Statistics and Its Interface Volume 13 (2020) 465–474

Meta-analysis of peptides to detect protein significance

Yuping Zhang*, Zhengqing Ouyang, Wei-Jun Qian,
Richard D. Smith, Wing Hung Wong, and Ronald W. Davis

Shotgun assays are widely used in biotechnologies to characterize large molecules, which are hard to be measured as a whole directly. For instance, in Liquid Chromatography – Mass Spectrometry (LC-MS) shotgun experiments, proteins in biological samples are digested into peptides, and then peptides are separated and measured. However, in proteomics study, investigators are usually interested in the performance of the whole proteins instead of those peptide fragments. In light of meta-analysis, we propose an adaptive thresholding method to select informative peptides, and combine peptide-level models to protein-level analysis. The meta-analysis procedure and modeling rationale can be adapted to data analysis of other types of shotgun assays.

Keywords and phrases: Meta-analysis, Adaptive thresholding, Shotgun technology.

1. INTRODUCTION

Classical meta-analysis refers to integrating multiple analysis results from individual studies to see if the overall effect is significant [9]. Meta-analysis plays an increasingly popular role in modern genomic research, such as combining multiple transcriptomic studies to identify differentially expressed genes [31], integrating multiple genomic studies for pathway enrichment analysis [30], and among others. To perform meta-analysis, it is crucial to appropriately collect a reasonable set of studies, and extract useful information from individual studies.

In this paper, we instead of performing meta-analysis in a classic application scenario, but adapt and extend the rationale of meta-analysis to model proteomic data from high-throughput shotgun assays. Shotgun proteomics has been used for identifying proteins in biological samples using a combination of high performance Liquid Chromatography (LS) combined with Mass Spectrometry (MS). It is named by analogy with the rapidly-expanding, quasi-random firing pattern of a shotgun. LC-MS has become one of main technologies for the emerging field of proteomics with applications in discovering novel disease-specific protein biomarkers, gaining better understanding of disease processes, and monitoring therapeutic responses [2, 1, 6, 7, 8, 19, 20, 26, 27, 28]. Typically, in LC-MS, protein samples are first digested into peptides by sequence-specific proteases such as trypsin. The resulting peptides are then separated by capillary LC and analyzed by tandem MS via an electrospray ionization interface. The detected LC-MS features contain the information on the mass, LC elution time, and intensity indicative of abundance for individual peptides. Many thousands of peptides can be identified in a single LC-MS or LC-MS with additional Mass spectrometry (LC-MS/MS) analysis using mass and time tag strategies [39] or bioinformatics approaches [3, 11, 12, 14, 16]. Peptide abundances are obtained based on either peak heights or peak areas of the detected LC-MS features [28]. A challenging aspect of the analysis is that measurement in peptide abundances can be affected not only by actual biological changes, but also by bias and noise. LC-MS reproducibility and quantification is affected by sample processing variations and LC-MS platform variations [33]. Moreover, different peptides derived from a given protein can have different responses and variations due to the differences in digestion and ionization efficiencies as well as protein modifications. The mapping between peptides and proteins is performed by searching existing protein sequence database. Mapping error is common in the mapping process.

Due to limited dynamic range of LC-MS detection and variation in platform sensitivity, low-abundance peptides may be detected in some samples but not in others even if they have the same concentrations within these samples. This leads to another significant challenge for LC-MS data modeling, namely, missing data. The degrees of missing data can be affected by protein abundances (as shown in Figure 1), which should be treated as non-random missing. The lower the abundance of a protein, the higher the missing rate of the peptides. Because of that, existing methods for handling randomly missing data such as k-nearest neighbors (KNN) [35], SVD based imputation method [25] or excluding the missing values directly, may lead to erroneous results [22]. A further challenge of proteomics is the variability of peptides for the same protein. Existing methods for protein level abundance estimation such as DAnTE [25] are based on averaging the intensities of all the peptides from a protein after some kinds of transformation. The most frequently observed peptide is often chosen

*Corresponding author. ORCID: 0000-0001-8986-0354.

as the reference peptide. Then, peptides originating from the same protein are scaled on the basis of the pre-chosen reference peptide (RRollup method) or with a modified z-sore approach (ZRollup method). After scaling, peptide intensities are averaged to obtain the relative protein abundance. These existing methods do not explicitly account for the issue of variability and missing data problems discussed above. In this paper, we present an additive mixed model to address the multiple sources of variance, and handle the heterogeneity of peptides using peptide-specific models. We begin with an additive model to obtain peptide-level significance and then adaptively select peptides to make protein-level inference through meta-analysis. We call our method PEAT – Protein Expression through Adaptive Thresholding. The software website is https://sites.google.com/site/statyuping/software/peat.

2. METHOD

To illustrate our modeling ideas, we plot peptides originated from one of spiking-in proteins in Figure 1 from a real dataset [21]. In the data, a dilution mixture of the tryptic digests of six nonhuman purified proteins was spiked into a complex sample background of human peptides isolated by solid-phase N-glycopeptide captured from serum. Figure 1 shows the intensities of those peptides from protein Adolase A. There are six levels of protein abundances injected per sample, which are 25, 50, 100, 200, 400, and 800 (fmol) (from left to right in Figure 1). For each protein concentration level, the data contains three samples, which are binned within the corresponding condition in Figure 1. As showed in Figure 1, different peptides from the same protein can provide vastly different signals. Peptides from different runs may have different missing rates and intensities, even when they belong to the same biological condition. Thus, a peptide-specific model is needed to address this heterogeneity. We consider two types of signals that a peptide may carry in the differential analysis, which include peptide intensities and observation rates. Consequently, we build two types of models, one is the intensity model, and the other is the observation-rate model.

Explicitly, we first check whether the peptide was observed in every condition. If each condition has at least one observation, we check whether there are missing data, and if so, we impute the missing data. In this case, we use the intensities of peptides to test whether they are differentially expressed across biological conditions (intensity model). If there is one condition without observations, we will use the observation rate of peptides to test whether they are differentially expressed across biological conditions (observation-rate model). The reason we consider both intensity model and observation-rate model is that peptides can be either absent in a sample or present at levels below the detection limit of the MS instrument. Finally, in order to obtain the protein-level statistics, we propose an adaptive thresholding statistic and use a permutation test to select appropriate thresholds. Below, we explain the details of each step.

