The Annals of Applied Statistics
2017, Vol. 11, No. 1, 114–138
DOI: 10.1214/16-AOAS994

A MIXED-EFFECTS MODEL FOR INCOMPLETE DATA FROM LABELING-BASED QUANTITATIVE PROTEOMICS EXPERIMENTS

BY LIN S. CHEN[*,1,2], JIEBIAO WANG[*,1,2], XIANLONG WANG[†,3] AND PEI WANG[‡,1,3,4]

University of Chicago[*], Fred Hutchinson Cancer Research Center[†] and Icahn School of Medicine at Mount Sinai[‡]

In mass spectrometry (MS) based quantitative proteomics research, the emerging iTRAQ (isobaric tag for relative and absolute quantitation) and TMT (tandem mass tags) techniques have been widely adopted for high throughput protein profiling. In a typical iTRAQ/TMT proteomics study, samples are grouped into batches, and each batch is processed by one multiplex experiment, in which the abundances of thousands of proteins/peptides in a batch of samples can be measured simultaneously. The multiplex labeling technique greatly enhances the throughput of protein quantification. However, the technical variation across different iTRAQ/TMT multiplex experiments is often large due to the dynamic nature of MS instruments. This leads to strong batch effects in the iTRAQ/TMT data. Moreover, the iTRAQ/TMT data often contain substantial batch-level nonignorable missing entries. Specifically, the abundance measures of a given protein/peptide are often either observed or missing altogether in all the samples from the same batch, with the missing probability depending on the combined batch-level abundances. We term this unique missing-data mechanism as the Batch-level Abundance-Dependent Missing-data Mechanism (BADMM). We introduce a new method—`mixEMM`—for analyzing iTRAQ/TMT data with batch effects and batch-level nonignorable missingness. The `mixEMM` method employs a linear mixed-effects model and explicitly models the batch effects and the BADMM. With simulation studies, we showed that, compared with existing approaches that utilize relative abundances and ignore the missing batches under the missing-completely-at-random assumption, the `mixEMM` method achieves more accurate parameter estimation and inference. We applied the method to an iTRAQ proteomics data from a breast cancer study and identified phosphopeptides differentially expressed between different breast cancer subtypes. The method can be applied to general clustered data with cluster-level nonignorable missing-data mechanisms.

Received June 2015; revised September 2016.

[1]Supported in part by R01GM108711 and U24 CA210993.

[2]Supported in part by R03CA174984.

[3]Supported in part by SUB-CA160034.

[4]Supported in part by NIH Grant P01CA53996.

Key words and phrases. Mixed-effects models, the expectation-conditional-maximization (`ECM`) algorithm, Batch-level Abundance-Dependent Missing-data Mechanism (BADMM).

1. Introduction.

1.1. *Quantitative proteomics research based on iTRAQ/TMT data.* Proteins are complex macromolecules responsible for nearly every task of cellular life and essential for the structures and functions of human tissues and organs. However, the discovery of protein biomarkers in cancer diagnosis, prevention and treatment has achieved only modest success, partially because the abundances of proteins are difficult to quantify. To date, MS-based platforms still serve as the workhorses in quantitative proteomics research. Traditional high-throughput mass-spectrometry (MS) experiments usually process samples one by one; and the process of each sample involves extensive fractionation, resulting in weeks of experimental time. The long time and high cost required for such experiments greatly limit the scale of most proteomics studies.

To improve the efficiency of MS-based protein quantification, labeled multiplex proteomics experiments, such as the iTRAQ (isobaric Tag for relative and Absolute Quantitation) and TMT (tandom mass tags), were introduced about a decade ago and have become increasingly popular in recent years [Ross et al. (2004), Werner et al. (2014), Wiese et al. (2007)]. For example, in an iTRAQ-MS-based study, samples are first grouped into batches (4 or 8 samples per batch), and then each batch is processed by one iTRAQ multiplex experiment consisting of three steps: (1) intact proteins of each sample are enzymatically digested into smaller segments of amino acid sequences, that is, peptides; (2) peptides from different samples in one batch are labeled with different isotope-coded covalent tags and are mixed together; (3) the mixtures are introduced into MS instruments, where peptides from different samples in the same batch are identified and quantified together. In this way, multiple samples can be processed together, which greatly reduces the overall quantification time and cost. For instance, Paulo et al. (2014) successfully quantified the protein abundances in nine mice using 10-plex TMT experiments across three tissue types and studied the effects of two mitogen-activated protein/extracellular signal-regulated kinase inhibitors on protein abundances. In addition, McAlister et al. (2014) have shown that multiplexed quantitation via iTRAQ/TMT enables more accurate quantification of protein/peptide abundances. Other successful examples include Franken et al. (2015), in which the authors used 10-plex TMT experiments to examine the changes in protein thermal stability across the proteome. Rauniyar and Yates III (2014) reviewed the studies based on multiplexed experiments and suggested combinations of experimental design and optimal data acquisition methods to increase the precision and accuracy for obtaining the relative protein abundances in multiplexed quantitative proteomics studies.

1.2. *Motivating iTRAQ proteomics data from the CPTAC project.* To improve our ability to diagnose, treat and prevent cancer, the National Cancer Institute launched the Clinical Proteomic Tumor Analysis Consortium (CPTAC,

http://proteomics.cancer.gov) to systematically identify proteins that are derived from alterations in cancer genomes [Ellis et al. (2013), Liebler et al. (2014), Mertins et al. (2016), Paulovich et al. (2010), Zhang et al. (2016)]. The CPTAC has recently conducted global proteome and phosphoproteome profiling of a subset of breast, colon and ovarian cancer samples that have been extensively characterized in The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov) [The Cancer Genome Atlas Network (2012)]. So far, this is the first attempt to characterize protein activities in cancer samples using sophisticated proteomics experiments on a large scale. Specifically, in the breast cancer project, 108 breast cancer tumor samples from 105 breast cancer patients (three of them have two tumor samples) have been analyzed with iTRAQ experiments to identify proteins related to breast cancer clinical variables and outcomes.

Another aim of the CPTAC project is to "set standards, establish procedures, and provide reagents to enable cancer researchers to effectively and reproducibly use proteomics approaches" [Ellis et al. (2013), Paulovich et al. (2010)]. Advances in methods and tools, especially the ones accounting for the unique characteristics of proteomics data like the method proposed in this paper, will better facilitate the achievement of those missions and will in turn lead to improved diagnostics, therapies and potentially preventive measures for cancer.

In this paper, we will focus on analyzing the phosphoproteomics data from the breast cancer CPTAC study. Phosphorylation is a key post-translational modification and plays a central role in many biological processes. Phosphorylation at different sites of one protein could induce different biological activities. Our goal is to identify individual phosphorylated peptides, that is, phosphopeptide, up- or down-regulated in triple negative breast cancer tumors compared to other subtypes of breast cancer. The investigation will provide important insights into breast cancer etiology and help identify protein biomarkers.

1.3. *Batch effect and batch-level nonignorable missing data.* Given the popularity and the efficiency of the iTRAQ/TMT technique, there is a pressing need for tailored methods for analyzing data from multiplex iTRAQ/TMT experiments. Though the iTRAQ/TMT-based batch-processing greatly reduces the cost and improves the efficiency of data generation, the consequent batch effects are substantial due to the dynamic nature of the MS instrument. To alleviate this problem, a general practice is to include a common reference sample in each batch for quality control. For example, in the 4-plex iTRAQ experiments of the CPTAC breast cancer study, each batch consisted of 3 breast tumor samples and a common reference sample. The reference sample was created by combining 40 tumor samples in the CPTAC breast cancer study.

Conventional approaches analyze the relative abundances of proteins/peptides in the target samples relative to the reference sample in the same batch. For example, in the aforementioned works [Franken et al. (2015), McAlister et al. (2014), Paulo et al. (2014), Rauniyar and Yates III (2014)], after proper normalization

across multiplex experiments, relative abundances were used to perform ANOVA, t-test or fold-change analyses. Largely, this strategy assumes that, in each multiplex run, experimental noises affect the reference sample in the same way as they affect the targeted samples, and thus calculating relative abundances removes the batch-level experimental variations. However, due to the complicated process of protein/peptide identification and quantification in the MS instruments, the target samples and reference sample could be subject to different experimental variations [Karp et al. (2010)]. Therefore, relative abundance measures cannot fully capture these data features.

Furthermore, most of the analyses in the literature were performed based on only observed data, and the missing protein abundances were largely ignored. It is well known that, in the general MS experiments, the lower the abundance of a given peptide, the more likely the peptide is missing in the output data [Chen, Prentice and Wang (2014), Wang et al. (2006)]. This missing-data mechanism is nonignorable [Rubin (1976)], and ignoring those missing protein/peptide abundances may lead to biased estimation and inference. More uniquely, data from iTRAQ/TMT-MS experiments often have a substantial amount of "batch-level" nonignorable missing data. Since all the samples in a batch are processed together in these experiments, a given peptide is either detected and quantified or missing from all the samples in the same batch. The missing probability of a peptide largely depends on the combined abundances of the peptides from all the samples in the same batch (the batch-level abundance). We term this missing-data mechanism the "Batch-level Abundance-Dependent Missing-data Mechanism (BADMM)." Figure 1 shows an illustration of the iTRAQ data on one peptide and its BADMM. Subsequently, protein quantification is often obtained as a summary of the peptide abundances in the protein and is also subject to the BADMM. In addition to BADMM, sporadic missingness may occur at the individual sample level. Sporadic missingness refers to the scenario where a peptide/protein is missing in some but not all of the samples in the same batch. Since the proportion of sporadic missing data is usually small (e.g., $<1\%$ sporadic versus $>99\%$ batch-level missingness in the motivating CPTAC data set), we assume these sporadic missing-values are missing-completely-at-random and are ignorable [Rubin (1976)].

