how many proteins in the top 150 from the HNSCC dataset are shared by the 3 different statistical methods. Ranking was based on FDR values.

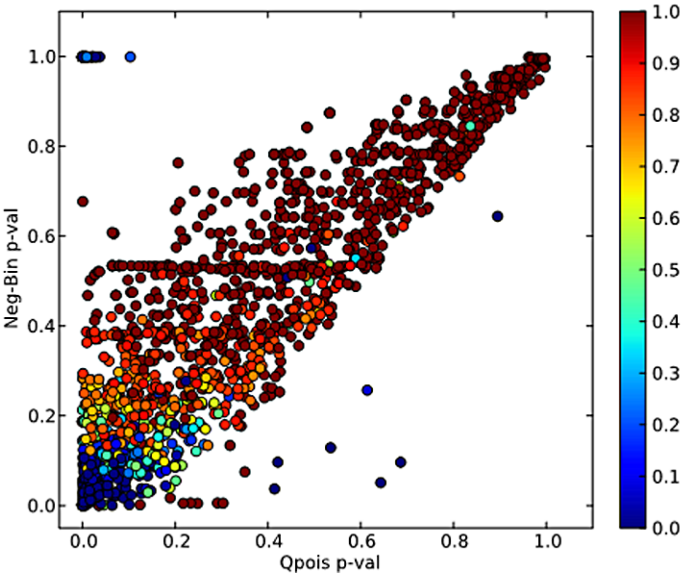

Figure 4. Comparisons of the p-values for quasi-Poisson and negative binomial methods for the HNSCC dataset with QSpec FDR values as the color label.

Figure 4 shows the relationship of the p-values of the quasi-Poisson *vs* negative binomial methods with data points color-based on the FDR values from the QSpec method. The blue color shows low FDR values and red implies higher FDR values (as computed by QSpec). It is seen that QSpec is a comparatively conservative method, as it gives proteins high p-values where the other two methods are rewarding these same proteins with low p-values.

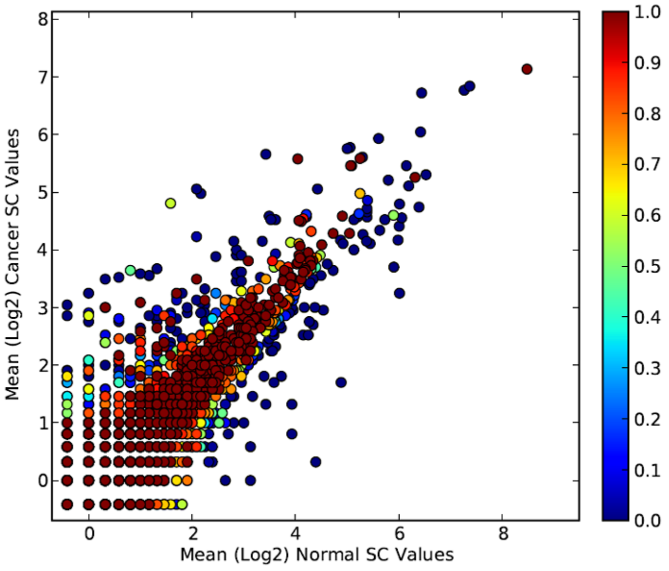

Figure 5. Log2 mean-mean plot for the cancer vs normal spectral count for the proteins in the HNSCC dataset. The coloring of the points is based on the FDR values of the QSpec method.

Also the under-dispersion in the HNSCC dataset may imply that there exists a negative correlation in the count data set, which is there, and the Pavelka et al. 2008 data on the other hand clearly shows a clustering based on the over-dispersion.

4.3 Evaluation of QSpec method on the HNSCC dataset

We can see that most of the data points with low p-values are spread out near the highest means, and that there are more high p-values found within the normal data. There is a general bias of spectral counts towards the normal data, as Figure 5 is skewed toward the right side (the normal side). As expected, proteins with greater means and greater variances are assigned with lower p-values. The coefficient of variance for cancerous data is much more spread out than that of the normal dataset.