2.1 Peptide-specific models

2.1.1 Missing data handling

Let y_{gi} be the peptide intensity for peptide g ($g \in \{1, \cdots, G\}$) of sample i ($i \in \{1, \cdots, I\}$) which is nested in group k ($k \in \{1, \cdots, K\}$). We assume intensities of peptide g from biological group k follow the normal distribution $N(u_{gk}, \sigma_g^2)$. If one peptide has observations in every biological condition and has missing data as well, we can impute the missing data. As shown in Figure 1, while peptides from low-abundance proteins are more likely to be missing; high-abundance protein can also have missing peptides. We model missing data from low abundance proteins as non-random missing. Statistically, the signal peaks from each peptides are censored at the left at a threshold dependent on detective sensitivity. The probability that censoring occurs is modeled as the left-hand tail probability of the $N(u_{gk}, \sigma_g^2)$ distribution, evaluated at the censoring threshold c, denoted by $\phi((c-u_{gk})/\sigma_g)$, where c is the unknown detection threshold for a missing peptide in a LC-MS experiment. We use one-way ANOVA (cell means model):

$$\mathbf{y}_g = \mathbf{X}\mathbf{u}_g + \mathbf{e}_g \tag{1}$$

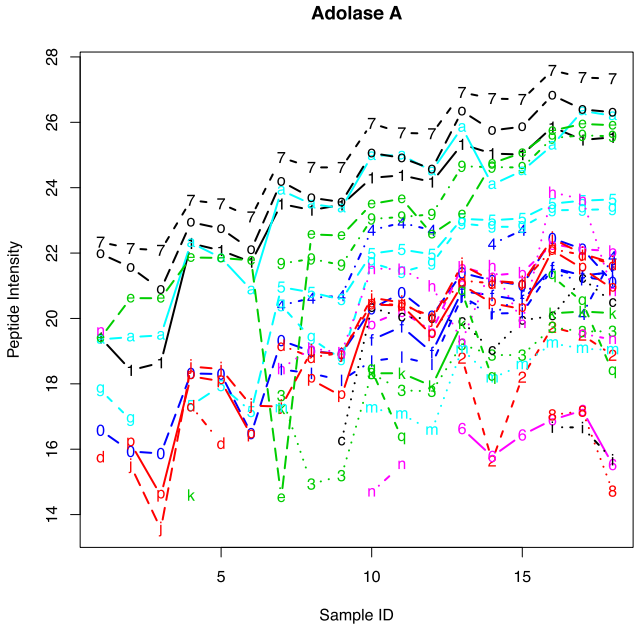

Figure 1. Intensities of peptides from Adolase A in spike-in data. The x-axis indicates samples. There are 18 samples from 6 conditions in total. Each condition has three samples. The samples are ordered by their conditions. Different condition has different spike-in protein concentrations, which are 25, 50, 100, 200, 400, and 800 (fmol), ordered from left to right in the figure. The y-axis indicates the log2 scaled peptide intensities. Different lines with different colors and types indicate different peptides.

to estimate $\mathbf{u_g}$, the vector consisting of u_{gk}, where $k \in \{1, \cdots, K\}$, K is the number of biological conditions, $\mathbf{y}_g$ is the vector consisting of the intensities of peptide g, and $\mathbf{X}$ is the design matrix for K groups.

Besides the intensity-dependent missingness, some peptides from high abundance proteins are missing completely at random due to technical factors such as ion-suppression effects [34], where some particular peptides dominate the LC-MS experiments and suppress the detection of other peptides. Incorrectly treating randomly missing peptides as intensity-dependent missing peptides or vice versa will result in biased estimates. Thus, we want to estimate the probabilities of "missing completely at random" and "missing not at random" from the entire collection of data.

We assume the probability of any peptide being randomly missed is π. Denote the intensity of peptide g from sample i by y_{gi}. Let W_{gi} be an indicator of whether y_{gi} is unobserved (0 if not missed, 1 if missed), which follows the Bernoulli distribution. Considering the two mechanisms of missing, the probability of a peptide is unobserved can be calculated as follows:

$$(2)\quad P(W_{gi}=1|u_{gk(i)}) = \pi+(1-\pi)\phi((c-u_{gk(i)})/\sigma_g) = q_{gi},$$

where $k(i)$ is the group index of sample i belongs to. Let θ denote the vector of unknown parameters, which consists of π, c and σ_g. The log-likelihood function for the above Bernoulli distribution is of the form: $l(\theta) = \sum_{i=1}^{I}\sum_{g=1}^{G}[(1-W_{gi})\log(1-q_{gi}) + W_{gi}\log q_{gi}]$.

For c, we use the minimum observed intensity of the whole dataset as its estimate $\hat{c}$. For π, we first fit a nonlinear regression model with the form of $m_g = f(\bar{y}_g)+\epsilon_g$, which m_g is the missing rate for peptide g, $\bar{y}_g$ is the average of all observed intensities of peptide g. In practice, we fit the nonlinear regression model using cubic splines. Then, we estimate the random missing probability π as $\hat{\pi} = f(\max_g \bar{y}_g)$, as illustrated in Figure 2. We then employ an iterative procedure to estimate the rest parameters and perform imputations. Specifically, we first assign initial values to the parameters. Let $\mathbf{y}_{gO}$ indicate the vector of observed intensities of peptide g. Let $\mathbf{X}_O$ be the design matrix corresponding to $\mathbf{y}_{gO}$. First, we obtain $\hat{\mathbf{u}}_{gk}^{(0)}$ by solving the regression model $\mathbf{y}_{gO} = \mathbf{X}_O\mathbf{u}_g + \mathbf{e}_g$. Then, $\hat{\mathbf{y}}_g^{(0)} \leftarrow \mathbf{X}\hat{\mathbf{u}}_g^{(0)}$. We then estimate the parameter σ_g as $\hat{\sigma}_g^{(0)} \leftarrow \sqrt{Var(\mathbf{y}_g - \hat{\mathbf{y}}_g^{(0)})(I-1)/(I-K)}$, where I is the number of samples, and K is the number of biological conditions. With the initial values, we then iterate the following steps for $l = 1, 2, \ldots$, until convergence.

1. For missing values, imputations are carried out by generating values at random by the following procedure. Suppose intensity y_{gi} from sample i and peptide g is missing. The probability of treating this missing value to be censored is as below:

$$(3)\quad P(y_{gi}^{(l-1)} < \hat{c}|W_{gi}=1) = \Phi(\zeta_{gi}^{(l-1)})/[\hat{\pi}+(1-\hat{\pi})\Phi(\zeta_{gi}^{(l-1)})],$$

which $\zeta_{gi}^{(l-1)} = (\hat{c} - \mathbf{x}_i\hat{\mathbf{u}}_g^{(l-1)})/\hat{\sigma}_g^{(l-1)}$, $\mathbf{x}_i$ is the i-th row vector of matrix $\mathbf{X}$ corresponding to sample i. We draw a random variable based on the Bernoulli distribution $B(P(y_{gi}^{(l-1)} < \hat{c}|W_{gi}=1))$. If the random sample is 1, we treat the missing value as censored missing. Then, the missing value $y_{gi}^{(l)}$ is imputed with a random draw from the normal distribution $N(\mathbf{x}_i\hat{\mathbf{u}}_g^{(l-1)}, \hat{\sigma}_g^{(l-1)})$ right-truncated at $\hat{c}$. Otherwise, we treat the missing value as completely random missing. Then, the missing value $y_{gi}^{(l)}$ is imputed with a random draw from the same Normal distribution, but without truncation at the estimated censoring point.

