

# Quasi-likelihood method for shotgun proteomic data

Ming Li<sup>1,2</sup>, Robbert Slebos<sup>3,5</sup>, Will Gray<sup>1</sup>, Daniel C. Liebler<sup>4,5</sup>, and Yu Shyr<sup>1,2\*</sup>

<sup>1</sup>Division of Cancer Biostatistics, <sup>2</sup>Department of Biostatistics, <sup>3</sup>Department of Cancer Biology, <sup>4</sup>Department of Biochemistry <sup>5</sup>Jim Ayers, Institute for Precancer Detection and Diagnosis, Vanderbilt University, Nashville, Tennessee 37232, USA.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXXX

## ABSTRACT

**Summary:** Recent advances in proteomic technologies have brought investigators a richer understanding of and capacity for detection of protein patterns and biomarkers. However, the complexity and high-dimensionality of LS-MS/MS (shotgun) data make quantitative analysis quite challenging. The statistical tools used to analyze such data are still immature. In this paper, we introduce a quasi-likelihood Poisson regression method to compare peptide/protein identifications between groups, with the unit of measurement defined as the frequency of a specific peptide/protein observation in a single LC-MS/MS analysis. To evaluate the performance of the proposed method, we applied it to a data set with known spiked-in proteins, allowing us to test the methodology on real data where the 'truth' is known by experimental design. Controlled by the FDR procedure, we compared the true detection and false positive rates of the proposed quasi-likelihood method with two recently developed methodologies, the Poisson likelihood model and the beta-binomial test. We conclude that the quasi-likelihood Poisson regression method retains the strength of both alternative methods and performs better on this spiked data set based on true detection and false positive rates.

**Availability:** The package is available upon request.

**Contact:** yu.shyr@vanderbilt.edu

## 1 INTRODUCTION

Shotgun proteome analysis platforms based on multidimensional liquid chromatography-tandem mass spectrometry (LC-MS/MS) provide a powerful tool to discover biomarker candidates in tissue specimens [Slebos *et al.*, 2008]. Given the fact that the spectral counts from label-free shotgun proteomics can be used for the protein abundance measurements [Old *et al.*, 2005; Liu *et al.*, 2004; Zybailov *et al.*, 2005], a protein/peptide frequency-based analysis approach has been adopted widely to compare spectral identifications between treatment groups, such as cancer vs. normal [Hanks *et al.*, 2006]. One major goal of statistical analysis on spectral count data is to identify certain differential proteins as 'winners,' potential biomarkers worthy of further investigation. Researchers have compared the ability of statistical tests such as the Fisher's exact test and the G-test to identify the 'winner' proteins in spectral count data [Zhang *et al.*, 2006]. Those simple tests seem

handy but cannot model the distribution of count data effectively, not to mention the fact that they lack the ability to account for the covariates associated with data. Most recently, two new methods have been developed that better suit the characteristics of the data and experiments: (1) a Poisson likelihood model in a Bayesian framework (Qspec) [Choi *et al.*, 2008] and (2) a beta-binomial test [Pham *et al.*, 2009]. One advantage of the Poisson likelihood method is that it can model complicated shotgun data structure from complex experimental designs; an obvious disadvantage, however, is the potential violation of the assumption of the equal mean-variance of Poisson distribution, which limits its applicability to the empirical fitting of shotgun data. The beta-binomial test considers the within and between-sample variations but is not flexible enough to take into account other settings from the experimental design of the data. In an attempt to combine the strengths of the two methods above while overcoming their weaknesses, we propose a quasi-likelihood Poisson model [McCullagh, 1983; Breslow, 1990] that will enable us to put our analysis in a regression framework but with relaxed restrictions on distribution assumptions. This approach is appropriate for modeling count data with overdispersion and/or underdispersion [Faddy *et al.*, 2001], which is typical of shotgun data.

This paper is organized as follows. Section 2 provides statistical details about quasi-likelihood method. Section 3 evaluates the proposed quasi-likelihood method by applying it to a yeast data set with known spike-in human proteins, subsequently comparing its performance to two recently developed methods, the Poisson likelihood model and the beta-binomial test. Section 4 discusses the analysis strategies and related issues and draws conclusions.