4.4 Evaluation of quasi-Poisson method on the HNSCC dataset

We have here the p-values from the quasi-Poisson model. We can see that generally the data points with low p-values are spread out near the highest means, and that there are more high p-values found within the normal data. However, this trend is less pronounced than in QSpec, with many proteins with high means given high p-values. However, proteins with greater means and greater variances are not necessarily rewarded with lower p-values. The coefficient of variance for a cancerous dataset is much more spread out than in the normal dataset, but to a lesser effect than QSpec. In this case we see also a trend in the Log2 mean-mean plot for the normal and cancer spectral counts based on the quasi-Poisson FDR values from Figure 6.

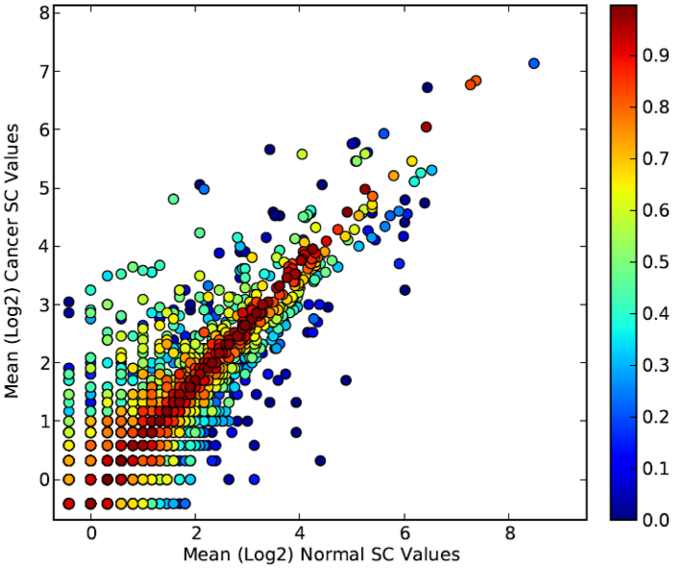

Figure 6. Log2 mean-mean plot for the cancer vs normal spectral count for the proteins in the HNSCC dataset. The coloring of the points is based on the FDR values of the quasi-Poisson method.

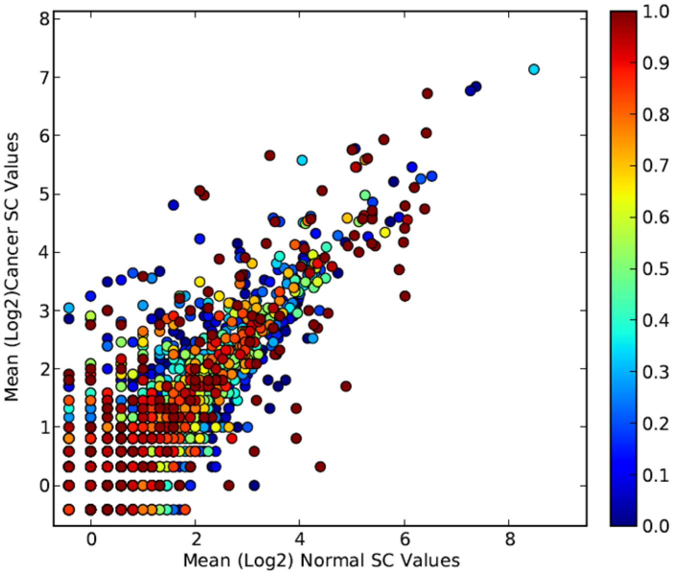

Figure 7. Log2 mean-mean plot for the cancer vs normal spectral count for the proteins in the HNSCC dataset. The coloring of the points is based on the FDR values of the negative binomial method.

4.5 Evaluation negative binomial method on the HNSCC dataset

We have here the p-values from the negative binomial model, Figure 7. The figure shows the log2 mean compar-