2. Obtain $\hat{\mathbf{u}}_g^{(l)}$ by solving the regression model $\mathbf{y}_g^{(l)} = \mathbf{X}\mathbf{u}_g + \mathbf{e}_g$, $\hat{\mathbf{y}}_g^{(l)} \leftarrow \mathbf{X}\hat{\mathbf{u}}_g^{(l)}$, $\hat{\sigma}_g^{(l)} \leftarrow \sqrt{Var(\mathbf{y}_g^{(l)} - \hat{\mathbf{y}}_g^{(l)})(I-1)/(I-K)}$.

2.1.2 Mixed regression model using peptide intensities

The variance of peptide intensities are affected by several factors including peptide intrinsic characteristics, experimental technique properties, and biological conditions. We propose a peptide-specific additive mixed model for the LC-MS data. Let μ denote the overall mean for all peptides under all conditions, v_i denote the main effect of each experimental run. Let α_g indicate the overall average effect for peptide g, and $t_{gk(i)}$ is the effect of biological conditions, such as disease groups, which is the effect we are mostly interested in. $k(i)$ is the group index, of which sample i belongs to. The additive mixed effect model is of the form

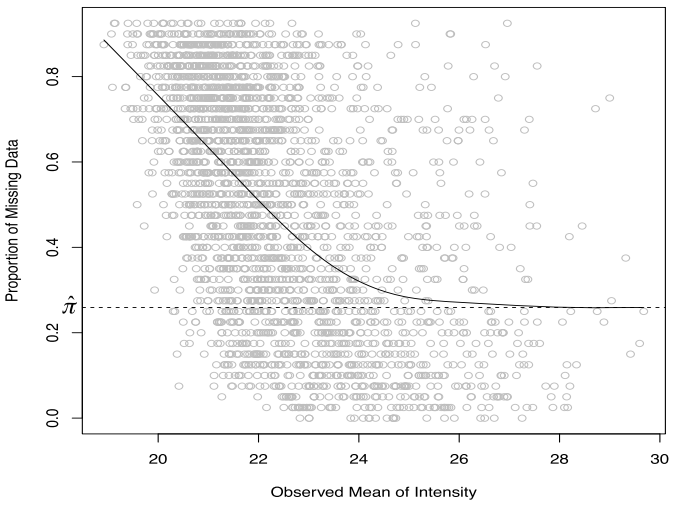

Figure 2. Random missing probability estimation. The x-axis indicates the average intensity of each peptide. The y-axis indicates the missing rate of each peptide. The solid curve is fitted by the cubic spline regression. The value on the y-axis for the dotted line indicates the estimated completely random missing probability $\hat{\pi}$.

$y_{gi} = \mu + v_i + \alpha_g + t_{gk(i)} + \varepsilon_{gi}$. In particular, μ, α_g and $t_{gk(i)}$ are fixed effects, while v_i is a normally distributed random effect with zero mean, and ε_{gi} is the error term assumed having a normal distribution with zero mean. The models have the following constraints: $\sum_{g=1}^{G} \alpha_g = \sum_{k=1}^{K} t_{gk(i)} = 0$. The null hypothesis is that the peptide is not differentially expressed, i.e., $t_{g1} = \cdots = t_{gK} = 0$.

2.1.3 Logistic regression model using peptide observation probabilities

For peptides completely missing in one biological condition, i.e. data is not observed for all of the subjects that belong to some biological condition(s), the above model does not apply. This is because that the regression coefficient cannot be obtained without observations for one biological condition. We will not throw this subset of data away, because some protein within this subset could be biologically differentially expressed, e.g. a protein expressed in one condition but not expressed in other condition(s). Thus, we include this type of data in our analysis and propose a logistic regression model to test the significance of differential expression. In the logistic regression model, the binary outcome variable (denoted as o_{gi}) indicates the observation status of peptide g in sample i (1 observed, 0 unobserved). Let p_{gi} be the probability of peptide g observed in sample i. The logistic regression model is of the form: $\text{logit}(E[o_{gi}|x_{g1}, \cdots, x_{gK}]) = \text{logit}(p_{gi}) = \ln[p_{gi}/(1-p_{gi})] = \beta_{g0} + \sum_{k=1}^{K} \beta_{gk} x_{ik}$, where β_{gk} reflects the biological condition effect in group k, x_{ik} is an indicator of whether sample i belongs to group k.

2.1.4 Peptide-level significance analysis

We define the null hypothesis (H_0: the peptide of interest is not differentially expressed) in the sense that there is no difference in intensity (H_{01}) and no difference in observation rate (H_{02}). The corresponding alternative hypothesis H_A is defined as the peptide is differentially expressed, i.e. there is difference in intensity (H_{A1}) or there is difference in observation rate (H_{A2}). Let l_{01} and l_{02} denote the log-likelihood function of the null model under H_{01} and H_{02}, respectively. Let l_{A1} and l_{A2} denote the log-likelihood function of the unconstrained model under H_{A1} and H_{A2}, respectively. We use the negative log-likelihood ratio test statistic $-2(l_{01} - l_{A1})$ and $-2(l_{02} - l_{A2})$ to detect the differentially expressed peptides, which both asymptotically follow the $\chi^2_{(K-1)}$ distribution under H_{01} and H_{02}, respectively.

2.2 Meta-analysis of peptide-level models to obtain protein-level significance

Our goal is to detect differentially expressed proteins across multiple biological conditions. The number of peptides mapped to one protein can range from several to several hundreds. In the situation of multiple peptides per protein, a sophisticated model is needed. Given p-values of peptides mapped to one protein, we want to obtain the protein-level p-value. Moreover, not every observed peptide mapped to one protein represents the true signal of the protein equally, due to the complexity of proteolytic processing and post-translational modifications as well as potential mapping errors. We thus want to select good peptides that are informative.

Considering protein j, we assume there are m_j peptides mapped to this protein. Suppose peptide g is mapped to protein j. Let p_g denote the p-value of peptide g differentially expressed across different biological conditions, which is obtained from the above peptide-specific models. Let H_0^g denote that peptide g is not differentially expressed across different biological conditions. H_0^g is true either because the protein is not differentially expressed across different biological conditions, or because peptide g is not informative on protein level due to technical factors or mapping errors. Let H_A^r denote the hypothesis that there are exactly r peptides mapped to a protein carrying the true signal. The alternative hypothesis is written as $H_A = H_A^1 \cup \cdots \cup H_A^{m_j}$. We rank the peptides according to their p-values in an increasing order. Intuitively, if the true signal lies in H_A^r, we can improve the power by only including peptides with the top r smallest p-values in peptide-to-protein summarization. However, for one protein, we do not know in advance the number of peptides with the true signals. Moreover, different proteins may have different number of informative peptides. Because of these difficulties, we propose the following adaptive thresholds with the aim to improve the power of the testing. Let $p_{(1)}, \cdots, p_{(m_j)}$ denote the ordered p-values. We define a combined statistic as $C_{j(r)} = -\sum_{g=1}^{m_j} \log(p_{(g)} I(g \le r))$. r is chosen to minimize $p_r = P(C_r)$, which is the p-value of the observed C statistic. The adaptive thresholding statistic V is defined as the minimal p-value among p_r, $r \in \{1, \cdots, m_j\}$. $V = \min_{r \in \{1,\cdots,m_j\}} P(C_r)$. Finally, the significance of the observed value of V is obtained by permutation analysis.