Given the presence of substantial batch effects, batch-level missingness (about 50%–80% per sample in our motivating CPTAC data) and small sample sizes in most labeled proteomics data, it is essential to account for the batch design and the nonignorable missingness deliberately to improve the precision of estimation and inference with such data. In this work, we propose to directly model the absolute abundances of proteins/peptides and their variance structures due to the batch design. By modeling the absolute abundances instead of the relative abundances, we can better characterize the variance of protein abundances in target samples and improve the power of statistical tests. This strategy has been employed for analyzing other types of proteomics data from mass spectrometry experiments [Chang

Experiment	Reference sample	Sample 1	Sample 2	Sample 3	Experiment random-effects	Missing Indicator
1	Y_{1R}	Y_{11}	Y_{12}	Y_{13}	b_1	$M_1 = 0$
2	Y_{2R}	Y_{21}	Y_{22}	Y_{23}	b_2	$M_2 = 1$
i	Y_{iR}	Y_{i1}	Y_{i2}	Y_{i3}	b_i	$M_i = 0$
N	Y_{NR}	Y_{N1}	Y_{N2}	Y_{N3}	b_N	$M_N = 0$

FIG. 1. *An illustration of a* 4*-plex iTRAQ data matrix of one peptide. Let* $\mathbf{Y}_{N\times 4}$ *be the abundance data for the peptide. And* $3 \times N$ *tumor samples are randomly grouped into* N *batches and are processed by* N *iTRAQ experiments. In each iTRAQ experiment* i $(i = 1, \ldots, N)$*, besides three tumor samples, a common reference sample is also processed together. Due to the sampling mechanism in mass spectrometry instruments, usually a peptide is either observed or missing in all four samples in one experiment. If missing, the missing indicator for the* i*th batch,* M_i*, is set to be* 1*. The missing probability of the batch relates to the total peptide abundance level in the batch. The lower the total abundance, the more likely the peptide will be missing in the experiment (batch).*

et al. (2012)]. Since samples in the same batch are subject to the same experimental conditions and procedures, a mixed-effects model with a random effect for each batch is a natural way to account for the experimental design [Laird and Ware (1982)].

The BADMM in the iTRAQ/TMT data hinders the direct application of a mixed-effects model. With BADMM, the probability of a protein/peptide being missing in a batch depends on the combined abundance of the protein/peptide in the batch. The missing data are not missing at random and are nonignorable [Rubin (1976)]. To obtain unbiased estimation and valid inference, the missing-data mechanism needs to be properly modeled and accounted for. Existing work on modeling the nonignorable missingness in iTRAQ/TMT data [Hill et al. (2008), Luo et al. (2009)] and the selection model for longitudinal data with nonignorable missingness [Ibrahim and Molenberghs (2009)] consider the probability of missingness for a protein/peptide in each sample independently. Since the probability of missingness for a protein/peptide is not independent between samples within the same batch, a simple linear shift or quantile normalization may not appropriately normalize the data. Oberg et al. (2008) proposed an approach which iterates between estimating the batch/sample effects and estimating the protein/peptide effects. Specifically, the batch/sample effects were estimated using the entire set of data and were considered fixed for computational feasibility. In contrast, we propose to model the batch-level missing-data pattern (BADMM) and incorporate it into a mixed-effects model. We model the probability of a protein/peptide being missing (in all of the samples) in a batch as a function of the total protein/peptide

abundance in the batch. This probabilistic missing-data mechanism provides an attractive way to account for the characteristics of iTRAQ/TMT and MS experimental complexities. Compared to a censoring model [Little and Rubin (2002)], it does not depend on a fixed detection threshold; it is more flexible and better depicts the experimental procedure.

1.4. *Outline.* To properly analyze iTRAQ/TMT data as characterized by the output from the CPTAC project, we introduce `mixEMM`—a mixed-effects model coupled with the probabilistic BADMM in Section 2. In Section 3, we use an Expectation and Conditional Maximization (`ECM`) algorithm to estimate the fixed and random effects in `mixEMM`. We also present an alternative probability function for BADMM that may be suitable for more general settings. In Section 4, we perform simulations to evaluate the performance of the `mixEMM` method. In Section 5, we apply the proposed method to the motivating CPTAC iTRAQ data and identify phosphopeptides related to breast cancer subtypes. In Section 6, we summarize the work as a useful tool for analyzing iTRAQ/TMT proteomics data and, moreover, as a general framework to handle cluster-level nonignorable missing-data patterns for data with repeated or clustered measures.

2. A mixed-effects model for data with batch-level nonignorable missingness. First of all, we assume the data to be analyzed have been properly preprocessed and normalized. In the motivating CPTAC breast cancer data, each phospho-modification of one peptide (i.e., each phosphosite) is viewed as an analysis unit because different modifications can result in different protein functions. We refer to these analysis units as *features* in this and the following sections. For one feature of interest, let $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{Q}$ denote the complete (observed and missing) abundances for this feature in all of the samples in the Q batches (in the CPTAC data, $Q = 36$ and $\mathbf{Y}$ is a vector of $\sum_{i=1}^{Q} p_i = 144$ elements). Specifically, $\mathbf{y}_i$ is a $p_i \times 1$ data vector of the ith batch, where p_i is the number of samples in the batch, y_{i1} represents the abundance in the reference sample, and $y_{i2}, \ldots, y_{ip_i}$ represents the abundances in the targeted samples in the batch. Note that since the raw abundance measurements from mass spectrometry instruments often follow a very heavy-tailed distribution, the raw abundance measurements are usually subject to log transformation in the data preprocessing. If so, then $\mathbf{Y}$ represents the log transformed abundances.

Suppose this feature is observed in only Q_{obs} batches ($Q_{\text{obs}} \le Q$). Let $\mathbf{y}_{\text{obs}}$ and $\mathbf{y}_{\text{mis}}$ denote the observed and the missing data, respectively, and $\mathbf{Y} = \{\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}\}$. Samples in the same batch are processed by one multiplex experiment, are subject to the same experimental procedure, and are correlated. We use a linear mixed-effects model to account for such correlations:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \tag{1}$$

where $\mathbf{X}_i$ is a known fixed design matrix with dimension $p_i \times k$, $\boldsymbol{\alpha}$ is a $k \times 1$ vector of parameters for fixed effects, $\mathbf{Z}_i$ is a known covariate matrix of dimension $p_i \times h$

for random effects, $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}_{h\times h})$ represents the random effect coefficient specific to each batch of samples, and $\mathbf{e}_i \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \mathbf{R}_i)$ is a diagonal covariance matrix of dimension $p_i \times p_i$. In our data application, $\mathbf{X}_i$ consists of a column of 1s, an indicator variable for the reference sample, and a set of clinical variables (for example, cancer subtype indicators); $\mathbf{b}_i$ is of length 1 ($h = 1$); $\mathbf{Z}_i$ is a vector of 1s; and for 4-plex iTRAQ experiments, $\mathbf{R}_i$ has diagonal elements $\{\sigma_0^2, \sigma^2, \sigma^2, \sigma^2\}$, where σ_0^2 is the variance corresponding to the reference sample and σ^2 is the variance of the other three samples. Since the reference sample was created by combining 40 tumor samples in the CPTAC breast cancer study, we expect it to have a different variance than other individual tumor samples.

According to (1), we have $\mathbf{y}_i \sim N(\mathbf{X}_i\boldsymbol{\alpha}, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \mathbf{R}_i$. Our goals are to obtain the maximum likelihood estimates (MLEs) of the fixed and random effects while accounting for the nonignorable BADMM and to draw inferences on the fixed effects for identifying features related to clinical variables ($\mathbf{X}_i$).

As described in the previous section, for a given feature, the lower its combined abundance across all of the samples in one batch, the more likely all measures of the feature in the batch will be missing during the experiment. Let M_i be the missing indicator of this feature in the ith batch: $M_i = 1$ if the feature is missing in the ith batch, and $M_i = 0$ otherwise. We model this BADMM using an exponential probabilistic model:

$$\Pr(M_i = 1|\mathbf{y}_i) = g\big(\mathbf{1}^T\mathbf{y}_i; \gamma_0, \gamma\big) = \exp\big(-\gamma_0 - \gamma \cdot \mathbf{1}^T\mathbf{y}_i - \boldsymbol{\gamma}_2 \cdot \mathbf{C}_i\big), \tag{2}$$

where γ_0 is a scaling constant determined by the missing rate of each protein/peptide; and γ models the relationship between protein/peptide abundances versus missing probabilities. In (2), $\mathbf{C}_i$ is a set of covariates associated with the experiment i (or the ith batch) and $\boldsymbol{\gamma}_2$ is the corresponding coefficient. In our motivating example, we do not have any experiment-specific (or batch-specific) covariate, and thus the last term is not considered. Since $\mathbf{y}_i$ is an abundance measure and all positive, $\mathbf{y}_i > 0$, the missing-data parameters (γ_0, γ) are non-negative, and the above probability function always takes value between 0 and 1. When $\gamma > 0$, the missing-data mechanism is nonignorable, and when $\gamma = 0$, the missing-data mechanism is missing at random or missing completely at random if no batch-level covariates are considered.

We first treat γ_0 and γ as known missing-data mechanism parameters. We discuss extensions to scenarios in which those parameters are unknown in Section 3.3. Moreover, in Section 3.5, we discuss other flexible probability functions for BADMM.

3. An `ECM` algorithm to calculate MLEs. To obtain the MLEs that maximize the observed-data likelihood function considering the missing-data mechanism, we employ an `ECM` algorithm and term the proposed method `mixEMM` (Mixed-Effects Models with BADMM).

Let $\boldsymbol{\Omega} = \{\boldsymbol{\alpha}, \sigma_0^2, \sigma^2, \mathbf{D}\}$ denote the set of parameters of interest. If $\mathbf{y}_{\text{mis}}$ and $\mathbf{b}_i$ were observed, the MLEs for $\mathbf{R}_i$, $\mathbf{D}$ and $\boldsymbol{\alpha}$ based on the likelihood of the complete data $(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \mathbf{b}, \mathbf{M})$ can be easily calculated. Thus, we employ an ECM algorithm [Meng and Rubin (1993)]: in the expectation (E) step of the $(t+1)$th iteration, we calculate $Q(\boldsymbol{\Omega}|\boldsymbol{\Omega}^{(t)})$—the expected value of the log-likelihood given the observed data and current parameter estimates. In the conditional maximization (CM) step, we obtain the current parameter estimates $\hat{\boldsymbol{\Omega}}^{(t+1)}$ by maximizing $Q(\boldsymbol{\Omega}|\boldsymbol{\Omega}^{(t)})$. Given the proposed BADMM in equation (2), closed-form solutions are available in the CM step. By iterating through the E and CM steps, the likelihood of the observed data will always increase, and we will obtain the MLEs at the convergence [Chen, Prentice and Wang (2014)].