## 2 QUASI-POISSON MODEL

To compare spectral counts for different clinical groups (such as normal vs. tumor), we can model the data in a regression framework. Let  $Y$  denote spectral counts and  $x$  stand for group. Since  $Y$  represents spectral count, it is not appropriate to assume a Gaussian distribution; instead, generalized linear models (GLM) will be applied to handle such non-normal responses [Nelder *et al.*, 1972]. Specifically, Poisson distribution, a distribution from the exponential family of distributions, is usually assumed for count data. This model can be expressed as:

\*to whom correspondence should be addressed

$$\log(Y) = \beta_0 + \beta_1 X_1 + \epsilon \quad (1)$$

Eq (1) is fitted by maximizing the Poisson likelihood function, and the group effect can be accessed by testing the significance of  $\beta_1$ . However, for specifying a Poisson distribution, we put an equal mean-variance assumption on the data, which is usually not held in the empirical fitting. To alleviate this assumption, we propose to use quasi-Poisson maximum likelihood. The key idea is that instead of claiming that  $Y$  is from a known distribution, we assume only knowledge of the first and second moments. In other words, for fitting such 'count' data, we are able to specify the link and variance function of the model, but we do not have a strong idea on an appropriate distribution form for the response. The important part of the model specification is the link and the variance, and the outcome is less sensitive to the distribution of the response, given a reasonable sample size [Faraway, 2006]. For quasi-Poisson method, the regression model can be specified as in Eq (1). The fitting procedure is an analogy of fitting the model using Poisson likelihood.

Let's derive the quasi-likelihood Poisson model as following: for the  $i^{th}$  response  $Y_i$ , we have  $E(Y_i) = u_i$  and  $Var(Y_i) = \varphi V(u_i)$ . Now define a score,  $U_i$  as

$$U_i = \frac{Y_i - u_i}{\varphi V(u_i)}$$

$$-E\left(\frac{\partial U_i}{\partial u_i}\right) = -E\left[\frac{-\varphi V(u_i) - (Y_i - u_i)\varphi V'(u_i)}{[\varphi V(u_i)]^2}\right]$$

It is easy to show that

$$E(U_i) = 0 \quad (2)$$

$$-E\left(\frac{\partial U_i}{\partial u_i}\right) = V(U_i) \quad (3)$$

We notice that the properties for score  $U_i$  shown in Eq (2) and (3) are shared by the derivatives of the log-likelihood, which suggests that the integration  $U_i$  would serve as a good surrogate for likelihood. It is natural for us to define a log quasi-likelihood for  $Y_i$  as:

$$Q_i = \int_{y_i}^{u_i} \frac{y_i - t}{\varphi V(t)} dt(t)$$

Then the log quasi-likelihood for all  $n$  observations will be

$$Q = \sum_{i=1}^n Q_i$$

It is not hard to verify that  $Q$  behaves just like log-likelihood and the estimation of  $\beta$  is obtained by maximizing  $Q$ . We summarize some features of the quasi-Poisson model as:

1. The usually asymptotic properties expected from maximum likelihood estimators also hold for quasi-likelihood based estimators [McCullagh, 1983]. Theoretically, these properties are assuring and desirable.
2. The quasi-likelihood Poisson method allows for the dispersion  $\varphi$  to be a free parameter. This parameter is useful in modeling overdispersion, which is typical of shotgun data [Pham *et al.*, 2009]. The quasi-likelihood method can model the variation of such data with increased accuracy.
3. The quasi-likelihood premise broadens our modeling possibilities for more real world data types: When we do not have a clear idea about the distribution of the data, we still can model such data with only knowledge of link and variance [Faraway, 2006].
4. In addition, quasi-likelihood allows us to model the data in a regression framework that is easily extended to model more complicated data from complicated experiments, such as repeated measurements, longitudinal data, etc. The quasi-likelihood method provides generally consistent estimates of regression coefficients even if the variance function is misspecified [Moore *et al.*, 1991].

