Submitted to the Annals of Applied Statistics

SUPPLEMENTARY TEXT FOR "THE EFFECTS OF NON-IGNORABLE MISSING DATA ON LABEL-FREE MASS SPECTROMETRY PROTEOMICS EXPERIMENTS"*

By Jonathon J. O'Brien[‡] Harsha P. Gunawardena[†] Joao A. Paulo[‡] Xian Chen[†] Joseph G. Ibrahim[†] Steven P. Gygi[‡] and Bahjat F. Qaqish[†]

University of North Carolina at Chapel Hill[†] *and Harvard Medical School*[‡]

1. Supplementary Text A: Proofs.

1.1. *Deriving the full conditional distribution for a missing value.* If Y_m is a missing intensity and $X\theta_{[m]} = E(Y_m)$. Then,

$$\begin{aligned} f_{(Y_m|\cdot)}(y_m) &\propto (1 - \Phi\left(a + b(y_m)\right)) \exp(-\frac{1}{2\sigma^2}(y_m - \mathbf{X}\theta_{[m]})^2) \\ &= \Phi\left(-a - by_m\right) \exp(-\frac{1}{2\sigma^2}(y_m - \mathbf{X}\theta_{[m]})^2) \end{aligned}$$

which is the kernel of an extended skew normal distribution defined as

$$f_{skew}(x) = \frac{\phi(\frac{x-\mu_x}{\sigma})\Phi(\omega\sqrt{1+c^2} + c(\frac{x-\mu_x}{\sigma}))}{\sigma\Phi(\omega)}$$

Where

$$\mu_x = \mathbf{X}\theta_{[m]}$$

and

$$\begin{aligned} \Phi\left(-by_m - a\right) &= \Phi\left(\frac{-b\sigma}{\sigma}(y_m - \mu_x + \mu_x) - a\right) \\ &= \Phi\left(-b\sigma\frac{(y_m - \mu_x)}{\sigma} - b\mu_x - a\right) \end{aligned}$$

Thus,

$$-b\sigma = c, \quad \omega\sqrt{1+c^2} = -b\mu_x - a$$

*Supported in part by NCI grant 5T32CA106209-07, and NIDDK grant DK098285

$$\Rightarrow \omega = \frac{-a - b\mu_x}{\sqrt{1+\sigma^2 b^2}}$$

Therefore,

$$\begin{aligned} f_{(Y_m|\cdot)}(x) &= \frac{\phi(\frac{x-\mu_x}{\sigma})\Phi(\omega\sqrt{1+(-b\sigma)^2} - b\sigma(\frac{x-\mu_x}{\sigma}))}{\sigma\Phi(\omega)} \\ &= \frac{\phi(\frac{x-\mu_x}{\sigma})\Phi(-a-bx)}{\sigma\Phi(\omega)}. \end{aligned}$$

1.2. *Deriving a general full conditional distribution for the Gaussian parameters.* Each of the parameters relating to the mean model has a similar structure and the full conditional distribution is given by one generalized formula.

Let the ith entry of θ, θ_i be a mean parameter with a Gaussian prior. Let β_i, τ_i^2 be the mean and variance of θ_i and let $\mathbf{X}\theta^*$ be the product of the design matrix and parameter vector with θ_i removed from θ and the column $\mathbf{X}_{[\cdot,i]}$ removed from $\mathbf{X}$. Here the subscript $[\cdot, i]$ is used to reference a submatrix of X containing all rows but only the ith column. Finally, let $j, \cdots, J$ represent the row indices for which $\mathbf{X}_{[\cdot,i]} = 1$. Then

$$\begin{aligned} f_{(\theta_i|\cdot)} &\propto f_{(\mathbf{Y}|\cdot)} f_{\theta_i} \\ &\propto \exp(-\frac{1}{2\tau_i^2}(\theta_i - \beta_i)^2) \prod_{j=1}^{J} \exp\left(-\frac{1}{2\sigma^2}\left((y_j - \theta_i - \mathbf{X}\theta^*_{[j]})^2\right)\right) \\ &\propto \exp\left(-\frac{1}{2\tau_i^2}(\theta_i^2 - 2\theta_i\beta_i) - \frac{1}{2\sigma^2}\sum_{j=1}^{J}\left(\theta_i^2 - 2\theta_i(y_j - \mathbf{X}\theta^*_{[j]})\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\theta_i^2\left(\frac{1}{\tau_i^2} + \frac{J}{\sigma^2}\right) - \theta_i\left(\frac{2\beta_i}{\tau_i^2} + \frac{2\sum_{j=1}^{J}(y_j - \mathbf{X}\theta^*_{[j]})}{\sigma^2}\right)\right)\right) \\ &= \exp\left(-\frac{\sigma^2 + \tau_i^2 J}{2\tau_i^2\sigma^2}\left(\theta_i^2 - \frac{2\theta_i}{\sigma^2 + \tau_i^2 J}\left(\sigma^2\beta_i + \tau_i^2\sum_{j=1}^{J}(y_j - \mathbf{X}\theta^*_{[j]})\right)\right)\right) \\ &\propto \exp\left(-\frac{\sigma^2 + \tau_i^2 J}{2\tau_i^2\sigma^2}\left(\theta_i - \frac{\sigma^2\beta_i + \tau_i^2\sum_{j=1}^{J}(y_j - \mathbf{X}\theta^*_{[j]})}{\sigma^2 + \tau_i^2 J}\right)^2\right) \end{aligned}$$

Therefore,

$$(\theta_i|\cdot) \sim N\left(\frac{\sigma^2\beta_i + \tau_i^2\sum_{j=1}^{J}(y_j - \mathbf{X}\theta^*_{[j]})}{\sigma^2 + \tau_i^2 J}, \frac{\tau_i^2\sigma^2}{\sigma^2 + \tau_i^2 J}\right)$$

2. Supplementary Text B: Simulations.

2.1. *Simulation details.* For the two sample simulation study, 500 data sets were generated as follows.