3.1. *E step.* In the E step, the expected log-likelihood function for the complete data given the observed data and the current parameter estimates can be written as

$$\begin{aligned} Q(\boldsymbol{\Omega}|\boldsymbol{\Omega}^{(t)}) &= \mathrm{E}_{\mathbf{y}_{\text{mis}},\mathbf{b}|\mathbf{y}_{\text{obs}},\mathbf{M};\boldsymbol{\Omega}^{(t)}}\big[\log L(\boldsymbol{\Omega}; \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \mathbf{b}, \mathbf{M})\big] \\ &= \sum_{i\in\mathbf{O}} \mathrm{E}_{\mathbf{b}_i|\mathbf{y}_i,M_i;\boldsymbol{\Omega}^{(t)}}\ell(\mathbf{y}_i, \mathbf{b}_i, M_i = 0; \boldsymbol{\Omega}) \\ &\quad + \sum_{i\notin\mathbf{O}} \mathrm{E}_{\mathbf{y}_i,\mathbf{b}_i|M_i;\boldsymbol{\Omega}^{(t)}}\ell(\mathbf{y}_i, \mathbf{b}_i, M_i = 1; \boldsymbol{\Omega}) \\ &= I1 + I2, \end{aligned}$$

where $\mathbf{O}$ denotes the set of indices of the observed batches. Existing literature on modeling the nonignorable missingness in iTRAQ/TMT data [Hill et al. (2008), Luo et al. (2009)] and the selection model for longitudinal data with nonignorable missingness [Ibrahim and Molenberghs (2009)] consider the probability of a feature's missingness in each sample independently. Those methods can handle sample-level nonignorable missing data, but are not directly applicable to iTRAQ/TMT data with batch-level missingness. In contrast, we will take the missing batches into account by explicitly modeling the BADMM—a major innovation of the proposed method.

For the observed batches,

$$\begin{aligned} I1 = \sum_{i\in\mathbf{O}} \mathrm{E}_{\mathbf{b}_i|\mathbf{y}_i,M_i;\boldsymbol{\Omega}^{(t)}}\big\{&\log\big[f(\mathbf{y}_i|\boldsymbol{\alpha}, \mathbf{R}_i, \mathbf{b}_i)\big] + \log\big[f(\mathbf{b}_i|\mathbf{D})\big] \\ &+ \log\big[f(M_i = 0|\mathbf{y}_i)\big]\big\}. \end{aligned}$$

The last term $\log[f(M_i = 0|\mathbf{y}_i)]$ does not involve parameters of interest.

To obtain the conditional expectation, we first calculate the conditional distribution of $\mathbf{b}_i$ for $i \in \mathbf{O}$ as a normal distribution with mean and variance

$$\mathbf{b}_i^{(t)} = \mathrm{E}\big(\mathbf{b}_i|\mathbf{y}_i, M_i = 0, \boldsymbol{\Omega}^{(t)}\big) = \mathbf{D}^{(t)}\mathbf{Z}_i^T\mathbf{W}_i^{(t)}\big(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\alpha}^{(t)}\big), \tag{3}$$

$$\boldsymbol{\Delta}_i^{(t)} = \mathrm{var}\big(\mathbf{b}_i|\mathbf{y}_i, M_i = 0, \boldsymbol{\Omega}^{(t)}\big) = \mathbf{D}^{(t)} - \mathbf{D}^{(t)}\mathbf{Z}_i^T\mathbf{W}_i^{(t)}\mathbf{Z}_i\mathbf{D}^{(t)}, \tag{4}$$

where $\mathbf{W}_i^{(t)} = (\boldsymbol{\Sigma}_i^{(t)})^{-1} = (\mathbf{Z}_i\mathbf{D}^{(t)}\mathbf{Z}_i^T + \mathbf{R}_i^{(t)})^{-1}$. It follows that

$$I1 = \text{const} - 1/2 \sum_{i\in\mathbf{O}} \big(\log|\mathbf{R}_i| + \big(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\alpha} - \mathbf{Z}_i\mathbf{b}_i^{(t)}\big)^T \mathbf{R}_i^{-1}\big(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\alpha} - \mathbf{Z}_i\mathbf{b}_i^{(t)}\big) + \operatorname{tr}\big(\mathbf{V}_i^{(t)}\mathbf{R}_i^{-1}\big) + \log|\mathbf{D}| + \mathbf{b}_i^{(t)T}\mathbf{D}^{-1}\mathbf{b}_i^{(t)} + \operatorname{tr}\big(\mathbf{D}^{-1}\boldsymbol{\Delta}_i^{(t)}\big)\big),$$

where $\mathbf{V}_i^{(t)} = \operatorname{var}(\mathbf{e}_i|\mathbf{y}_i, M_i = 0, \boldsymbol{\Omega}^{(t)}) = \mathbf{Z}_i\boldsymbol{\Delta}^{(t)}\mathbf{Z}_i$ for $i \in \mathbf{O}$.

To calculate $I2$, we first compute the conditional expectation and variance of $\mathbf{y}_i$ and $\mathbf{b}_i$ for $i \notin \mathbf{O}$. Given $\Pr(M_i = 1|\mathbf{y}_i)$ in equation (2), it is easy to see that, for $i \notin \mathbf{O}$,

$$\mathbf{y}_i^{(t)} = \mathrm{E}\big(\mathbf{y}_i|M_i = 1, \boldsymbol{\Omega}^{(t)}\big) = \mathbf{X}_i\boldsymbol{\alpha}^{(t)} - \gamma\boldsymbol{\Sigma}_i^{(t)}\mathbf{1}, \tag{5}$$

$$\operatorname{var}\big(\mathbf{y}_i|M_i = 1, \boldsymbol{\Omega}^{(t)}\big) = \boldsymbol{\Sigma}_i^{(t)}, \tag{6}$$

where $\boldsymbol{\Sigma}_i^{(t)} = \mathbf{Z}_i\mathbf{D}^{(t)}\mathbf{Z}_i^T + \mathbf{R}_i^{(t)}$. It follows that, for $i \notin \mathbf{O}$,

$$\begin{aligned}\mathbf{b}_i^{(t)} &= \mathrm{E}\big(\mathbf{b}_i|M_i = 1, \boldsymbol{\Omega}^{(t)}\big) = \mathrm{E}\big(\mathrm{E}\big(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\Omega}^{(t)}\big)|M_i = 1, \boldsymbol{\Omega}^{(t)}\big)\\ &= \mathbf{D}^{(t)}\mathbf{Z}_i^T\mathbf{W}_i^{(t)}\big(\mathbf{y}_i^{(t)} - \mathbf{X}_i\boldsymbol{\alpha}^{(t)}\big),\end{aligned} \tag{7}$$

$$\begin{aligned}\boldsymbol{\Delta}_i^{(t)} &= \operatorname{var}\big(\mathbf{b}_i|M_i = 1, \boldsymbol{\Omega}^{(t)}\big)\\ &= \mathrm{E}\big(\operatorname{var}\big(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\Omega}^{(t)}\big)|M_i = 1, \boldsymbol{\Omega}^{(t)}\big) + \operatorname{var}\big(\mathrm{E}\big(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\Omega}^{(t)}\big)|M_i = 1, \boldsymbol{\Omega}^{(t)}\big)\\ &= \mathbf{D}^{(t)},\end{aligned} \tag{8}$$

$$\mathbf{V}_i^{(t)} = \operatorname{var}\big(\mathbf{e}_i|M_i = 1, \boldsymbol{\Omega}^{(t)}\big) = \mathbf{R}_i^{(t)}. \tag{9}$$

Then we can obtain the following for the missing batches of samples:

$$\begin{aligned}I2 &= \sum_{i\notin\mathbf{O}} \mathrm{E}_{\mathbf{y}_i,\mathbf{b}_i|M_i;\boldsymbol{\Omega}^{(t)}}\big\{\log\big[f(\mathbf{y}_i|\boldsymbol{\alpha}, \mathbf{R}_i, \mathbf{b}_i)\big] + \log\big[f(\mathbf{b}_i|\mathbf{D})\big] + \log\big[f(M_i = 1|\mathbf{y}_i)\big]\big\}\\ &= \text{const} - 1/2 \sum_{i\notin\mathbf{O}} \big(\log|\mathbf{R}_i| + \big(\mathbf{y}_i^{(t)} - \mathbf{X}_i\boldsymbol{\alpha} - \mathbf{Z}_i\mathbf{b}_i^{(t)}\big)^T \mathbf{R}_i^{-1}\big(\mathbf{y}_i^{(t)} - \mathbf{X}_i\boldsymbol{\alpha} - \mathbf{Z}_i\mathbf{b}_i^{(t)}\big)\\ &\quad + \operatorname{tr}\big(\mathbf{V}_i^{(t)}\mathbf{R}_i^{-1}\big) + \log|\mathbf{D}| + \mathbf{b}_i^{(t)T}\mathbf{D}^{-1}\mathbf{b}_i^{(t)} + \operatorname{tr}\big(\mathbf{D}^{(-1)}\mathbf{D}^{(t)}\big) + 2\gamma\cdot\mathbf{1}^T\mathbf{y}_i^{(t)}\big).\end{aligned}$$

3.2. *CM step.* In the CM step, we sequentially maximize the expected complete-data log-likelihood for the parameters of interest. In the first step of CM, we obtain the estimate for $\mathbf{D}$ that maximizes $Q(\boldsymbol{\Omega}|\boldsymbol{\Omega}^{(t)})$:

$$\mathbf{D}^{(t+1)} = \frac{1}{Q}\sum_{i=1}^{Q}\big(\mathbf{b}_i^{(t)}\mathbf{b}_i^{(t)T} + \boldsymbol{\Delta}_i^{(t)}\big). \tag{10}$$

Then, conditioned on the current $\mathbf{R}_i^{(t)}$, the estimate for $\boldsymbol{\alpha}$ is given by

$$\boldsymbol{\alpha}^{(t+1)} = \left(\sum_{i=1}^{Q} \mathbf{X}_i^T \left(\mathbf{R}_i^{(t)}\right)^{-1} \mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{Q} \mathbf{X}_i^T \left(\mathbf{R}_i^{(t)}\right)^{-1} \left(\mathbf{y}_i^{(t)} - \mathbf{Z}_i \mathbf{b}_i^{(t)}\right)\right), \tag{11}$$

where $\mathbf{y}_i^{(t)} = \mathbf{y}_i$ when $M_i = 0$, and $\mathbf{y}_i^{(t)} = \mathbf{X}_i \boldsymbol{\alpha}^{(t)} - \gamma \boldsymbol{\Sigma}_i^{(t)} \mathbf{1}$ when $M_i = 1$.