### 3 METHOD EVALUATION

Researchers have concluded that the beta-binomial test performs the best among all existing methods [Pham *et al.*, 2009], and that the Poisson likelihood method is preferred in a regression setting. The key idea by Qspec [Choi *et al.*, 2008]) provides a framework that is flexible and easy to extend. Therefore, rather than a comprehensive comparison to all existing methods, we will focus on comparing the proposed quasi-likelihood method with these two advanced methods: beta-binomial and Poisson likelihood model. For method evaluation criteria, we will use power of detection (true positive rate, sensitivity) and type I error (false positive rate) as applied in [Pham *et al.*, 2009]. We also apply the FDR controlling procedure for handling multiple comparisons during simultaneous testing of thousands of proteins [Benjamin *et al.*, 1995]. We provide the power and false positive control rates for these methods, corresponding to certain FDR values.

#### 3.1 The Data

It is crucial to select an appropriate data set for method evaluation since the evaluation on either power or false positive rate depends on knowledge of 'ground truth.' Simulation is a common choice, but it can miss the true characteristics of the data in real world. Using a list provided by experts who know the ground truth may be an alternative, but it is still merely an approximation [Pham *et al.*, 2009]. The data we choose is specifically designed for method evaluation: A mixture of 48 human proteins was spiked into a known yeast reference proteome at 5 different concentration levels (A: 0.24, B: 0.67, C: 2.54, D: 6.7 and E: 20 fmol/ug yeast protein. More details of the data is described in [Paulovich *et al.*, 2009]). For such spiked data, we know for certain the ground truth about the existing difference; thus the comparisons based on such a data set will be fair to any of these tests.

### 3.2 True Detection Rate

Figures 1, 2, and 3 show the power at different FDR values of quasi-likelihood, Poisson likelihood, and beta-binomial methods respectively. First of all, as expected, they share the same patterns: The power increases as the spiked human protein intensity goes higher. Secondly, by a head-to-head comparison from these 3 plots, we see quasi-likelihood method performs better at all 5 concentration levels, especially when the concentration is low. At levels C, D, and E, the quasi-likelihood method can detect most of the human proteins with a reasonably small FDR value. When the in-spiked level is low, the quasi-likelihood model yields the best results, but all three methods have low detection rates.

Table 1 summarizes the powers for three methods at a small fixed FDR value 0.05. Again, we see that at all levels, the quasi-likelihood method performs best. Especially, when the concentration is low (Yeast vs. A; Yeast vs. B), quasi-likelihood method has the power to detect such a small difference. When the difference is high, all three methods have similar performance. In supplemental figures, we provide results on powers among the different levels of comparison at different FDR for the three methods. All give evidence that the quasi-likelihood method is an improvement over existing methods.

### 3.3 False Positive Rate

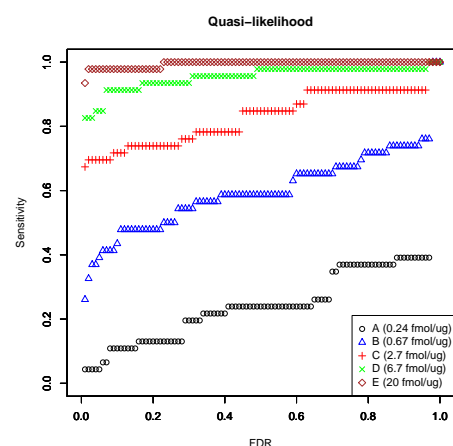
Figures 4, 5, and 6 show the false positive rate (FPR) at different FDR values of the quasi-likelihood, Poisson likelihood, and beta-binomial methods respectively. They all perform well, while quasi-likelihood has a slightly higher but still reasonable FPR at one concentration level. Table 2 summarizes the FPR for three methods at a small fixed FDR value 0.05. The values in the table show good performance for all three methods.