Below we illustrate the detailed procedure for the adaptive thresholding statistic when applied to the detection of differentially expressed proteins.

1. Peptide-specific p-value calculation
   (a) If the missing values for the peptide of interest is imputable, we impute the missing values and calculate the p-value using the likelihood ratio test based on standard regression models with peptide intensities as outcomes.
   (b) If the missing values for the peptide of interest is not imputable, we calculate the p-value using likelihood ratio test based on the logistic regression models.
2. Calculate the adaptive-thresholding statistic V:
   (a) Given r, the observed combined statistic C for protein j is $C_{j(r)} = -\sum_{g=1}^{m_j} \log(p_{(g)}) I(g \le r)$. Define the permuted combined statistic $C_{j(r)}^{(b)} = -\sum_{g=1}^{m_j} \log(p_{(g)}^{(b)}) I(g \le r^{(b)})$ from permutation b with group indices permuted.

(b) Estimate the p-value of the observed C_j as

$$P(C_{j(r)}) = \frac{\sum_{b=1}^{B}\sum_{j'=1}^{J} I\{C_{j'(r)}^{(b)} \geq C_{j(r)}\}}{B \cdot J},$$

where J is the number of proteins, B is the number of permutations. Similarly, for the permutation b, we have

$$P(C_{j(r)}^{(b)}) = \frac{\sum_{b'=1}^{B}\sum_{j'=1}^{J} I\{C_{j'(r)}^{(b')} \geq C_{j}^{(b)}\}}{B \cdot J}.$$

(c) Calculate the optimal r for protein j as

$$r^* = \arg\min_{r\in\{1,\ldots,m_j\}} P(C_{j(r)}).$$

To find the optimal r^*, the computational complexity is $O(m_j)$. The computational complexity is lower than the adaptive weight statistic proposed in existing literature [18], which is $O(2^{m_j})$. Similarly,

$$r^{(b)*} = \arg\min_{r\in\{1,\ldots,m_j\}} P(C_{j(r)}^{(b)}).$$

Define the adaptive thresholding statistic V_j as $V_j = P(C_{j(r)^*})$. Similarly, $V_j^{(b)} = P(C_{j(r)^{(b)*}}^{(b)})$.

3. Assess the p-value and q-value of the adaptive-thresholding statistic V

(a) The p-value of V_j is calculated as

$$P_V(V_j) = \frac{\sum_{b=1}^{B}\sum_{j'=1}^{J} I\{V_{j'}^{(b)} \leq V_j\}}{B \cdot J}.$$

(b) Estimate π_0, the proportion of not differentially expressed proteins, as

$$\hat{\pi}_0 = \frac{\sum_{j=1}^{J} I\{P_V(V_j) \in A\}}{J \cdot l(A)}.$$

We choose $A = [0.5, 1]$ and $l(A) = 0.5$ as suggested by the literature [32].

(c) Estimate the q-value for each protein as

$$q(V_j) = \frac{\hat{\pi}_0 \sum_{b=1}^{B}\sum_{j'=1}^{J} I\{V_{j'}^{(b)} \leq V_j\}}{B\sum_{j'=1}^{J} I\{V_{j'} \leq V_j\}}.$$

2.3 Estimation of protein-level expression

Protein-level expression is summarized from the selected peptide intensities. For protein j with m_j mapped peptides, we calculate its expression for sample i as

$$E_{ji} = \frac{1}{\sum_{g=1}^{m_j} I(g \leq r^*)} \sum_{g=1}^{m_j} \delta_{gi} y_{gi} I(g \leq r^*),$$

where δ_{gi} is the scaling factor. The selected peptides mapped to the same protein are scaled on the basis of the reference peptide to bring all peptide profiles across biological conditions to the same level. To remove outlying values, a Grubb's outlier test is performed. The Grubb's test is used to detect if the sample dataset contains one outlier, statistically different than the other values [10]. The test is based on calculating a score (the difference between outlier and the mean divided by standard deviation) of this outlier and comparing it to an appropriate critical value. The critical value for this test is calculated according to the approximation given by Pearson and Sekar [24]. Let $\tau_{gi} = (y_{gi} - \bar{y}_g)/s$, where $\bar{y}_g = \sum_{i=1}^{I} y_{gi}/I$, $s = \sqrt{\sum_{i=1}^{I}(y_{gi} - \bar{y}_g)^2/I}$. If y_{gi} is an observation arbitrarily selected from a random sample of I drawn from an infinite normal population, then the elementary probability distribution of τ is

$$p(\tau_g) = \frac{\Gamma(\frac{I-1}{2})}{\sqrt{(I-1)\pi}\Gamma(\frac{I-2}{2})}(1 - \frac{\tau_g^2}{I-1})^{\frac{I-4}{2}}.$$

The probability that the absolute value of τ_i is greater than a specified value τ_0 is

$$P\{|\tau_{gi}| > \tau_{g0}\} = 2\int_{\tau_{g0}}^{\sqrt{I-1}} p(\tau_g)d\tau_g.$$

The critical value τ_{g0} is calculated by reversing the above formula with a specified p-value cutoff (we use 0.05 as the p-value cutoff).

3. POWER AND ADMISSIBILITY

In this section, we study the power and admissibility of the proposed adaptive thresholding statistic under some assumptions. We assume independence among peptides mapped to one protein of interest. For simplicity, we consider two-sample test of means of two Gaussian distributions with known variance and without missing data.

Let

$$Z_g = \frac{\bar{X}_{2g} - \bar{X}_{1g}}{\sigma_g\sqrt{1/n_1 + 1/n_2}}, \tag{4}$$

$g = 1, \cdots, m_j$ be the statistic for peptide g in protein j, where $\bar{X}_{1g} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$, $\bar{X}_{2g} = \frac{1}{n_2}\sum_{i=1}^{n_2} X_i$, $X_{gi} \sim N(0, \sigma_g^2)$ when $1 \leq i \leq n_1$, and $X_{gi} \sim N(\theta_g, \sigma_g^2)$ when $n_1+1 \leq i \leq n_2$. The p-value for peptide g is $P_g = Pr(|Z_g| \geq |z_g||\theta_g = 0)$, where Z is the standard normal distribution. Denote the null hypothesis by $H_0 = \{\theta_1 = \cdots = \theta_{m_j} = 0\}$ and alternative hypothesis by $H_A = \{\text{at least one } \theta_g \neq 0\}$. Let $\beta^{AT}(\theta;\alpha)$ be the power of a test controlled at level α for the adaptive thresholding statistic given $\theta \in H_A$, we have

$$\begin{aligned}\beta^{AT}(\theta;\alpha) &= Pr(V \leq V_\alpha|\theta)\\ &= 1 - \int_{\Omega^{AT}} \Pi_{g=1}^{m_j} p(P_g|\theta)dP_1\cdots dP_{m_j},\end{aligned}$$

where V_α is the solution of v to the equation $P(V \le v|H_0) = \alpha$, $\Omega^{AT} = \{P(C_{j(r^*)} > V_\alpha)\} = \cap_{r=1}^{m_j}\{P(C_{j(r)}) > V_\alpha\} = \cap_{r=1}^{m_j}\{C_{j(r)} < \chi_{2r}^{-2}(1-V_\alpha)\}$, and χ_{2r}^{-2} is the inverse CDF of a χ_{2r}^2 with the degrees of freedom $2r$.