Last, we can obtain the estimates for $\sigma_0^{2(t+1)}$ and $\sigma^{2(t+1)}$ conditioned on $\boldsymbol{\alpha}^{(t+1)}$:

$$\sigma_0{}^{2(t+1)} = \frac{1}{Q} \sum_{i=1}^{Q} \left[\left(y_{i1}^{(t)} - \mathbf{X}_{i1} \boldsymbol{\alpha}^{(t+1)} - \mathbf{Z}_{i1} \mathbf{b}_i^{(t)}\right)^2 + v_{i11}^{(t)}\right], \tag{12}$$

and

$$\begin{aligned} \sigma^{2(t+1)} = \Bigg\{ \sum_{i=1}^{Q} \Bigg[\sum_{j=2}^{p_i} \left(y_{ij}^{(t)} - \mathbf{X}_{ij} \boldsymbol{\alpha}^{(t+1)} - \mathbf{Z}_{ij} \mathbf{b}_i^{(t)}\right)^2 \\ + \left(\operatorname{tr} \mathbf{V}_i^{(t)} - v_{i11}^{(t)}\right) \Bigg] \Bigg\} \Bigg/ \left(\sum_{i=1}^{Q} p_i - Q\right), \end{aligned} \tag{13}$$

where $v_{i11}^{(t)}$ denotes the first diagonal element of $\mathbf{V}_i^{(t)}$. By iterating through the E- and CM-steps, MLEs for the fixed effects and variance components can be obtained.

In addition, through computing the information matrix of the log-likelihood function of the observed data, we can estimate the variance of $\hat{\boldsymbol{\alpha}}$ using

$$\widehat{\operatorname{var}}(\hat{\boldsymbol{\alpha}}) = \left(\sum_{i \in \mathbf{O}} \mathbf{X}_i \mathbf{W}_i \mathbf{X}_i\right)^{-1}. \tag{14}$$

We can then perform the Wald test to detect nonzero $\boldsymbol{\alpha}$.

3.3. *Estimation of the missing-data mechanism parameter.* In real applications, the missing-data mechanism parameter $\boldsymbol{\Gamma} = \{\gamma_0, \gamma\}$ in (2) is often unknown and needs to be estimated. One simple approach is to use the missing percentage and sum of abundance based on available data of each feature to model the relationship between the probability of missingness and the abundance. Specifically, we assume all of the features in one data set are subject to the same missing-data mechanism. We calculate the average batch-level abundance for each feature j based on the observed data and denote it as t_j and also obtain the missing percentage of feature j as $\pi_j = 1 - Q_{j,\text{obs}}/Q$, where $Q_{j,\text{obs}}$ is the number of batches in which feature j is quantified. We can estimate $\boldsymbol{\Gamma}$ in (2) by

$$\hat{\boldsymbol{\Gamma}} = \underset{\boldsymbol{\Gamma} = \{\gamma_0, \gamma\}}{\arg\min} \sum_j \left(\log(\pi_j) + \gamma_0 + \gamma t_j\right)^2. \tag{15}$$

Alternatively, one can also employ the profile likelihood approach proposed in Chen, Prentice and Wang (2014) to jointly estimate the parameters of interest and the missing-data mechanism parameters. Let $L_{\mathbf{\Gamma}}(\mathbf{\Omega}) = L(\mathbf{y}_{\text{obs}}, \mathbf{M}; \mathbf{\Omega}, \mathbf{\Gamma})$. One can evaluate $L_{\mathbf{\Gamma}}(\mathbf{\Omega})$ at different $\mathbf{\Gamma}$ values and choose the $\mathbf{\Gamma}$ that gives the maximum over the likelihood profile. As shown in Chen, Prentice and Wang (2014) with both simulations and real data examples, the estimated $\mathbf{\Gamma}$ based on available case estimates of protein abundance is very close to the profile likelihood estimates, especially when the sample size is limited as in most proteomics studies. Moreover, in Section 4.3, we demonstrate that the available case estimate of $\mathbf{\Gamma}$ is very close to the true values under all the simulation settings considered in this paper. Thus, we use the available case estimates of the missing-data mechanism parameter in our data analysis.

3.4. *An outline of the algorithm to fit the* `mixEMM` *model*. In summary, we implement an `ECM` algorithm to fit the `mixEMM` model for analyzing iTRAQ/TMT proteomics data. An outline of the `ECM` algorithm is provided in Algorithm 1. Note that, for the small amount of sporadic missingness, we treat them as missing-completely-at-random [Rubin (1976)] and remove the corresponding data points from the evaluation of the likelihood function. Specifically, if a protein is measured in l ($l < 4$) samples in a 4-plex iTRAQ experiment, we will set $p_i = l$ and apply the proposed method.

3.5. *Logit probability functions for BADMM*. The probability of missingness in (2) is designed to characterize the BADMM for abundance data from iTRAQ/TMT or other proteomics experiments. By using an exponential function, the probability of missingness in (2) can be naturally integrated with the density function of normal distributions. Thus, closed-form solutions can be obtained in the `ECM` algorithm, which makes the computation efficient.

Alternatively, a logistic function is often used to model the probability of missingness as a function of protein/peptide abundances and other experiment-specific

Algorithm 1 An algorithm to fit the `mixEMM` model

1. Estimate missing-data mechanism parameter $\mathbf{\Gamma}$ by (15).
2. Obtain the initial estimate $\mathbf{\Omega}^{(0)}$ for fixed effects and variance components.
3. E-step: For the exponential missing-data mechanism function, given $\hat{\mathbf{\Gamma}}$, calculate the conditional expectations and variances of $\mathbf{y}_{\text{mis}}, \mathbf{e}_i, \mathbf{b}_i$ given the observed $\mathbf{y}_{\text{obs}}$, $\mathbf{M}$, and the current parameter estimates $\hat{\mathbf{\Omega}}^{(t-1)}$ according to (3), (4), (5) and (6).
4. CM-step: Given the estimated sufficient statistics, obtain the current estimates of $\mathbf{D}$, $\boldsymbol{\alpha}$, σ_0^2, and σ^2 using (10), (11) and (12) and (13), respectively.
5. Repeat 3–4 until convergence.

covariates [Little and Rubin (2002), Luo et al. (2009)]:

$$\text{logit}\big(\Pr(M_i = 1|\mathbf{y}_i)\big) = \gamma_0 + \gamma \cdot \mathbf{1}^T\mathbf{y}_i + \boldsymbol{\gamma}_2 \cdot \mathbf{C}_i. \tag{16}$$

The interpretations of the missing-data mechanism parameters, γ_0, γ and $\boldsymbol{\gamma}_2$ are similar to those for the exponential missing-data mechanism in (2), except that those parameters are not required to be non-negative in the logit missing-data mechanism.

For the logit missing-data mechanism in (16), we will use numeric integration [Pinheiro and Bates (1995)] to obtain the conditional means and variances for $\mathbf{y}_i^{(t)}$'s in the missing batches, and replace the corresponding terms in (7), (8) and (9) with the following:

$$\begin{aligned}
\mathbf{y}_i^{(t)} &= \mathrm{E}\big(\mathbf{y}_i|M_i = 1, \boldsymbol{\Omega}^{(t)}\big) \\
&= \frac{\int \mathbf{y}_i P(M_i = 1|\mathbf{y}_i)\phi(\mathbf{y}_i, \mathbf{X}_i\boldsymbol{\alpha}^{(t)}, \boldsymbol{\Sigma}_i^{(t)})\, d\mathbf{y}_i}{\int P(M_i = 1|\mathbf{y}_i)\phi(\mathbf{y}_i, \mathbf{X}_i\boldsymbol{\alpha}^{(t)}, \boldsymbol{\Sigma}_i^{(t)})\, d\mathbf{y}_i}, \\
\mathrm{var}\big(\mathbf{y}_i|M_i = 1, \boldsymbol{\Omega}^{(t)}\big) &= \mathrm{E}\big(\mathbf{y}_i\mathbf{y}_i^T|M_i = 1, \boldsymbol{\Omega}^{(t)}\big) - \mathbf{y}_i^{(t)}\mathbf{y}_i^{(t)^T}, \\
\mathbf{b}_i^{(t)} &= \mathrm{E}\big(\mathbf{b}_i|\mathbf{y}_i, M_i = 1, \boldsymbol{\Omega}^{(t)}\big) = \mathbf{D}^{(t)}\mathbf{Z}_i^T\mathbf{W}_i^{(t)}\big(\mathbf{y}_i^{(t)} - \mathbf{X}_i\boldsymbol{\alpha}^{(t)}\big), \\
\mathrm{var}\big(\mathbf{b}_i|\mathbf{y}_i, M_i = 1, \boldsymbol{\Omega}^{(t)}\big) &= \mathbf{D}^{(t)} - \mathbf{D}^{(t)}\mathbf{Z}_i^T\mathbf{W}_i^{(t)}\mathbf{Z}_i\mathbf{D}^{(t)} \\
&\quad + \mathbf{D}^{(t)}\mathbf{Z}_i^T\mathbf{W}_i^{(t)}\,\mathrm{var}\big(\mathbf{y}_i|M_i = 1, \boldsymbol{\Omega}^{(t)}\big)\mathbf{W}_i^{(t)}\mathbf{Z}_i\mathbf{D}^{(t)}, \quad \text{and} \\
\mathbf{V}_i^{(t)} &= \mathrm{var}\big(\mathbf{e}_i|\mathbf{y}_i, M_i = 1, \boldsymbol{\Omega}^{(t)}\big) \\
&= \mathbf{Z}_i\mathbf{D}^{(t)}\mathbf{Z}_i^T - \mathbf{Z}_i\mathbf{D}^{(t)}\mathbf{Z}_i^T\mathbf{W}_i^{(t)}\mathbf{Z}_i\mathbf{D}^{(t)}\mathbf{Z}_i^T \\
&\quad + \mathbf{R}_i\mathbf{W}_i^{(t)}\,\mathrm{var}\big(\mathbf{y}_i|M_i = 1, \boldsymbol{\Omega}^{(t)}\big)\mathbf{W}_i^{(t)}\mathbf{R}_i.
\end{aligned}$$