**Table 1.** Powers (%) of Three Methods at FDR 0.05

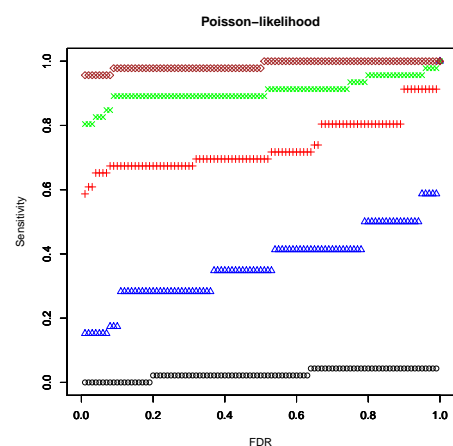
	Quasi	Poisson	Beta-Binomial
Yeast vs. Yeast+Spike			
A	4	0	4
B	39	15	20
C	70	65	65
D	85	83	80
E	98	96	93
3-fold difference			
D vs. E	67	57	54
C vs. D	35	17	28
9-fold difference			
C vs. E	93	93	89
B vs. D	72	70	70
27-fold difference			
B vs. E	96	96	93
A vs. D	83	80	80

## 4 DISCUSSION AND CONCLUSION

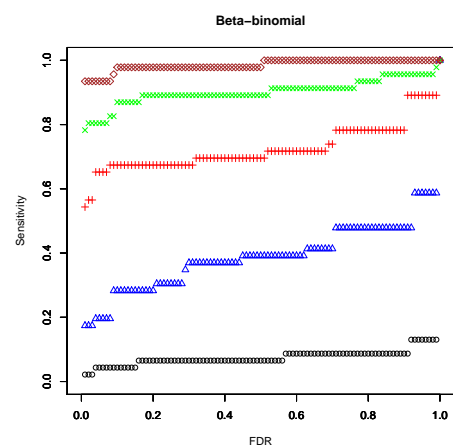
The analysis goal for shotgun data cannot be achieved completely by providing a method or test that is only suitable to the data type itself. Indeed, as with all types of high-throughput data, we face



**Fig. 1.** Quasi Powers at 5 Different Spiked Concentration



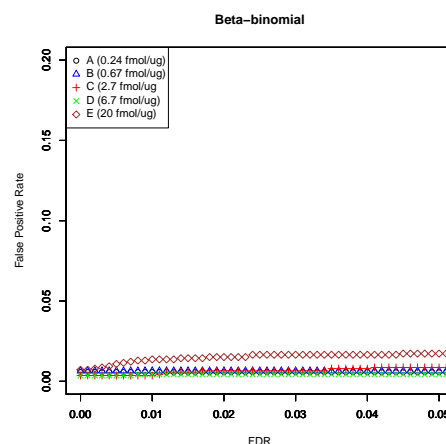
**Fig. 2.** Poisson Powers at 5 Different Spiked Concentration



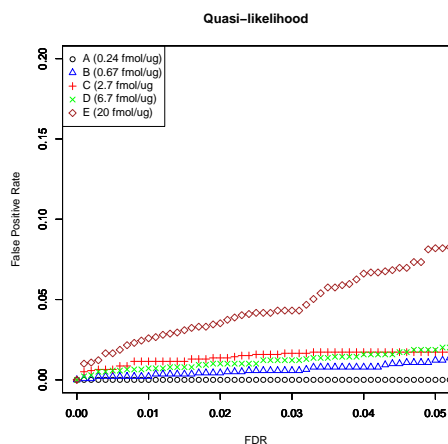
**Fig. 3.** Beta-binomial Powers at 5 Different Spiked Concentration

**Table 2.** False Positive Rate (%) at FDR 0.05

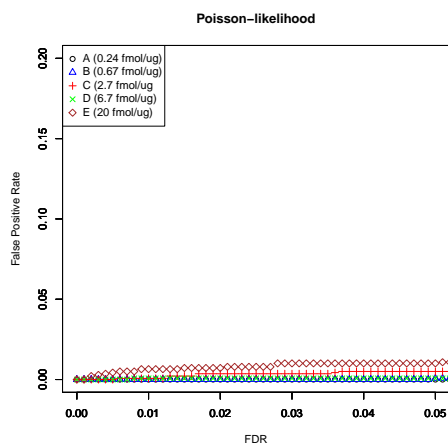
	Quasi	Poisson	Beta-Binomial
<b>Yeast vs. Yeast+Spike</b>			
A	0	0	1
B	1	0	1
C	2	1	1
D	2	0	0
E	8	1	0
<b>3-fold difference</b>			
D vs. E	2	0	0
C vs. D	2	0	1
<b>9-fold difference</b>			
C vs. E	9	2	2
B vs. D	2	0	0
<b>27-fold difference</b>			
B vs. E	9	2	2
A vs. D	2	0	0