When H_0 is true, the individual P_g is uniformly distributed on $[0,1]$. The density of the p-value under H_A is as below

$$p(P|\theta) = \frac{p(x|\theta)}{p(x|0)} \quad (0 \le P \le 1),$$

where $x = g(P)$ indicates the solution of $P = \int_x^1 f(x|0)dx$ [23]. In above simplified setting, the density of P_g is

$$p(P_g|\theta_g) = \frac{1}{2}\exp\left\{\frac{c_g}{2}[2\Phi^{-1}(1-P_g/2) - c_g]\right\} + \frac{1}{2}\exp\left\{\frac{c_g}{2}[-2\Phi^{-1}(1-P_g/2) + c_g]\right\}, \tag{5}$$

where $c_g = \frac{\theta_g}{\sigma_g\sqrt{1/n_1+1/n_2}}$, $g = 1, \cdots, m_j$.

Without peptide-selection, the power of Fisher's combined probability test is

$$\begin{aligned}\beta^{Fisher}(\theta;\alpha) &= Pr(-\sum_{g=1}^{m_j} P_g \le \chi_{2m_j}^2(1-\alpha)|\theta \in H_A)\\ &= 1 - \int_{\Omega^{Fisher}} \Pi_{g=1}^{m_j} p(P_g|\theta)dp_1\cdots dp_{m_j},\end{aligned}$$

where $\Omega^{Fisher} = \{C_{j(m_j)} \le \chi_{2m_j}^2(1-\alpha)\}$, and $p(P_g|\theta_g)$ is determined by Equation (5).

Obviously $\Omega^{AT} \le \Omega^{Fisher}$, thus $\beta^{AT} \ge \beta^{Fisher}$. This means, peptide selection can improve the power with the existence of uninformative peptides mapped to the protein of interest, due to technical factors or potential mapping errors.

Theorem 3.1 ([4]). *Under H_A and the test statistic is in the exponential family, the necessary and sufficient condition for a combined test procedure to be admissible is that the corresponding acceptance region is convex.*

Corollary 1. *The acceptance region of adaptive thresholding statistic (AT) is convex and, thus, AT is admissible under H_A and assumption (4).*

Proof. Denote the two-sided p-value by $p_g = 2(1-\Phi(|z_g|))$, where $\Phi(x) = \int_{-\infty}^x \phi(x)$, and $\phi(x)$ is the density of the standard normal distribution. Below, we prove that $f(z_g) = -\log(p_g) = -\log(1-\Phi(|z_g|)) - \log 2$ is convex.

With simple calculation, we have $f''(z) = \frac{\phi(z)}{[1-\Phi(|z|)]^2}\{\phi(z) - |z|[1-\Phi(|z|)]\}$ when $z \ne 0$. Because $1-\Phi(z) \le \phi(z)/z$, for $z > 0$, thus, $f''(z) > 0$, when $z \ne 0$. In addition, $f(z)$ is continuous at $z = 0$, so $f(z)$ is convex in z. Because the sum of convex functions is convex, we can further obtain $f(z_1, \cdots, z_r) = -\sum_{g=1}^r \log(p_g)$ for $\forall g \ge 1$ is convex.

For the adaptive thresholding statistic (AT), the acceptance region is $\{z_1, \cdots, z_{m_j} : \min_{1\le r\le m_j} p(c_j(r)) > c\}$, where $p(c_{j(r)})$ is the right-sided p-value of $C_{j(r)}$.

$$\begin{aligned}&\{z_1, \cdots, z_{m_j} : \min_{1\le r\le m_j} p(c_j(r)) > c\}\\ &= \bigcap_{r=1}^{m_j}\{z_1, \cdots, z_r : p(-\sum_{g=1}^{r}\log(p_{(g)})) > c\}\\ &= \bigcap_{r=1}^{m_j}\{z_1, \cdots, z_r : p(\max_{g\in G_r}(-\sum_{g}\log(p_g))) > c\}\\ &= \bigcap_{r=1}^{m_j}\{z_1, \cdots, z_r : \min_{g\in G_r} p(-\sum_{g}\log(p_g)) > c\}\\ &= \bigcap_{r=1}^{m_j}\bigcap_{g\in G_r}\{z_1, \cdots, z_r : p(-\sum_{g}\log(p_g)) > c\}\\ &= \bigcap_{r=1}^{m_j}\bigcap_{g\in G_r}\{z_1, \cdots, z_r : -\sum_{g\in G_r}\log(p_g) < \chi_{2r}^{-2}(1-c)\},\end{aligned}$$

where G_r is the set including any r peptides. Thus, the acceptance region of adaptive thresholding statistic is convex, since the intersection of convex sets is convex. □

4. SIMULATION STUDIES

To study the specificity and sensitivity of our approach, we performed the following six simulation experiments. We mimicked the spike-in data to generate the peptides and proteins in our simulations. We generated 94 proteins, and each corresponding simulated protein had the same number of peptides as in the spike-in data [21]. The first 20 proteins were simulated to be differentially expressed. The rest proteins were simulated to be not differentially expressed. The random missingness π parameter was set to be 0.1 for simulations 1, 2, 5 and 6; but 0.2 for simulations 3 and 4. The censoring threshold was selected such that a total of 20% all measurements were missing for simulations 1, 2, 3, 4 and 6; but 30% all measurements were missing for simulation 5. For each protein, we randomly indicate no more than 40% of its mapped peptides are uninformative for simulations 1, 2, 3, 4 and 5, but no more than 30% of its mapped peptides are uninformative for simulation 6. For simulations 1, 3, 5 and 6, we simulated 100 samples. The first 50 samples were from group 1; the second 50 samples were from group 2. For simulations 2 and 4, we simulate 20 samples and two biological conditions. Each biological condition contains 10 samples. The expression levels of proteins were generated via the following procedure.