4. Simulations.

4.1. *Comparison of modeling absolute abundance via* `mixEMM` *versus modeling relative abundance.* To remove batch effects in iTRAQ/TMT-based proteomics analyses, a standard practice is to analyze the relative abundance of a protein/peptide in the target samples relative to the abundance level of the protein/peptide in the reference sample from the same batch, and assess the association of relative abundance of each protein/peptide with the phenotype. In this simulation section, we will show that directly modeling absolute abundance with `mixEMM` is better than the conventional analysis based on relative abundances.

4.1.1. *Simulation* 1. *Generate data based on multivariate normal distributions.* We simulated 1000 multivariate normal data sets $\mathbf{y}_i \sim N(\mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i, \mathbf{R})$ with $p = 4$ for a batch size of $Q = 40$ and $Q = 200$. In each batch, we assume that

the first sample is the reference sample and the rest are target samples. The fixed effects are $\boldsymbol{\alpha}$. Here $\mathbf{X}_i$ is a $p \times (k+1)$ $(k=2)$ covariate matrix for each observation i with the first column being $\mathbf{1}$ and $\boldsymbol{\alpha} = (10, -a, a)^T$. In assessing the type I error rate, we set $a = 0$. In evaluating the power, we set $a = 0.7$ when $Q = 40$ and $a = 0.3$ when $Q = 200$. Here we included only a random intercept for each batch. The random effect is $b_i \sim N(0, D)$, and $\mathbf{Z}_i$ is a vector of 1's. The matrix $\mathbf{R}$ is a diagonal matrix with diagonal elements $(\sigma_0^2, \sigma^2, \sigma^2, \sigma^2)$. Note that σ_0^2 represents the variance of the reference sample, and it is purely due to experimental variation across the different iTRAQ/TMT multiplex. While σ^2 represents the variance of the target samples, it is a combination of both biological and experimental variation. Thus, the reference sample variance is often smaller than the variance of other tumor samples. We simulated three settings: (I) the sample variation and experimental variation are large, $\sigma_0^2 = 2, \sigma^2 = 4, D = 3$; (II) the sample variation and experimental variation are small, $\sigma_0^2 = 1, \sigma^2 = 2, D = 1$; and (III) the experimental variation for the reference sample is extremely small, $\sigma_0^2 = 0.01, \sigma^2 = 4, D = 3$. Note that simulation setting III is the setting ideal for linear regression based on relative abundance. We generated approximately 40% missing data at the batch level by the mechanism in (2) with $\gamma_0 = 0$ and $\gamma = 0.1$. We also generated an additional 5% sporadic (random) missingness.

When applying the `mixEMM` method, based on the estimated MLEs for the fixed effects and their variance estimates in (14), we first obtained the Wald test statistics for testing $H_0 : \boldsymbol{\alpha}_{-1} = \mathbf{0}$, where $\boldsymbol{\alpha}_{-1}$ stands for the fixed effects other than the mean (i.e., the intercept). We then derived the p-values by approximating the null distribution through permuting the order of batches of response variables. We compared the proposed `mixEMM` method incorporating BADMM with $\gamma = 0.1$ versus the `mixEMM` model with $\gamma = 0$. Note that, when $\gamma = 0$, the missing mechanism is treated as missing at random (or missing completely at random) and ignorable.

We also compared the performance of `mixEMM` with that of the conventional analysis based on relative abundances. Specifically, we assumed the simulated $\mathbf{y}_i$ representing log abundances and calculated the relative abundance measures as $y_{ij} - y_{i1}$ for $j = 2, 3, 4$. We then treated relative abundances as responses and fitted linear regressions to detect significant associations (regression coefficients). Again, p-values were derived through permutation tests in the same way as we did for `mixEMM`.

Table 1(a) shows that with permutation-based p-values, all three methods can control type I error rates in different scenarios at the p-value threshold of 0.05. Compared with the conventional approach of analyzing relative abundances, both versions of `mixEMM` enjoyed much improved power [Table 1(b)]. In particular, when experimental variation is large, the improvement of power can be up to 3-fold. Even when the variation of the reference samples is extremely small and the batch effects have limited impact on relative abundances, the power of `mixEMM` is still more than twice that of linear regression analysis based on relative abundances. We repeated the simulation at the p-value threshold of 0.01 and reached

TABLE 1

Simulation 1 *results. We compare* (a) *the type I error rates and* (b) *the power of the* `mixEMM` *method with and without considering BADMM, as well as linear regressions using relative abundances as responses*

# batch (experiment)	Simulation setting	Methods		
		The proposed `mixEMM` with $\gamma = 0.1$ in BADMM	**The `mixEMM` with $\gamma = 0$**	**Linear regression on relative abundance**
	(a) *Type* I *error rates at the p-value threshold of 0.05*			
40	Large sample/experimental variation ($\sigma_0^2 = 2, \sigma^2 = 4, D = 3$)	0.055	0.056	0.044
	Small sample/experimental variation ($\sigma_0^2 = 1, \sigma^2 = 2, D = 1$)	0.053	0.053	0.040
	Minimum experimental variation ($\sigma_0^2 = 0.01, \sigma^2 = 4, D = 3$)	0.042	0.043	0.054
200	Large sample/experimental variation ($\sigma_0^2 = 2, \sigma^2 = 4, D = 3$)	0.045	0.045	0.061
	Small sample/experimental variation ($\sigma_0^2 = 1, \sigma^2 = 2, D = 1$)	0.071	0.070	0.052
	Minimum experimental variation ($\sigma_0^2 = 0.01, \sigma^2 = 4, D = 3$)	0.054	0.046	0.053
	(b) *Power at the p-value threshold of 0.05*			
40	Large sample/experimental variation ($\sigma_0^2 = 2, \sigma^2 = 4, D = 3$)	0.437	0.442	0.176
	Small sample/experimental variation ($\sigma_0^2 = 1, \sigma^2 = 2, D = 1$)	0.780	0.773	0.293
	Minimum experimental variation ($\sigma_0^2 = 0.01, \sigma^2 = 4, D = 3$)	0.663	0.657	0.230
200	Large sample/experimental variation ($\sigma_0^2 = 2, \sigma^2 = 4, D = 3$)	0.491	0.472	0.190
	Small sample/experimental variation ($\sigma_0^2 = 1, \sigma^2 = 2, D = 1$)	0.838	0.842	0.349
	Minimum experimental variation ($\sigma_0^2 = 0.01, \sigma^2 = 4, D = 3$)	0.696	0.695	0.244

the same conclusion (results not shown). These results clearly demonstrated the advantage of modeling the batch design through a mixed-effects model, which helps to characterize the variance structure in the data more precisely. The two versions of `mixEMM` ($\gamma = 0.1$ v.s. $\gamma = 0$) enjoy similar power in all settings. This

suggests that BADMM has only limited impacts on the testing results. However, in the next section, we will demonstrate that incorporating BADMM will improve parameter estimation.

4.1.2. *Simulation* 2. *Generate data based on real iTRAQ outputs*. In this simulation, we generated data based on the phosphoproteomics profiles from the CPTAC breast cancer study (please see Section 5 for details). Specifically, we started with the abundance measurements of 3182 phosphosites fully observed in all 36 batches. Each batch has three tumor samples and one reference sample. We randomly selected 1000 out of the 3182 phosphosites, and then randomly split all 108 tumor samples into two groups of equal sizes. After that, we randomly picked 300 out of the 1000 phosphosites. For each of these 300 phosphosites, we added $\delta = s\sigma_i$ to the abundances of the samples in the first group, where s was set to 0.5 or 0.7 to reflect different signal sizes, and σ_i was the standard deviation (SD) of the abundance for the ith phosphosite. We then generated 40% batch-level missing values using model (2) with $\gamma_0 = 0$ and $\gamma = 0.043$. We repeated this process and obtained 10 replicate data sets.