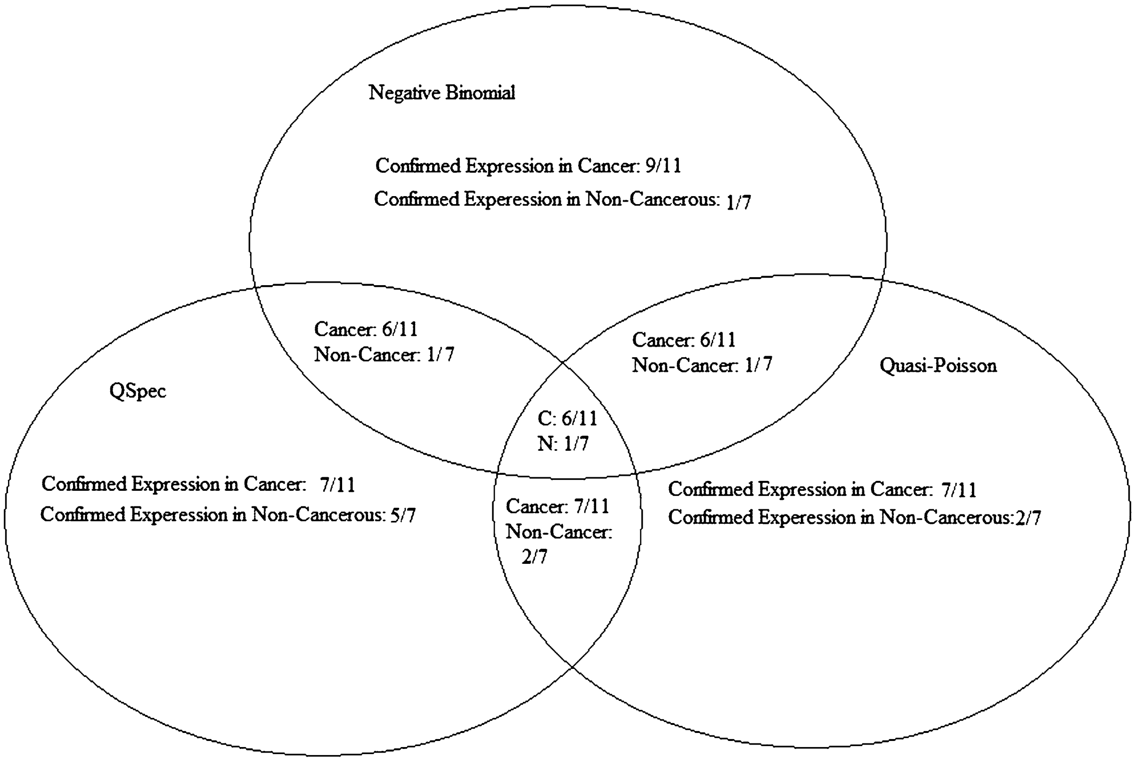
Negative Binomial
Confirmed Expression in Cancer: 9/11
Confirmed Experession in Non-Cancerous: 1/7
Cancer: 6/11
Non-Cancer: 1/7
Cancer: 6/11
Non-Cancer: 1/7
QSpec
Quasi-Poisson
C: 6/11
N: 1/7
Confirmed Expression in Cancer: 7/11
Confirmed Experession in Non-Cancerous: 5/7
Cancer: 7/11
Non-Cancer: 2/7
Confirmed Expression in Cancer: 7/11
Confirmed Experession in Non-Cancerous:2/7

Figure 8. A Venn diagram showing the number of top ranked differentially expressed HNSCC dataset proteins that were ranked as one of the top 100 proteins by each of the three statistical methods we used. In designating common proteins for all three methods we have used abbreviations of C for cancerous, and N for normal datasets.

isons for the normal and cancerous spectral counts based on the negative binomial FDR values. We can see that generally the data points with low p-values are spread out near the highest means, and that there are more high p-values found within the normal data. The trend is less pronounced than in QSpec or even quasi-Poisson, with a large number of proteins with high means given high p-values. However, proteins with greater means and greater variances are not necessarily rewarded with lower p-values. The coefficient of variance for cancer is much more spread out than in the normal, but to a lesser effect than QSpec.