Let G_0 indicate the set of peptides that are differentially expressed. Let v_i indicate the effect of LC-MS experiment for sample i, α_g indicate the effect of peptide g, $t_{gk(i)}$ indicate the group effect of peptide g and group $k(i)$, and ε_{gi} indicate the error effect. We generated the data according to the distributions as below.

$$y_{gi} = 15 + v_i + \alpha_g + \alpha_g * t_{gk(i)} + \varepsilon_{gi},$$

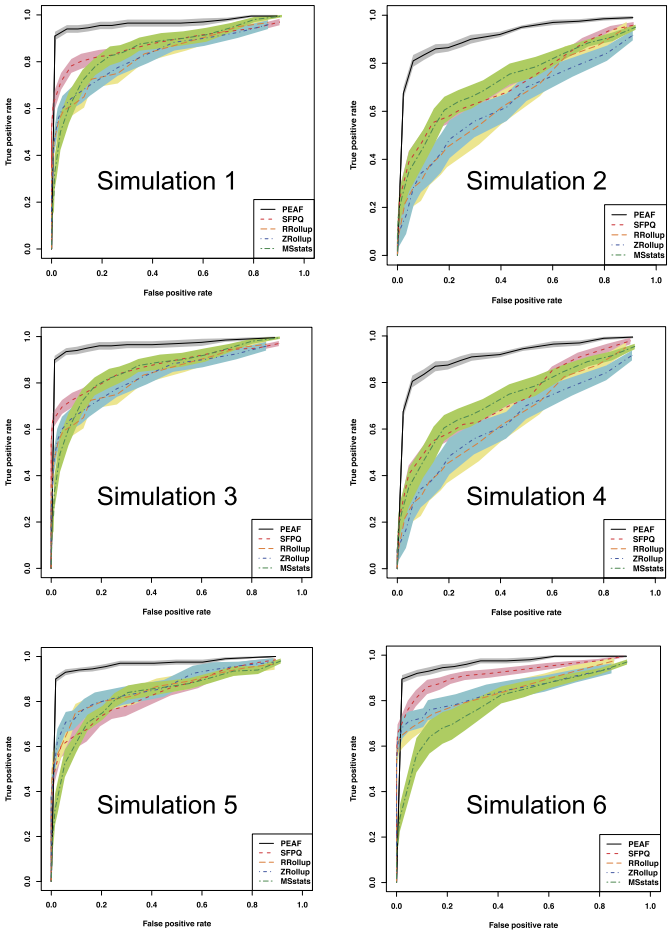

Figure 3. The Receiver Operating Characteristic (ROC) plots for six simulation studies. The x-axes indicate false positive rate, and the y-axes indicate true positive rate. The curves with different colors and line types show the average true positive rates across 10 times replicates for each method respectively. The corresponding shadows show the standard errors. Black solid lines indicate PEAT results; Red dashed lines indicate SFPQ results; Orange longdash lines indicate RRollup results; Blue dotdash lines indicate ZRollup results; Green twodash lines indicate MSstats results.

where $v_i \sim N(0,1)$, $\alpha_g \sim N(0,1)$, $\varepsilon_{gi} \sim N(0,1)$ and $t_{gk(i)}$ is generated based on the following procedure:

$$t_{gk(i)} = \begin{cases} N(1, 0.1) & \text{if } g \in G_0 \text{ and } k(i) = 1 \\ N(-1, 0.1) & \text{if } g \in G_0 \text{ and } k(i) = 2 \\ N(0, 0.1) & \text{else.} \end{cases}$$

We run 10 times for each simulation. The performance for these six simulations is illustrated in Figure 3. For comparisons, we also applied the RRollup and ZRollup methods presented in the DAnTE software [25], the MSstats [5] and the method denoted by SFPQ [15]. One can see that for each simulation, our method has the best performance. For smaller sample size, which is common in real application situation, our method has larger improvement comparing to existing methods.

Table 1. Dilution outline of the six purified proteins in the spike-in data set. Myoglobin: sp|P68082|MYG_HORSE; Carbonic anhydrase: sp|P00921|CAH2_BOVIN; Cytochrome c: sp|P00004|CYC_HORSE; Lysozyme: sp|P00698|LYSC_CHICK; Alcohol dehydrogenase: sp|P00330|ADH1_YEAST; Adolase A: sp|P00883|ALDOA_RABIT

Protein name	Protein injected (fmol) per sample					
Myoglobin	800	25	50	100	200	400
Carbonic anhydrase	400	800	25	50	100	200
Cytochrome c	200	400	800	25	50	100
Lysozyme	100	200	400	800	25	50
Alcohol dehydrogenase	50	100	200	400	800	25
Adolase A	25	50	100	200	400	800

5. APPLICATIONS TO REAL DATA

5.1 Application to spike-in data

We used a spike-in dataset [21] to illustrate the real application and compare PEAT with other methods. In this spike-in dataset, a dilution mixture of the tryptic digests of six nonhuman purified proteins was spiked into a complex sample background of human peptides isolated by solid-phase N-glycopeptide captured from serum. The dilution were designed and performed according to statistical principles spanning a dynamic range of two orders of magnitude from 25 to 800 fmol injected (as shown in Table 1). The concentration combinations of six spike-in nonhuman proteins lead to six biological conditions. We applied PEAT to this dataset, and estimated the protein-level abundances based on the selected informative peptides. For comparison, we also applied SFPQ, RRollup, ZRollup and MSstats to this spike-in dataset. We found that all the methods can detect the six non-human proteins are significantly different among the six conditions. To assess the performance of protein abundance estimation, we compared the estimated protein abundances with the real spike-in concentrations of the nonhuman proteins. We used linear regression models to calibrate the estimated protein abundances through each method with the true concentrations of proteins. The R^2 values of the regressions were used to characterize how good the fits were. Table 2 shows the R^2 values for the regressions between log 2-transformed concentrations and the log 2-transformed estimated abundances of proteins for all the methods. Overall, PEAT outperforms other methods.

5.2 Application to burn data

To demonstrate its application in clinical research, we applied PEAT to a human plasma proteome study following severe burn injury. Blood plasma samples from 10 healthy control subjects and 16 burn patients were used [29]. Samples

Table 2. Method comparisons using spike-in data. Proteins 1, 2, 3, 4, 5 and 6 are sp|P68082|MYG_HORSE, sp|P00921|CAH2_BOVIN, sp|P00004|CYC_HORSE, sp|P00698|LYSC_CHICK, sp|P00330|ADH1_YEAST and sp|P00883|ALDOA_RABIT, respectively. The table shows the R^2 values for the regressions between $\log 2$-transformed concentrations and the $\log 2$-transformed estimated abundances of proteins for each method

Proteins	1	2	3	4	5	6
PEAT	0.994	0.992	0.987	0.992	0.991	0.984
RRollup	0.982	0.988	0.975	0.951	0.977	0.943
ZRollup	0.930	0.936	0.967	0.985	0.987	0.979
MSstats	0.970	0.966	0.912	0.973	0.983	0.981

from burn patients were collected at two time points. Thus, the study contains 3 biological conditions – control, burn early time point and burn later time point. Peptide samples from individual healthy subjects or burn patients were analyzed using LC-MS. LC-MS features were identified by the AMT tag strategy and the details of data analysis were previously described [29]. We used the label-free MS intensities for each patient sample in our study without considering the 18O-labeled reference spiked into each sample. We preselected proteins with two or more unique peptides. In total, 316 proteins with 3282 peptides were studied. We applied PEAT to detect differentially expressed proteins across the three biological conditions. With the q-value < 0.1 and p-value < 0.031 criterion, 42 significant proteins were identified by PEAT. We studied the functions of these significant proteins. They are most related to the following functions: acute phase response signaling, LXR/RXR activation, complement system, coagulation system, intrinsic prothrombin activation pathway, atherosclerosis signaling, and clathrin-mediated endocytosis signaling. These findings are in good agreement with previous studies [13, 29, 36, 37, 38]. Among these proteins, FLT4 had been verified as a drug target for inflammation [17]. Figure 4 shows the heatmap of these significant proteins detected by PEAT. According to the trend of protein abundance changes between burn patients and healthy subjects, these proteins are divided into two groups – early responding proteins and late responding proteins. Early responding proteins have larger perturbations at the first time point than the second time point. And vice versa for late responding proteins.