We applied two versions of `mixEMM` with the missing-data parameter $\gamma = 0$ or γ estimated from the data. For comparison, a linear regression based on the relative abundance was also performed. For all methods, we derived permutation based p-values. In Figure 2, we reported the power and false discovery rate (FDR) when the

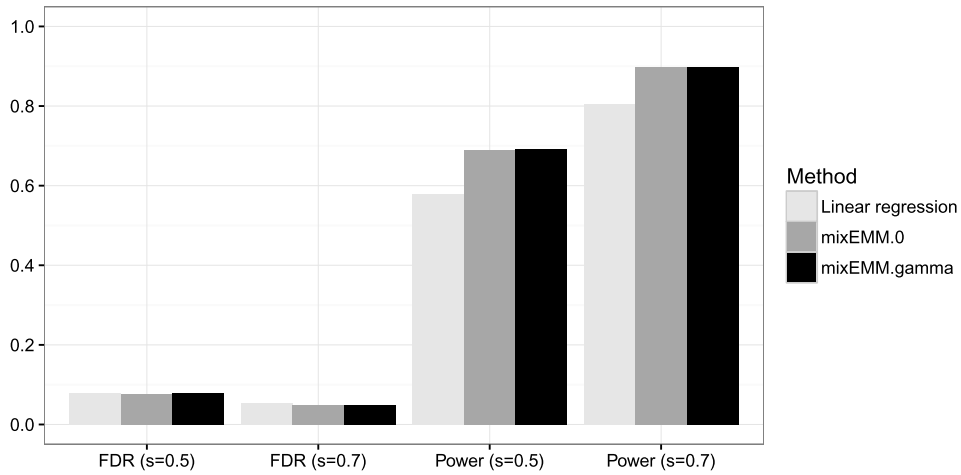


FIG. 2. *Simulation* 2 *results. Based on simulation data generated from real iTRAQ outputs, we compare the performance of the* `mixEMM` *methods and the linear regression which used relative abundances as responses. The colors of the bars represent different methods. The heights of the bars represent either FDRs or powers of various methods. Bars corresponding to different signal levels were labeled with "*$s = 0.5$*" and "*$s = 0.7$*" in the x-axis, respectively. For* `mixEMM.0`, $\gamma = 0$ *was used. For* `mixEMM.gamma`, $\hat{\gamma}$ *based on observed data was used.*

targeted FDR was set at 0.05. For different signal sizes, both versions of `mixEMM` are more powerful than the linear regression based on relative abundances.

4.2. *The BADMM modeling in* `mixEMM`. We simulated 1000 multivariate normal data sets similar to before with $\boldsymbol{\alpha} = (10, -1, 1)^T$, $\sigma_0^2 = 2$, $\sigma^2 = 4$. We generated approximately 40% missing data at the batch level by the mechanism in (2) with $\gamma_0 = 0$ and $\gamma = 0.1$, and an additional 5% sporadic (random) missingness.

Table 2 shows the relative Mean Squared Errors (MSEs) of `mixEMM` incorporating BADMM ($\gamma = 0.1$) versus `mixEMM` without considering BADMM ($\gamma = 0$) on estimates for the fixed effects and variance with different sample sizes. The relative MSEs for the fixed effects estimates are approximately 0.8 for $Q = 40$ and 0.5 for $Q = 200$. This suggests that, by taking into account the missing batches, the proposed `mixEMM` method provides more accurate estimates for fixed effects in both the limited and large sample scenarios. The relative MSEs for variance estimates are very close to 1, indicating that modeling the nonignorable missingness mainly helps to correct the biases in the fixed effects estimates rather than the variance estimates.

In addition to the simulations above, we also reanalyzed the simulated data using the logit missing-data mechanism function in (16) and compared the relative MSEs. Note that the simulated data were generated from the exponential BADMM in (2) and that we used the logit function to analyze the data with $\gamma_0 = 0$ and $\gamma = 0.1$; that is, the missing-data mechanism is potentially misspecified. The relative MSEs based on the logit function are close to those based on the true BADMM with only a minor loss of efficiency. Since the logit function is quite flexible and fits the observed missing-data pattern well, the overall biases of the fixed effects estimates are quite small. This suggests that the logit BADMM function is a general and flexible missing-data mechanism function. When data are generated by

TABLE 2

The comparison of relative MSEs and computation time for estimates of fixed effects and variance components obtained from incorporating BADMM ($\gamma = 0.1$) relative to those assuming missing at random ($\gamma = 0$) in `mixEMM`. *The missing data are generated according to the exponential BADMM in (2). We compare the relative MSEs when the true missing-data mechanism is accounted for in the estimation of the* `mixEMM` *algorithm and when the logit BADMM is used in the estimation with estimated missing-data mechanism parameters. The results are based on* 1000 *repeated simulations*

Methods	# experiment Q	α	σ_0^2	σ^2	D	Computation Time (in hours)
`mixEMM` with	40	0.848	1.014	1.006	1.184	0.287
exponential BADMM	200	0.492	1.016	1.006	1.015	1.514
`mixEMM` with	40	0.851	1.004	1.002	1.047	4.460
logit BADMM	200	0.538	1.007	1.004	0.983	24.243

logit BADMM and reanalyzed by exponential BADMM, as long as the exponential pattern nicely fits the observed missing-data pattern, the conclusions are similar (results not shown).

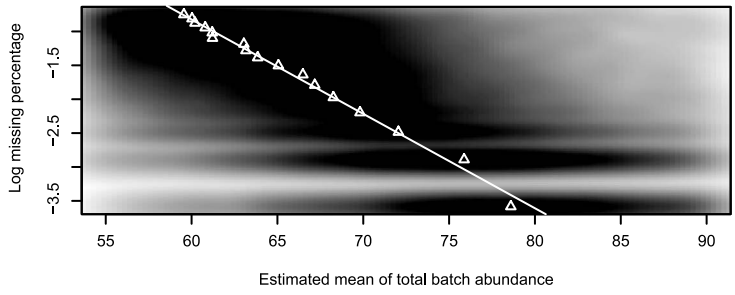
When the two BADMM mechanisms produce similar patterns, the exponential function is about 15 times faster than the logit function. Specifically, in terms of computation time, it takes 0.287 and 1.514 hours for a single node computer to analyze 1000 features based on the exponential BADMM when sample sizes are 40 and 200, respectively, whereas it takes 4.460 and 24.243 hours for the analysis based on the logit BADMM. The computation time increases rapidly with dimensionality p and sample size n. When jointly analyzing multiple features, for example, multiple peptides of the same protein, the superiority of the exponential BADMM becomes more substantial. On the other hand, the logit BADMM would be useful when the exponential pattern does not fit well.

The fit of the selected and estimated BADMM pattern should often be checked before using the `mixEMM` method in the estimation and inference. For example, in our real data application, we evaluate the fit of the exponential BADMM in Figure 3 before the subsequent analysis.

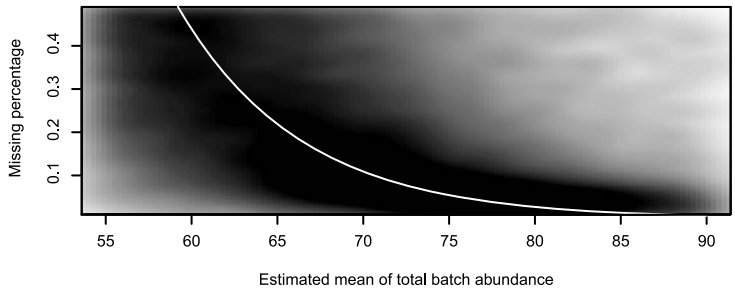
4.3. *Evaluating available-case based missing-data mechanism parameter estimates*. When applying the `mixEMM`, in the simulations above, we either used the true missing-data mechanism with true parameters or we use a misspecified mechanism with misspecified parameters. In this subsection, we evaluated the estimation of the missing-data mechanism parameter. Specifically, we pooled all features together and obtained the available case sum of abundance estimates and the proportion of missing batch for each feature. We estimated the missing-data mechanism parameter based on (15) for the exponential BADMM in (2) given the data.

We simulated 1000 features with means randomly sampled from $N(10, 2^2)$, other parameters similar to those in previous sections and the number of batches $Q = 40$ and $Q = 200$. We generated batch-level missingness by (2) with $\gamma_0 = 0$ and $\gamma = 0.1$, and calculated $\hat{\gamma}_0$ and $\hat{\gamma}$ based on the 1000 features. We repeated the procedure 100 times. Table 3 lists the distribution of $\hat{\gamma}_0$ and $\hat{\gamma}$. The estimates for γ are reasonably accurate, while $\hat{\gamma}_0$ could have substantial biases. However, since γ_0 does not affect the E- or the CM-step, the overall performance of available-case based missing-data mechanism parameters is almost identical to that of using true parameters (data not shown).

5. Application to the CPTAC proteomics data to identify proteins related to triple negative breast cancer tumors. Triple negative breast cancer (TNBC) refers to breast cancer that does not express the genes for the estrogen receptor, progesterone receptor or Her2/neu. TNBC patients have a much higher risk of relapse in the first 3–5 years compared to other types of breast cancer patients. It is also more difficult to treat TNBC since most chemotherapies target one of the three receptors. More effective treatment strategies for TNBC patients are highly



(a)



(b)

FIG. 3. *An illustration of BADMM based on CPTAC breast cancer phosphoproteomics data*. (a) *A smoothed density representation of the scatter plot of the log percentage of missing batches for each phosphopeptide* (*y-axis*) *versus its estimated mean of total batch abundances based on the observed data* (*x-axis*). *Note that, for one phosphospeptide and one batch, the total batch abundance is defined as the sum of abundance measurements of this phosphopeptide in all samples of this batch* [*i.e.*, $\mathbf{1}^T\mathbf{y}_i$ *in equation* (2)]. *The mean value is then estimated using the average across all batches. This plot is generated using the R function* `smoothScatter`. *The darker the shade is, the higher the density is. The triangular points indicate medians of estimated means of total batch abundance of all phosphopeptides with the same missing percentage. The white line represents the linear regression fit of the triangular points*. (b) *A similar plot as* (a) *except that the y-axis is on the original scale. The curve corresponds to the line in* (a).

TABLE 3
The distribution of available case-based estimated missing-data mechanism parameters based on 100 *repeated simulations*

# batch	Parameter	True value	min	Median	Mean	Max
40	γ	0.1	0.093	0.101	0.101	0.107
	γ_0	0	−0.119	−0.059	−0.055	0.029
200	γ	0.1	0.097	0.104	0.104	0.108
	γ_0	0	−0.134	−0.094	−0.093	−0.014

desirable. In this section, we applied the proposed `mixEMM` algorithm to the motivating proteomics data set from the CPTAC breast cancer project [Mertins et al. (2016)] to identify phosphopeptides up- or down-regulated in TNBC tumors compared to other types of breast cancer tumors. Such information would shed light on the disease mechanism of TNBC, which may then lead to better clinical practice for TNBC diagnosis and treatment.