4.6 Known differential expression for the HNSCC dataset

Li and colleagues used multiple reaction monitoring to validate 18 of the dysregulated proteins in the HNSCC dataset suggested by the quasi-Poisson model using spectral counts (Li et al. 2010). Of these, 11 proteins were upregulated in the cancerous dataset and 7 were upregulated in the normal dataset. To compare the three methods in our case, we generate the top hundred proteins from quasi-Poisson, negative binomial and QSpec, and compare the results. Figure 8 summarized the comparison in a Venn diagram. It shows the number of verified, differentially expressed proteins that were ranked amongst the top 100 proteins in each of the three methods as well as how many were shared between the three methods. QSpec places 12 out of the 18 confirmed proteins in its top 100 list of dysregulated proteins. The corresponding numbers for the negative binomial and quasi-Poisson models are 10 and 9, respectively. 7 of the 18 proteins (6 in cancerous and 1 in normal datasets) are among the top ranked 100 proteins by all three methods. Note that, to show the number of proteins in the intersection of all three methods, we have used abbreviations C for cancerous, N for normal datasets. The results in this case show a better performance for QSpec, though the prediction cases of the cancerous proteins was found to be the best for the negative binomial method.

5. THE STRENGTHS AND WEAKNESS OF EACH OF THE THREE STATISTICAL METHODS

The major findings of our research can be summarized in these following statements: (i) The quasi-Poisson statistical model is significantly more liberal in determining a protein as differentially expressed than the QSpec or negative binomial are. (ii) Negative binomial is not as vulnerable to the “zero variance problem” as quasi-Poisson and QSpec. (iii) The negative binomial model is sensitive to cases when there are no observations in one of the groups.

5.1 How conservative is each model relative to each other?

Out of the three methods, quasi-Poisson seems to be the most liberal with the greatest number of proteins with p-values smaller than 0.01. QSpec selects similar proteins types as quasi-Poisson does, but is more conservative. The negative binomial method tends to reward spectral count observations with greater variance with lower p-values than quasi-Poisson or QSpec do. The negative binomial method turned out to be the most conservative method of all.

5.2 How each model handles proteins that have zero variance?

The top ten proteins for quasi-Poisson all exhibit a similar nature of small means and variances. Spectral counts of c(2, 2, 2, 2) for non-cancerous and c(0, 0, 0, 0) for cancerous, and c(1, 1, 1, 1) non-cancerous and c(2, 2, 2, 2) for cancerous make up 7 of the top ten proteins for quasi-Poisson. Quasi-Poisson has a strong bias towards proteins that exhibit low variance. For example, a protein with a spectral count observation of c(2, 2, 2, 2) for non-cancerous and c(0, 0, 0, 0) for cancerous data would be treated as more significantly expressed than a protein that has c(100, 120, 80, 100) for non-cancerous and c(23, 34, 12, 16) for cancerous data. This is because Poisson based models have a great difficulty with handling low variance, and proteins whose spectral counts have zero variance will have its importance exaggerated by the quasi-Poisson model. c(2, 2, 2, 2) for non-cancerous and c(0, 0, 0, 0) for cancerous data and c(0, 0, 0, 0) for non-cancerous and c(2, 2, 2, 2) for cancerous data are given different p-values by the three methods. This is because within the HNSCC dataset, there is a bias toward normal datasets, with them accounting for a significantly greater number of spectra than the cancerous data. The models compensate for this by taking the natural log of the total number of observations. When the variance is not equal to zero, all three methods give rather high p-values to proteins which have low mean spectral count numbers for the cancerous and non-cancerous data. All three methods heavily favor proteins having large mean spectral count in the cancerous dataset and low mean spectral count in the normal dataset. The opposite is also true, but the p-values are higher due to the normalization for the bias towards normal dataset. QSpec does not have any zero variance samples within its top ten proteins, and seems to have a heterogeneous mixture of proteins with respect to variance and mean. Amongst the three methods quasi-Poisson model had the highest number of zero variance proteins in its top ten protein list. The negative binomial method is less vulnerable to the zero variance problem.

The coefficients of variance for the negative binomial distribution both increased with an increase in the p-value, but this correlation is quite weak. This was true with the other two methods as well, although QSpec demonstrated the strongest correlation for coefficient of variance (although that was still rather weak). Generally proteins with the lowest p-values for the negative binomial distribution tended to have greater variances, with few replications having identical sample sizes. This was true of QSpec as well, with very few zero values found for either QSpec or the negative binomial distribution. The opposite phenomena occurred for quasi-Poisson, with very low variance occurring between replicates for the highest ranked proteins. A great number of samples with four identical replicates occurred, but samples with slight variance in replicates were less likely to get ranked favorably. The means diverge as p-values rise in the quasi-Poisson method, the means converge as p-values rise in the QSpec, and little correlation was noticed concerning the means of the proteins ranked highest by the negative binomial method. Quasi-Poisson is inclined to treat spectral count replicates with greater variance amongst them as significantly expressed. This is similarly true for negative binomial.