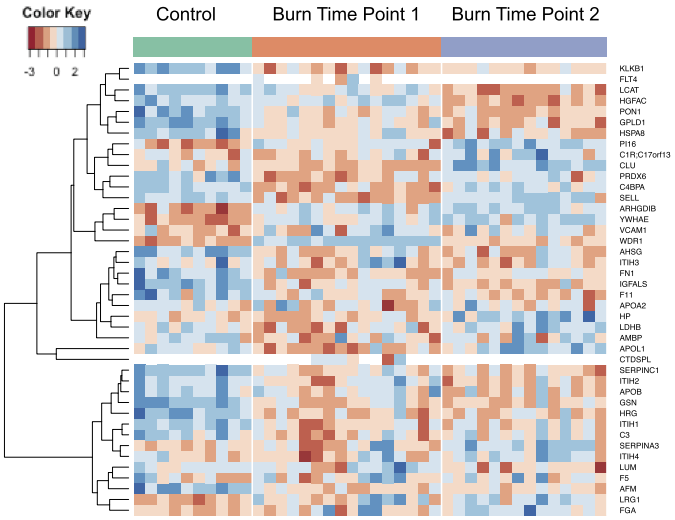

Figure 4. Heatmap of the estimated protein expression for the burn injury study. Each row indicates one protein, each column indicates one sample. The samples are from three categories – healthy controls, burn patients from the first time point, burn patients from the second time point. We add a white line among different biological conditions. The white cells in the heatmap indicate that no protein expression values were estimated due to peptide observations were missing for the entire group.

6. DISCUSSION

In light of meta-analysis, we developed a new method through an adaptive thresholding statistic, PEAT, for data analysis arising from shotgun assays. We illustrated it in proteomics studies and demonstrated the utility for LC-MS data analysis. We considered the mechanisms of different types of missing data and the variations associated with LC-MS experiments at the peptide level and their effects on the protein-level variations. PEAT was designed to combine peptide-level models, select informative peptides, and then perform protein-level analysis. PEAT can be used in label-free MS data analysis, and also serves a good complementary analysis tool for labeled MS experiments. The proposed meta-analysis procedure can be adapted to data from other shotgun technologies, where the large molecules of interest are divided into small components so that they can be measured.

ACKNOWLEDGMENT

This work was supported by National Institutes of Health (NIH) Grant HG 000250 (to R.W.D) and NIH grant P41GM103493 (to R.D.S).

Received 28 April 2019

REFERENCES

[1] Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature* **422** 198–207.

[2] Belczacka, I., Latosinska, A., Metzger, J., Marx, D., Vlahou, A., Mischak, H. and Frantzi, M. (2019). Proteomics biomarkers for solid tumors: Current status and future prospects. *Mass spectrometry reviews* **38** 49–78.

[3] Bellew, M., Coram, M., Fitzgibbon, M., Igra, M., Randolph, T., Wang, P., May, D., Eng, J., Fang, R., Lin, C., Chen, J., Goodlett, D., Whiteaker, J., Paulovich, A. and

McIntosh, M. (2006). A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics* **22** 1902–9.

[4] Birnbaum, A. (1954). Combining independent tests of significance. *Journal of the American Statistical Association* 559–74. MR0065101

[5] Clough, T., Key, M., Ott, I., Ragg, S., Schadow, G. and Vitek, O. (2009). Protein quantification in label-free LC-MS experiments. *Journal of proteome research* **8** 5275–84.

[6] Diamandis, E. P. (2004). Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Molecular & cellular proteomics: MCP* **3** 367–78.

[7] Engwegen, J. Y., Gast, M. C., Schellens, J. H. and Beijnen, J. H. (2006). Clinical proteomics: searching for better tumour markers with SELDI-TOF mass spectrometry. *Trends in pharmacological sciences* **27** 251–9.

[8] Fortier, M. H., Bonneil, E., Goodley, P. and Thibault, P. (2005). Integrated microfluidic device for mass spectrometry-based proteomics and its application to biomarker discovery programs. *Analytical chemistry* **77** 1631–40.

[9] Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational researcher* **5** 3–8.

[10] Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics* **11** 1–21.

[11] Hsieh, E. J., Hoopmann, M. R., MacLean, B. and MacCoss, M. J. (2010). Comparison of database search strategies for high precursor mass accuracy MS/MS data. *Journal of proteome research* **9** 1138–43.

[12] Jaffe, J. D., Mani, D. R., Leptos, K. C., Church, G. M., Gillette, M. A. and Carr, S. A. (2006). PEPPeR, a platform for experimental proteomic pattern recognition. *Molecular & cellular proteomics: MCP* **5** 1927–41.

[13] Jeschke, M. G., Chinkes, D. L., Finnerty, C. C., Kulp, G., Suman, O. E., Norbury, W. B., Branski, L. K., Gauglitz, G. G., Mlcak, R. P. and Herndon, D. N. (2008). Pathophysiologic response to severe burn injury. *Annals of surgery* **248** 387–401.

[14] Kall, L., Canterbury, J. D., Weston, J., Noble, W. S. and MacCoss, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature methods* **4** 923–5.

[15] Karpievitch, Y., Stanley, J., Taverner, T., Huang, J., Adkins, J. N., Ansong, C., Heffron, F., Metz, T. O., Qian, W.-J., Yoon, H. and Smith, R. D. (2009). A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics* **25** 2028–34.

[16] Klammer, A. A., Yi, X., MacCoss, M. J. and Noble, W. S. (2007). Improving tandem mass spectrum identification using peptide retention time prediction across diverse chromatography conditions. *Analytical chemistry* **79** 6111–8.

[17] Leedom, A. J., Sullivan, A. B., Dong, B., Lau, D. and Gronert, K. (2010). Endogenous LXA4 circuits are determinants of pathological angiogenesis in response to chronic injury. *The American journal of pathology* **176** 74–84.

[18] Li, J. and Tseng, G. C. (2011). An Adaptively Weighted Statistic for Detecting Differential Gene Expression When Combining Multiple Transcriptomic Studies. *Annals of Applied Statistics* **5** 994–1019. MR2840184

[19] Major, M. B., Camp, N. D., Berndt, J. D., Yi, X., Goldenberg, S. J., Hubbert, C., Biechele, T. L., Gingras, A. C., Zheng, N., Maccoss, M. J., Angers, S. and Moon, R. T. (2007). Wilms tumor suppressor WTX negatively regulates WNT/beta-catenin signaling. *Science* **316** 1043–6.