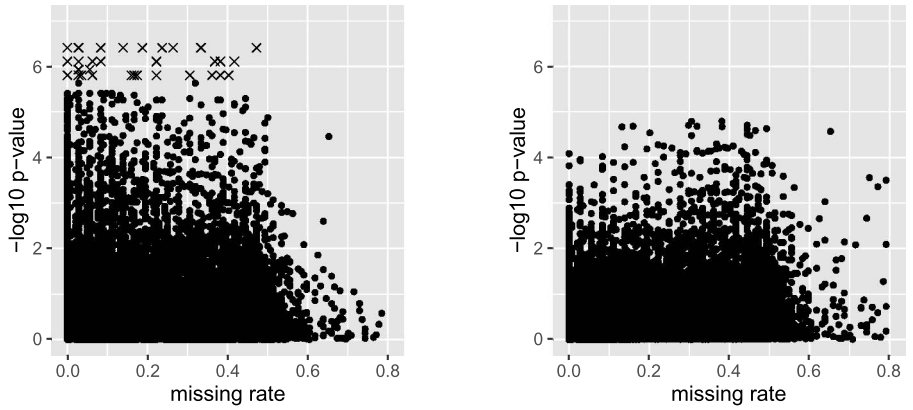
In the CPTAC breast cancer project, we analyzed 108 tumor samples from 105 breast cancer patients, with 3 patients having two tumor samples. Protein profiles were obtained through 36 four-plex iTRAQ experiments generated at Dr. Carr's lab at the Broad Institute of MIT and Harvard in Boston, U.S. Each iTRAQ experiment processed 3 breast tumor samples and the reference sample, which was created by combining 40 of these tumors. The iTRAQ-labeled peptides were fractionated and chemically enriched for phosphopeptides. The resulting samples were processed using high-resolution MS instruments (LS-MS/MS on Thermo Q-Exactive). Phosphopeptide identification and quantification were performed using Spectrum Mill software (Agilent Technologies, Santa Clara, CA). Specifically, for phosphopeptide quantification, the charge state with the best peptide-spectrum match score across all fractions and samples of one phosphosite was used as the representative state for that phosphosite to derive the corresponding quantification measurements. This strategy helps to reduce the impact of false phosphopeptide identifications on the quantification results.

The phosphoproteomics data were downloaded from the CPTAC Data Coordinating Center (http://proteomics.cancer.gov/programs/cptacnetwork) sponsored by the National Cancer Institute. In total, 63,698 phosphopeptides were identified and quantified in at least one sample. However, only 3182 (5.0%) phosphopeptides had complete measurements in all the samples. The missing rates of each sample ranged from 46.78% to 79.56%. The substantial amount of missing values in phosphoproteomics data raises a pressing need for statistical methods properly incorporating nonignorable missingness. Among all missing observations, 97.3% were batch-level missingness, that is, a phosphopeptide was missing in all four samples of an iTRAQ experiment. Thus, the BADMM pattern suits these data sets well.

Since the distribution of the raw intensity measurements has a very heavy right tail, we performed the analysis based on log transformed abundances. Note that all log transformed data still take positive values. We normalized each sample to have the same median and median absolute deviation. We then filtered out the low-quality observations and focused on the 25,961 phosphopeptides that were observed in at least 25 (70%) of the 36 runs of the reference sample. The missing rates of each sample for these 25,961 phosphopeptides ranged from 8.40% to 56.78% with a mean value of 19.60%. Figure 3 illustrates the relationships between the missing percentage and estimated mean of total batch abundance of each phosphopeptide [i.e., $\mathbf{1}^T\mathbf{y}_i$ in equation (2)]. The exponential probabilistic model in equation (2) accurately reflects the BADMM pattern in the data.

Additionally, we fitted a standard linear mixed-effects model with random intercepts for the batches to the observed abundances of each phosphopeptide. The intra-class correlations, which quantify how strongly abundance measures for the same peptides resemble each other, for all phosphopeptides have a median of 0.788 and a mean of 0.733. Over 93.3% of the phosphopeptides have significant random intercepts by ANOVA tests at the Bonferroni adjusted p-value threshold of 0.05. Those suggest that the batch effects in the current data are quite strong. Our proposed `mixEMM` model considering BADMM well suits the data.

We applied the proposed `mixEMM` method to identify the phosphopeptides up- or down-regulated in TNBC tumors relative to other breast cancer tumors. In the mixed-effects model, we included a random effect, for each iTRAQ multiplex experiment, and three fixed effects: an intercept, an indicator for the reference sample and an indicator for triple negative subtype. We also conducted the analysis using linear regression models based on relative abundances for comparison. The resulting p-values of all 25,961 phosphopeptides from both methods based on permutation are illustrated in Figure 4. At the same Bonferroni adjusted p-value threshold of 0.05, the `mixEMM` algorithm considering BADMM identified 44 phosphosites corresponding to 29 unique genes as being significantly up- or down-regulated in TNBC relative to other types of breast cancer tumors. Only 3 of these 44 phosphosites have complete observations in all 108 samples. Nine and three of the 44 phosphosites have a missing rate greater than 30% and 40%, respectively. In contrast, the conventional analysis based on relative abundances failed to detect any



(a) mixEMM (b) Linear regression

FIG. 4. *The relationship between p-values and missing rates*. (a) *The results from* `mixEMM`*-based analysis using absolute abundances and considering BADMM*; (b) *the results from linear regression analysis using relative abundances. In both* (a) *and* (b), *the X-axis represents the missing rates of phosphosites and the Y-axis represents the negative* $\log_{10}$ *of p-values. Phosphopeptides are indicated as "X"s if their p-values are below the Bonferroni corrected p-value cutoff* (0.05/25961). *Otherwise, they are plotted as dots.*

significant phosphosite at the same significance threshold. Given the p-values of both mixEMM and the conventional analysis were calculated based on permutation and simulation results had shown that both approaches can control the type I error rates, we concluded that the mixEMM method enjoys improved power over conventional methods at the same p-value threshold. These results are consistent with what we observed in the simulations.

The phosphosite with the most significant p-value corresponds to the gene *FOXA1*, a transcription factor. The gene *FOXA1* is known to be associated with breast cancer risk [Meyer and Carroll (2012)]. A more recent work further suggests that *FOXA1* silencing increases migration and invasion of breast cancer cells [Bernardo et al. (2013)]. This is consistent with our finding that phosphoprotein of *FOXA1* was significantly down-regulated in TNBC tumors, and TNBC tumors are usually more aggressive than other subtypes of breast cancer. Moreover, according to the STRING data base [Szklarczyk et al. (2014)], *FOXA1* interacts with another gene, *SOX10*, in the significant 29 gene list. The gene *SOX10* is a neural crest transcription factor. Based on a recent immunohistochemistry study [Cimino-Mathews et al. (2013)], *SOX10* has been reported to be preferentially expressed in TNBC and was also validated as a sensitive diagnostic marker for basal-like TNBC [Ivanov et al. (2013)]. The proposed mixEMM method detects these known TNBC genes, which strengthens our confidence that the mixEMM method will help to reveal biological relevant information underlying iTRAQ data. Further investigation on how *FOXA1*, *SOX10* and the other 27 significant genes function may help us better understand the disease mechanism and improve the development of novel diagnostic and therapeutic tools for TNBC.

6. Discussion. In this paper, we propose a new method—mixEMM—for analyzing data from iTRAQ/TMT proteomics experiments. The proposed mixEMM method employs a mixed-effects model to characterize the variance structure for the abundance measurements from iTRAQ/TMT experiments. It uses an exponential probability function to model the batch-level nonignorable missing-data mechanism (BADMM) in the iTRAQ/TMT data. The goal of our analyses is to estimate the fixed effects for the associations between proteomic features and sample phenotypes (e.g., clinical outcomes). To achieve this goal, we implement an ECM algorithm to calculate the MLEs of the parameters of interest. The superior performance of the mixEMM method over the conventional approach is illustrated using both simulations and a real data example.

In practice, the experimental variation across different iTRAQ/TMT experiments is often large. In other words, even for the same reference sample, the protein/peptide abundance measurements in different batches measured by different iTRAQ/TMT experiments may differ substantially. The conventional approach directly analyzes relative abundance measures, which in some sense mixes up the variation in the target samples and the reference samples, and consequently causes a loss of efficiency and power. In contrast, mixEMM precisely characterizes the

experimental properties, accounts for the variations of samples across batches, and gains substantial power improvement in the subsequent tests.

While explicitly modeling BADMM has a limited impact on testing, it improves parameter estimations for fixed effects. In addition to the exponential probability function for BADMM, we have also investigated the use of the logit function for modeling the missing-data mechanism. When both functions fit the observed missing-data pattern well, the estimation accuracies of the two functions are comparable, and the computationally efficient exponential function is recommended. The logit BADMM function is more flexible and can be used in the analyses of log-ratio data or data with more complex missing-data patterns. Other flexible missing-data mechanism functions, such as spline functions, can be incorporated into the proposed framework, although numerical integration would be required.

In iTRAQ/TMT experiments, different proteins/peptides with different physical and chemical properties may be subjected to different experimental variations. Thus, it is reasonable to assume batch effects to be feature-specific as we did in `mixEMM`. Nevertheless, experimental factors, such as variations in sample loadings, might affect the whole sample measurements. Therefore, in practice, we suggest performing global normalization using all data to take care of major experimental shifts before fitting the proposed `mixEMM` model.

This work was motivated by phosphoproteomics data in which the natural analysis unit is each individual phosphopeptide and each phosphopeptide is directly quantified in the experiments. However, the proposed method can also be applied to any type of proteomics assay that uses iTRAQ, including global proteomics and glycol-proteomics. Note that, for global proteomics data, the quantification is obtained at the peptide level, while the target analysis unit is each individual protein. To perform inference at the protein level, one strategy is to apply the proposed `mixEMM` algorithm at the peptide-level data and then summarize the results of peptides within each protein. Another strategy is to first calculate protein abundances based on the mean or median of peptide abundances within each protein and then apply the proposed `mixEMM` method to the summary protein abundances. A more sophisticated treatment would be to perform a multivariate analysis and jointly model multiple peptides of the same protein. Clough et al. (2009) compared the two strategies based on proteomics data from label-free liquid chromatography-MS experiments. To our knowledge, such comparisons have not been done on labeled experiments with iTRAQ/TMT data. Research along this direction is ongoing.