**Fig. 6.** Beta-binomial FPR at 5 Different Spiked Concentration



**Fig. 4.** Quasi FPR at 5 Different Spiked Concentration



**Fig. 5.** Poisson FPR at 5 Different Spiked Concentration

additional challenges in the real world application, such as multiple comparisons, large number of zero observations, small sample size, and data normalization. To complete our work, presented below are the strategies we employed to address such issues when analyzing shotgun data in real world applications.

1. Shotgun proteomic data typically contains thousands of proteins. When we test thousands of hypotheses simultaneously, we need to be careful of a potential selection bias. Unadjusted P-values taken from single-inference procedures result in an increased rate of false positives. Family-wise error rate methods are too conservative and have less application value for discovery purposes because they are not suitable for the large-scale simultaneous hypothesis testing problems that arise from high-throughput technologies. To handle the complications presented by simultaneously testing thousands of proteins, we apply the False Discovery Rate (FDR) [Benjamin *et al.*, 1995] controlling procedure.
2. If we have strong biological knowledge to conclude some spectra are not believable prior to any statistical treatment, we should omit them from analysis. When we apply the FDR controlling procedure to adjust for multiple comparison issues later on, we will be less "punished" due to fewer attempts in searching out the significance. Appropriate thresholds for the first step data screening might be necessary.
3. Another complicating feature of shotgun proteomic data sets is the presence of large numbers of zero-values, e.g., no spectra observed for a given protein. When comparing two test groups, we can encounter data with all zeros in one group, which causes the estimated standard error to be enormous. In such a case, the corresponding Wald test statistics, based on estimated coefficients with standard errors, are not reliable. A conservative way to avoid this problem is to add one spectrum count to the group comprised only of zeros. Another way is to compare using nested models. When comparing models, the P-value of an F-test rather than a  $\chi^2$  test is calculated following the quasi-likelihood method [Faraway, 2006].

4. One limitation of LC-MS/MS based shotgun proteomics data sets is that due to both financial and time constraints, the number of analytic runs is usually small. Such small sample sizes complicate statistical analyses of the data, and an asymptotic distribution for the test statistics might not hold when the sample size is small. For example, tests based on the  $\chi^2$  distribution are only an approximation that becomes more accurate as the sample size increases. One possible solution is to provide an appropriate test statistic. An appropriate test statistic may involve a difference of deviance (performed on two nested models) that is generally more accurate than a goodness-of-fit test involving a single deviance.
5. To make sure the data are comparable, we need to 'normalize' it prior to comparison. The 'confident ids' from each run [Ma *et al.*, 2009] is added into the quasi-likelihood regression model at the offset. The offset serves as the 'size' variable, which determines the number of opportunities for proteins to occur, and by modeling such an offset, we normalize the shotgun data and make it comparable for each group.

Different methods have been developed for shotgun data, but until now there has been no 'gold standard' statistical method for analyzing such data. In addition to the methods either mentioned before or proposed in this paper, negative binomial and zero-inflated Poisson models are often viewed as solutions for count data with a large number of zeros. We did not choose them because for smaller sample sizes, such as the one presented in the current study, the assumption on mixture distributions will further complicate the estimation process; that is, we would be under-powered to estimate the mixtures. The analysis method and the set of strategies presented in this paper can be used as a general framework for this type of data, which is flexible for extension to deal with data coming from various shotgun proteomics experiment designs. A possible extension of the quasi-likelihood approach would be an adaptation for repeated measurements and longitudinal shotgun study designs through generalized estimating equations (GEE) [Liang *et al.*, 1986; Zeger *et al.*, 1986]. Our model is in a regression setting that can be easily extended to incorporate other covariates when needed.