5.3 The vulnerability of the negative binomial model to replicates with zero spectral counts

Label-free shotgun proteomics has been historically vulnerable to under-estimating proteins that exhibit zero spectral count in one of its two groups (Li et al. 2010). It should be noted that the negative binomial model is sensitive in these cases when there are no observations in one of the groups. For example in Table 1 the contaminant protein with the occupancy vector c(56, 52, 83, 26, 0, 0, 0, 0) produces non-significant result for the two state model. Probability for the slope from the z-test is 1.0. However, if we slightly "perturb" the occupancy vectors by introducing small, but non-zero observations values in the second state, c(56, 52, 83, 26, 0, 1, 0, 0), then the slope becomes important up to p-value of 10^{-7}! It is clear that in this case the model is not stable. In fact, addition of 1 spectrum to the second group should have made the slope term even less important if the first computation was right. Note that due to the nature of the spectral count experiments, cut-offs based on the scores, variations in tandem mass spectra, the number of spectra is never accurate up to one spectrum. Therefore, care should be practiced when analyzing data points where there is no observation in one of the states. The negative binomial model may not be accurate in this case. In general, we observed that one way to check a model would be to introduce small (up to the accuracy of the mass spectral counts, which is always more than 1) changes in the spectral count and check if the results are still consistent.

5.4 Summary and future directions

Here we have applied the QSpec and quasi-Poisson, two previously established statistical methods for label-free shotgun proteomics, to a data set known to have differential expression, as well as a control data set. We have also compared the negative binomial model to these two models to determine its effectiveness in describing spectral count data. We have seen that the quasi-Poisson statistical model is significantly more liberal in determining a protein as differentially expressed than the QSpec is, and the negative binomial is significantly more conservative. The negative binomial statistical model is not as vulnerable to the "zero variance problem" as quasi-Poisson and QSpec, but is still vulnerable when there are no observations in one of the groups.

In spite of the successes of spectral counting as a label-free method, in comparison to other methods like isotope labeling it has its limitations. The accuracy of the statistical methods relies on the role of various instrument control settings for controlling signal to noise ratio. There are limitations of the spectral count method with regards to a small number of replicates and bias toward high abundance proteins (Choi et al. 2008). To this end, future efforts should focus on improving accuracies in experimental protocols which will help to design more accurate multivariate statistical approaches that can effectively combine different abundance metrics leading to improved statistical power of detecting differential proteins.

There is great potential for future research in this field. Further testing of the negative binomial model's ability to describe label-free shotgun proteomic data is suggested, and there are other statistical models that like negative binomial have been well established in research but have never been thoroughly applied to label-free shotgun proteomics data.

ACKNOWLEDGEMENTS

This work was supported, in part, by HHSN2722008000 48C NIAID Clinical Proteomics Center (Allan R. Brasier, UTMB) and NIH-NLBIHHSN268201000037C NHLBI Proteomics Center for Airway Inflammation (Alex Kurosky, UTMB).

Received 1 April 2011

REFERENCES

Al-Shahrour, F., az-Uriarte, R. and Dopazo, J. (2004). FatiGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20** 578–580.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* **57** 289300. MR1325392

Bondarenko, P. V., Chelius, D. and Shaler, T. A. (2002). Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal. Chem.* **74** 4741–4749.

Cai, L., Huang, H., Blackshaw, S., Liu, J. S., Cepko, C. and Wong, W. H. (2004). Clustering analysis of SAGE data using a Poisson approach. *Genome Biol.* **5** R51.

Chelius, D. and Bondarenko, P. V. (2002). Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J. Proteome. Res.* **1** 317–323.

Choi, H., Fermin, D. and Nesvizhskii, A. I. (2008). Significance analysis of spectral count data in label-free shotgun proteomics. *Mol. Cell Proteomics* **7** 2373–2385.

Griffin, N. M., Yu, J., Long, F., Oh, P., Shore, S., Li, Y., Koziol, J. A. and Schnitzer, J. E. (2010). Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat. Biotechnol.* **28** 83–89.

Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H. and Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17** 994–999.

Hochberg, Y. and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Stat. Med.* **9** 811–818.

Huang, H., Cai, L. and Wong, W. H. (2008). Clustering analysis of SAGE transcription profiles using a Poisson approach. *Methods Mol. Biol.* **387** 185–198.

Li, M., Gray, W., Zhang, H., Chung, C. H., Billheimer, D., Yarbrough, W. G., Liebler, D. C., Shyr, Y. and Slebos, R. J. (2010). Comparative shotgun proteomics using spectral count data and quasi-likelihood modeling. *J. Proteome Res.* **9** 4295–4305.

Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M. and Yates, J. R., III (1999). Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17** 676–682.

Liu, H., Sadygov, R. G. and Yates, J. R., III (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76** 4193–4201.

Neal, D. J. and Simons, J. S. (2007). Inference in regression models of heavily skewed alcohol use data: A comparison of ordinary least squares, generalized linear models, and bootstrap resampling. *Psychol. Addict. Behav.* **21** 441–452.

Old, W. M., Meyer-Arendt, K., veline-Wolf, L., Pierce, K. G., Mendoza, A., Sevinsky, J. R., Resing, K. A. and Ahn, N. G. (2005). Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell Proteomics.* **4** 1487–1502.

Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A. and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell Proteomics.* **1** 376–386.

Paoletti, A. C., Parmely, T. J., Tomomori-Sato, C., Sato, S., Zhu, D., Conaway, R. C., Conaway, J. W., Florens, L. and Washburn, M. P. (2006). Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. *Proc. Natl. Acad. Sci. U.S.A* **103** 18928–18933.

Pavelka, N., Fournier, M. L., Swanson, S. K., Pelizzola, M., Ricciardi-Castagnoli, P., Florens, L. and Washburn, M. P. (2008). Statistical similarities between transcriptomics and quantitative shotgun proteomics data. *Mol. Cell Proteomics* **7** 631–644.

Pavelka, N., Pelizzola, M., Vizzardelli, C., Capozzoli, M., Splendiani, A., Granucci, F. and Ricciardi-Castagnoli, P. (2004). A power law global error model for the identification of differentially expressed genes in microarray data. *BMC. Bioinformatics* **5** 203.

Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A. and Pappin, D. J. (2004). Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell Proteomics.* **3** 1154–1169.

Schulz-Trieglaff, O., Machtejevas, E., Reinert, K., Schluter, H., Thiemann, J. and Unger, K. (2009). Statistical quality assessment and outlier detection for liquid chromatography-mass spectrometry experiments. *BioData. Min.* **2** 4.

Shah, A. R., Davidson, J., Monroe, M. E., Mayampurath, A. M., Danielson, W. F., Shi, Y., Robinson, A. C., Clowers, B. H., Belov, M. E., Anderson, G. A. and Smith, R. D. (2010). An efficient data format for mass spectrometry-based proteomics. *J. Am. Soc. Mass Spectrom.*

Ver Hoef, J. M. and Boveng, P. L. (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology* **88** 2766–2772.

Vogel, C. and Marcotte, E. M. (2008). Calculating absolute and relative protein abundance from mass spectrometry-based protein expression data. *Nat. Protoc.* **3** 1444–1451.

Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K., Florens, L. and Washburn, M. P. (2006). Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J. Proteome Res.* **5** 2339–2347.

Zybailov, B. L., Florens, L. and Washburn, M. P. (2011). Quantitative shotgun proteomics using a protease with broad specificity and normalized spectral abundance factors. *Mol. Biosyst.*

Matthew C. Leitch
Department of Biochemistry and Molecular Biology
Sealy Center for Molecular Medicine
The University of Texas Medical Branch
Galveston, TX 77555
USA

Indranil Mitra
Department of Biochemistry and Molecular Biology
Sealy Center for Molecular Medicine
The University of Texas Medical Branch
Galveston, TX 77555
USA

Rovshan G. Sadygov
Department of Biochemistry and Molecular Biology
Sealy Center for Molecular Medicine
The University of Texas Medical Branch
Galveston, TX 77555
USA
E-mail address: rovshan.sadygov@utmb.edu