In the CPTAC study, a common reference sample was included in each iTRAQ experiment. While the common reference design is preferable for clustering analysis, it may not be the most efficient design for differential expression analysis [Dobbin and Simon (2002)]. On the other hand, the iTRAQ protocol enables the use of randomized block design, which may lead to better control of experimental variance. The proposed `mixEMM` can be easily modified to handle data from randomized block designs.

BADMM is a unique property for data from iTRAQ/TMT experiments. For data from unlabeled proteomic experiments, however, missing events in different samples are independent. In these cases, a sample-level-abundance-dependent-missing probability model would be more suitable, and `mixEMM` can be easily modified to handle such structures. In addition, `mixEMM` can be extended to incorporate the same prior distribution on the variance parameters of batch effects for all features. Such a hierarchical model can help to stabilize the batch effect estimation, especially when the sample size is limited.

In our model, we assume batch effects and errors follow Gaussian distributions. While we cannot guarantee that the assumption of the Gaussian distribution holds in real data sets, results of simulation 2 (Section 4.1.2) suggest that the performance of `mixEMM` is robust to the violation of the Gaussian assumption.

The proposed framework is not limited to proteomics data and is generally applicable to data with repeated/clustered measures and cluster-level incomplete data. An R package `mixEMM` will be available through CRAN.

Acknowledgment. Mass spectrometry and proteomics data were acquired by the breast cancer project of the CPTAC consortium, which is led by Dr. Steve Carr from the Broad Institute of MIT and Harvard, and Dr. Amenda Paulovich from Fred Hucthinson Cancer Research Center, and is supported by NCI grant CA160034. We thank Drs. Chenwei Lin, D.R. Mani, Philipp Mertins, Yan Ping and others from the CPTAC consortium for their help on the proteomics data. We also thank Dr. Ross Prentice for helpful suggestions and comments. Moreover, we thank the Editor, Associate Editor and anonymous referees for their valuable suggestions that helped to improve the manuscript greatly.

REFERENCES

BERNARDO, G. M., BEBEK, G., GINTHER, C. L., SIZEMORE, S. T., LOZADA, K. L., MIEDLER, J. D., ANDERSON, L. A., GODWIN, A. K., ABDUL-KARIM, F. W., SLAMON, D. J. and KERI, R. A. (2013). FOXA1 represses the molecular phenotype of basal breast cancer cells. *Oncogene* **32** 554–563.

CHANG, C.-Y., PICOTTI, P., HÜTTENHAIN, R., HEINZELMANN-SCHWARZ, V., JOVANOVIC, M., AEBERSOLD, R. and VITEK, O. (2012). Protein significance analysis in selected reaction monitoring (SRM) measurements. *Mol. Cell. Proteomics* **11** M111.014662.

CHEN, L. S., PRENTICE, R. L. and WANG, P. (2014). A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation. *Biometrics* **70** 312–322. MR3258036

CIMINO-MATHEWS, A., SUBHAWONG, A. P., ELWOOD, H., WARZECHA, H. N., SHARMA, R., PARK, B. H., TAUBE, J. M., ILLEI, P. B. and ARGANI, P. (2013). Neural crest transcription factor Sox10 is preferentially expressed in triple-negative and metaplastic breast carcinomas. *Human Pathol.* **44** 959–965.

CLOUGH, T., KEY, M., OTT, I., RAGG, S., SCHADOW, G. and VITEK, O. (2009). Protein quantification in label-free LC-MS experiments. *J. Proteome Res.* **8** 5275–5284.

DOBBIN, K. and SIMON, R. (2002). Comparison of microarray designs for class comparison and class discovery. *Bioinformatics* **18** 1438–1445.

ELLIS, M., GILLETTE, M., CARR, S., PAULOVICH, A., SMITH, R., RODLAND, K., TOWNSEND, R., KINSINGER, C., MESRI, M., RODRIGUEZ, H., LIEBLER, D. and CPTAC (2013). Connecting genomic alterations to cancer biology with proteomics: The NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discovery* **3** 1108–1112.

FRANKEN, H., MATHIESON, T., CHILDS, D., SWEETMAN, G. M. A., WERNER, T., TÖGEL, I., DOCE, C., GADE, S., BANTSCHEFF, M., DREWES, G., REINHARD, F. B. M., HUBER, W. and SAVITSKI, M. M. (2015). Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry. *Nat. Protoc.* **10** 1567–1593.

HILL, E. G., SCHWACKE, J. H., COMTE-WALTERS, S., SLATE, E. H., OBERG, A. L., ECKEL-PASSOW, J. E., THERNEAU, T. M. and SCHEY, K. L. (2008). A statistical model for iTRAQ data analysis. *J. Proteome Res.* **7** 3091–3101.

IBRAHIM, J. G. and MOLENBERGHS, G. (2009). Missing data methods in longitudinal studies: A review. *TEST* **18** 1–43. MR2495958

IVANOV, S. V., PANACCIONE, A., NONAKA, D., PRASAD, M. L., BOYD, K. L., BROWN, B., GUO, Y., SEWELL, A. and YARBROUGH, W. G. (2013). Diagnostic SOX10 gene signatures in salivary adenoid cystic and breast basal-like carcinomas. *Br. J. Cancer* **109** 444–451.

KARP, N. A., HUBER, W., SADOWSKI, P. G., CHARLES, P. D., HESTER, S. V. and LILLEY, K. S. (2010). Addressing accuracy and precision issues in iTRAQ quantitation. *Mol. Cell. Proteomics* **9** 1885–1897.

LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974.

LIEBLER, D., ZHANG, B., WANG, J., WANG, X., ZHU, J., LIU, Q., SHI, Z., CHAMBERS, M. C. et al. (2014). Proteogenomic characterization of human colon and rectal cancer. *Nature* **513** 382–387.

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, Hoboken, NJ. MR1925014

LUO, R., COLANGELO, C. M., SESSA, W. C. and ZHAO, H. (2009). Bayesian analysis of iTRAQ data with nonrandom missingness: Identification of differentially expressed proteins. *Statistics in Biosciences* **1** 228–245.

MCALISTER, G. C., NUSINOW, D. P., JEDRYCHOWSKI, M. P., WÜHR, M., HUTTLIN, E. L., ERICKSON, B. K., RAD, R., HAAS, W. and GYGI, S. P. (2014). MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Analytical Chemistry* **86** 7150–7158.

MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. MR1243503

MERTINS, P., MANI, D. R., RUGGLES, K. V., GILLETTE, M. A., CLAUSER, K. R., WANG, P. et al. (2016). Proteogenomic connects somatic mutations to signaling in breast cancer. *Nature* **534** 55–62.

MEYER, K. B. and CARROLL, J. S. (2012). FOXA1 and breast cancer risk. *Nat. Genet.* **44** 1176–1177.

THE CANCER GENOME ATLAS NETWORK (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490** 61–70.

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, H. (2008). Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. *J. Proteome Res.* **7** 225–233.

PAULO, J. A., MCALLISTER, F. E., EVERLEY, R. A., BEAUSOLEIL, S. A., BANKS, A. S. and GYGI, S. P. (2014). Effects of MEK inhibitors GSK1120212 and PD0325901 in vivo using 10-plex quantitative proteomics and phosphoproteomics. *Proteomics* **15** 462–473.

PAULOVICH, A. G., BILLHEIMER, D., HAM, A. J., VEGA-MONTOTO, L., RUDNICK, P. A., TABB, D. L., WANG, P., BLACKMAN, R. K., BUNK, D. M. and CARDASIS, H. ET AL. (2010). Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol. Cell. Proteomics* **9** 242–254.

PINHEIRO, J. C. and BATES, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J. Comput. Graph. Statist.* **4** 12–35.

RAUNIYAR, N. and YATES III, J. R. (2014). Isobaric labeling-based relative quantification in shotgun proteomics. *J. Proteome Res.* **13** 5293–5309.

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.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196

SZKLARCZYK, D., FRANCESCHINI, A., WYDER, S., FORSLUND, K., HELLER, D., HUERTA-CEPAS, J., SIMONOVIC, M., ROTH, A., SANTOS, A., TSAFOU, K. P. et al. (2014). STRING v10: Protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* gku1003.

WANG, P., TANG, H., ZHANG, H., WHITEAKER, J., PAULOVICH, A. G. and MCINTOSH, M. (2006). Normalization regarding non-random missing values in high-throughput mass spectrometry data. *Pacific Symposium on Biocomputing* 315–326.

WERNER, T., SWEETMAN, G., SAVITSKI, M. F., MATHIESON, T., BANTSCHEFF, M. and SAVITSKI, M. M. (2014). Ion coalescence of neutron encoded TMT 10-plex reporter ions. *Anal. Chem.* **86** 3594–3601.

WIESE, S., REIDEGELD, K. A., MEYER, H. E. and WARSCHEID, B. (2007). Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. *Proteomics* **7** 340–350.

ZHANG, H., LIU, T., ZHANG, Z., PAYNE, S. H., ZHANG, B. and MCDERMOTT, J. E. et al. (2016). Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell* **166** 755–765.

L. S. CHEN
J. WANG
DEPARTMENT OF PUBLIC HEALTH SCIENCES
UNIVERSITY OF CHICAGO
5841 S MARYLAND AVE
CHICAGO, ILLINOIS
USA
E-MAIL: lchen@health.bsd.uchicago.edu
jwang88@uchicago.edu

X. WANG
DIVISION OF PUBLIC HEALTH SCIENCES
FRED HUTCHINSON CANCER RESEARCH CENTER
1100 FAIRVIEW AVE N
SEATTLE, WASHINGTON 98109
USA
E-MAIL: xwan2@fhcrc.org

P. WANG
ICAHN INSTITUTE OF GENOMICS AND MULTISCALE BIOLOGY
ICAHN SCHOOL OF MEDICINE AT MOUNT SINAI
1470 MADISON AVE, S8-102
NEW YORK, NEW YORK
USA
E-MAIL: pei.wang@mssm.edu