## ACKNOWLEDGEMENT

This research was supported in part by Jim Ayers Institute for Precancer Detection and Diagnosis (need code here), the Lung Cancer Special Program of Research Excellence (SPORE) (2P50 CA090949-06A1), GI SPORE (2P50 CA095103-06), and Cancer Center Support Grant (CCSG) (5P30 CA068485-12). The authors are grateful to collaborators for shotgun related projects that motivated this research. In addition, we wish to thank Lynne Berry and Yvonne Poindexter for their editorial work on this manuscript.

## REFERENCES

- Luo, W., Hill, S., Slebos, R., Li, M., Brbek, J., Amanchy, R., Pandey, A., Ham, A., and Hanks, S. (2008). The global impact of oncogenic Src on a phosphotyrosine proteome. *Journal of Proteome Research*, **XX**, XXX-XXXX.
- Old, W.M., Meyer-Arendt, K., Aveline-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. *Molecular & Cellular Proteomics*, **4**, 1487-1502.
- Zhang, B., VerBerkmoes, N., Langston, M.A., Uberbacher, E., Hettich, R.L., and Samatova, N.F. (2006). Detecting differential and correlated protein expression in label-free shotgun proteomics. *Journal of Proteome Research*, **5**, 2909-2918.
- Pavelka, N., Fournier, M.L., Swanson, S.K., Pelizzola, M., Ricciardi-Castagnoli, P., Florens, L., and Washburn, M.P. (2007). Statistical similarities between transcriptomics and quantitative shotgun proteomics data. *Molecular & Cellular Proteomics*, **x,xx**, xxx-xxxx.
- Choi, H., Fermin, D. and Nesvizhskii, A. (2008). Significance analysis of spectral count data in label-free shotgun proteomics. *Molecular & Cellular Proteomics*, **7**, 12, 2373-2385.
- Pham, T.V., Piersma, S.R., Warmoes, M., and Jimenez, C.R. (2009). On the beta binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. *Bioinformatics*, **xx**, 363-9.
- Slebos, R., Brock, J., Winters, N., Stuart, S., Martinez, M., Li, M., Chambers, M., Zimmerman, L., Ham, A., Tabb, D., and Liebler, D. (2008). Evaluation of strong cation exchange and isoelectric focusing of peptides for multidimensional liquid chromatography-tandem mass spectrometry. *Journal of Proteome Research* **xx**, 5286-94.
- McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics*, **11** 59-67.
- Faraway, J. (2006). Extending the linear model with R. *Chapman & Hall/CRC*.
- Benjamin, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, No. 1 289-300.
- Nelder, J. and Wedderburn, R. (1972). Generalized linear model. *Journal of the Royal Statistical Society, A* **132** 370-384.
- Breslow, N. (1990). Test of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *Quarterly of applied mathematics*, **85**, Providence: Brown University.
- Faddy, M.J. and Bosch, R.J. (2001). Likelihood-based modeling and analysis of data underdispersed relative to the Poisson distribution. *Biometrics*, **57**(2), 620-624.
- Moore, D.F. and Tsiatis, A. (1991). Robust estimation of the variance in moment methods for extra-binomial and extra-Poisson variation. *Biometrics*, **47**(2), 383-401.
- Liu, H., Sadygov, R.G. and Yates, J.R. 3rd (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem*, **76**(14), 4193-201.
- Zybailov, B. et al. (2005). Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Anal Chem*, **77**(19) 6218-24.
- Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73** 13-22.
- Zeger, S.L. and Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**(1) 121-30.
- Paulovich, A. G., Billheimer, D., Ham, A. J., Vega-Montoto, L. J., 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., Fenyó, D., Carr, S. A., Fisher, S. J., Gibson, B. W., Mesri, M., Neubert, T. A., Reginier, F. E., Rodriguez, H., Spiegelman, C., Stein, S. E., Tempst, P., and Liebler, D. C. (2009). A CPTAC inter-laboratory study characterizing a yeast performance standard for benchmarking LC-MS Platform performance. *Mol Cell Proteomics*.
- 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., and Tabb, D. L. (2009). IDPicker 2.0: Improved Protein Assembly with High Discrimination Peptide Identification Filtering. *Journal of proteome research*, **8** 3872-3881.