[20] Mayor, T., Graumann, J., Bryan, J., MacCoss, M. J. and Deshaies, R. J. (2007). Quantitative profiling of ubiquitylated proteins reveals proteasome substrates and the substrate repertoire influenced by the Rpn10 receptor pathway. *Molecular & cellular proteomics: MCP* **6** 1885–95.

[21] Mueller, L. N., Rinner, O., Schmidt, A., Letarte, S., Bodenmiller, B., Brusniak, M. Y., Vitek, O., Aebersold, R. and Müller, M. (2007). SuperHirn – a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **7** 3470–80.

[22] Oberg, A. L., Mahoney, D. W., Eckel-Passow, J. E., Malone, C. J., Wolfinger, R. D., Hill, E. G., Cooper, L. T., Onuma, O. K., Spiro, C., Therneau, T. M. and Bergen, r. H. R. (2008). Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. *Journal of proteome research* **7** 225–33.

[23] Pearson, E. (1938). The probability integral transformation for testing goodness of fit and combining independent tests of significance. *Biometrika* **30** 134–48.

[24] Pearson, E. S. and Sekar, C. C. (1936). The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika* **28** 308–20.

[25] Polpitiya, A. D., Qian, W. J., Jaitly, N., Petyuk, V. A., Adkins, J. N., Camp, n. D. G., Anderson, G. A. and Smith, R. D. (2008). DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics* **24** 1556–8.

[26] Pusch, W., Flocco, M. T., Leung, S. M., Thiele, H. and Kostrzewa, M. (2003). Mass spectrometry-based clinical proteomics. *Pharmacogenomics* **4** 463–76.

[27] Qian, W. J., Camp, D. G. and Smith, R. D. (2004). High-throughput proteomics using Fourier transform ion cyclotron resonance mass spectrometry. *Expert review of proteomics* **1** 87–95.

[28] Qian, W. J., Jacobs, J. M., Liu, T., Camp, D. G. and Smith, R. D. (2006). Advances and challenges in liquid chromatography-mass spectrometry-based proteomics profiling for clinical applications. *Molecular & cellular proteomics: MCP* **5** 1727–44.

[29] Qian, W. J., Petritis, B. O., Kaushal, A., Finnerty, C. C., Jeschke, M. G., Monroe, M. E., Moore, R. J., Schepmoes, A. A., Xiao, W., Moldawer, L. L., Davis, R. W., Tompkins, R. G., Herndon, D. N., Camp, n. D. G. and Smith, R. D. (2010). Plasma proteome response to severe burn injury revealed by 18O-labeled universal reference-based quantitative proteomics. *Journal of proteome research* **9** 4779–89.

[30] Shen, K. and Tseng, G. C. (2010). Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics* **26** 1316–23.

[31] Song, C. and Tseng, G. C. (2014). Hypothesis setting and order statistic for robust genomic meta-analysis. *The annals of applied statistics* **8** 777. MR3262534

[32] Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66** 187–205. MR2035766

[33] Tabb, D. L., Vega-Montoto, L., Rudnick, P. A., Variyath, A. M., Ham, A. J., Bunk, D. M., Kilpatrick, L. E., Billheimer, D. D., Blackman, R. K., Cardasis, H. L., Carr, S. A., Clauser, K. R., Jaffe, J. D., Kowalski, K. A., Neubert, T. A., Regnier, F. E., Schilling, B., Tegeler, T. J., Wang, M., Wang, P., Whiteaker, J. R., Zimmerman, L. J., Fisher, S. J., Gibson, B. W., Kinsinger, C. R., Mesri, M., Rodriguez, H., Stein, S. E., Tempst, P., Paulovich, A. G., Liebler, D. C. and Spiegelman, C. (2010). Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *Journal of proteome research* **9** 761–76.

[34] Tang, K., Page, J. S. and Smith, R. D. (2004). Charge competition and the linear dynamic range of detection in electrospray ionization mass spectrometry. *Journal of the American Society for Mass Spectrometry* **15** 1416–23.

[35] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics* **17** 520–5.

[36] Zhang, Y., Tibshirani, R. and Davis, R. (2012). Classification of patients from time-course gene expression. *Biostatistics* **14** 87–98.
[37] Zhang, Y., Tibshirani, R. J. and Davis, R. W. (2010). Predicting patient survival from longitudinal gene expression. *Statistical applications in genetics and molecular biology* **9** Article 41. MR2746023
[38] Zhou, B., Xu, W., Herndon, D., Tompkins, R., Davis, R., Xiao, W., Wong, W. H., Toner, M., Warren, H. S., Schoenfeld, D. A., Rahme, L., McDonald-Smith, G. P., Hayden, D., Mason, P., Fagan, S., Yu, Y. M., Cobb, J. P., Remick, D. G., Mannick, J. A., Lederer, J. A., Gamelli, R. L., Silver, G. M., West, M. A., Shapiro, M. B., Smith, R., Camp, n. D. G., Qian, W., Storey, J., Mindrinos, M., Tibshirani, R., Lowry, S., Calvano, S., Chaudry, I., Cohen, M., Moore, E. E., Johnson, J., Moldawer, L. L., Baker, H. V., Efron, P. A., Balis, U. G., Billiar, T. R., Ochoa, J. B., Sperry, J. L., Miller-Graziano, C. L., De, A. K., Bankey, P. E., Finnerty, C. C., Jeschke, M. G., Minei, J. P., Arnoldo, B. D., Hunt, J. L., Horton, J., Brownstein, B., Freeman, B., Maier, R. V., Nathens, A. B., Cuschieri, J., Gibran, N., Klein, M. and O'Keefe, G. (2010). Analysis of factorial time-course microarrays with application to a clinical study of burn injury. *Proceedings of the National Academy of Sciences of the United States of America* **107** 9923–8.
[39] Zimmer, J. S., Monroe, M. E., Qian, W. J. and Smith, R. D. (2006). Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass spectrometry reviews* **25** 450–82.

Yuping Zhang
Department of Statistics
University of Connecticut
Storrs, Connecticut 06269
USA
E-mail address: yuping.zhang@uconn.edu

Zhengqing Ouyang
Department of Biostatistics and Epidemiology
School of Public Health and Health Sciences
University of Massachusetts
Amherst, Massachusetts 01003
USA
E-mail address: ouyang@schoolph.umass.edu

Wei-Jun Qian
Biological Sciences Division and
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
Richland, Washington 99352
USA
E-mail address: weijun.qian@pnnl.gov

Richard D. Smith
Biological Sciences Division and
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
Richland, Washington 99352
USA
E-mail address: rds@pnnl.gov

Wing Hung Wong
Department of Statistics
Stanford University
Stanford, California 94305
USA
E-mail address: whwong@stanford.edu

Ronald W. Davis
Stanford Genome Technology Center
Stanford University
Palo Alto, California 94306
USA
E-mail address: krhong@stanford.edu