1. Generate 3 variance components. $\tau^2_{pep} \sim IG(1,1)$, $\tau^2_{fc} \sim IG(1.5,1)$, $\sigma^2 \sim IG(2,1)$. Where, IG is shorthand for the inverse gamma distribution. These distributions allow for a wide variety of combinations while preserving the general relationships observed when analyzing the breast cancer data ($\hat{\tau^2_{pep}} = 2.95$, $\hat{\tau^2_{fc}} = 1.38$ and $\hat{\sigma^2} = 0.77$).

2. Generate 200 protein fold-change estimates, indexed by i. $\mu_i \sim N(0, \tau_{fc})$.

3. Generate a random number of peptides to belong to each of the 200 proteins. $n_i = N_i + 1; N \sim Poisson(4)$.

4. For each peptide generate a value. $\theta_{j(i)} \sim N(18.5, \tau_{pep})$.

5. Observations in condition k, $k = 1, 2$ are computed as $y_{ijk} = \theta_{j(i)} + \epsilon_{ijk}$ for $k = 1$ and $y_{ijk} = \theta_{j(i)} + \mu_i + \epsilon_{ijk}$. Where $\epsilon_{ijk} \sim N(0, \sigma)$.

6. simulated missing values are randomly generate using one of the three algorithms described below.

2.2. *Simulated missingness.* Simulated missing values were generate with three techniques. In all cases the goal is to generate a vector of indicator variables R_{ijk} such that the missing values are non-ignorably missing, with a higher probability of being observed as intensity increases. Parameters for each method of generating missing data were selected to provide pre-defined overall percentages of missing values.

2.2.1. *Probit Missingness.* The first method for simulating missing values is to use the mechanism from the SMP model such that $p_{ijk} = \Phi(a + by_{ijk})$. In the simulation we set $a = -9$ and $b = 0.5$then we randomly drew Bernoulli random variables (R_{ijk}) according to those probabilities to identify which y_{ijk} are missing. This results in approximately 40-50 percent of the data being missing.

2.2.2. *Logit Missingness.* Here we take a similar approach but alter the mechanism. Instead of a probit function we use an inverse logit with a quadratic function of b, i.e. $p_{ijk} = \frac{\exp(a+by_{ijk}+b^2y_{ijk})}{\exp(a+by_{ijk}+b^2y_{ijk})+1}$. In the simulation

we set $a = -8$ and $b = 0.333$ then we randomly drew Bernoulli random variables (R_{ijk}) according to those probabilities to identify which y_{ijk} are missing. This results in approximately 40-50 percent of the data being missing.

2.2.3. *LOD/MCAR Missingness.* This method combines two distinct missing data mechanisms. First we generate a vector of random Bernouli variables to create a certain amount of MCAR missingness. In the simulation we achieve this by generating $Bernoulli(.04)$ random variables. We then create vectors of random limits of detection (LOD). These limits are compared to the outcome vectors and whenever the outcome is less than the corresponding LOD the value is marked as missing. In the simulation LOD's were generated as $N(18.25, .7)$ random variables which, when combined with the MCAR missing values, resulted in approximately 40-50% of the data being missing.

500 simulations were run for each missing data mechanism. Results were only recorded for proteins where the contrast was estimable. RMSE for the probit missing data mechanisms were shown in the main text. The results for the other two mechanisms are shown in Figure 1. The SMP model, provides a substantial reduction of error relative to the other methods with every missing data mechanism that we tried. This result is consistent with the gains seen in the analysis of breast cancer data and the dilution experiment.

3. Supplementary Text C: Method Implementation. These simulated data and the breast cancer data were analyzed with six different methods. The dilution experiment makes use of an additional 3 types of imputation. All 9 methods are detailed here.

oneway The one-way ANOVA is the simplest method studied. Using the notation from the main text, we fit the model $E(y_{jk}) = \beta_0 + \theta_k$ separately to each protein, with $\hat{\theta_1} = 0$ as a side constraint (guaranteeing that the first condition is the reference). This method implemented with the lm() function in R and only point estimates were used in the paper. This method was only implemented on proteins with estimable contrasts.

twoway The two-way ANOVA implemented is also fit separately to each protein with the model $E(y_{jk}) = \alpha_j + \theta_k$, again $\hat{\theta_1} = 0$ and without an intercept this model reduces to a One-Way ANOVA whenever $J(i) = 1$. Confidence intervals were obtained using the confint() function. This method was only implemented on estimable contrasts.

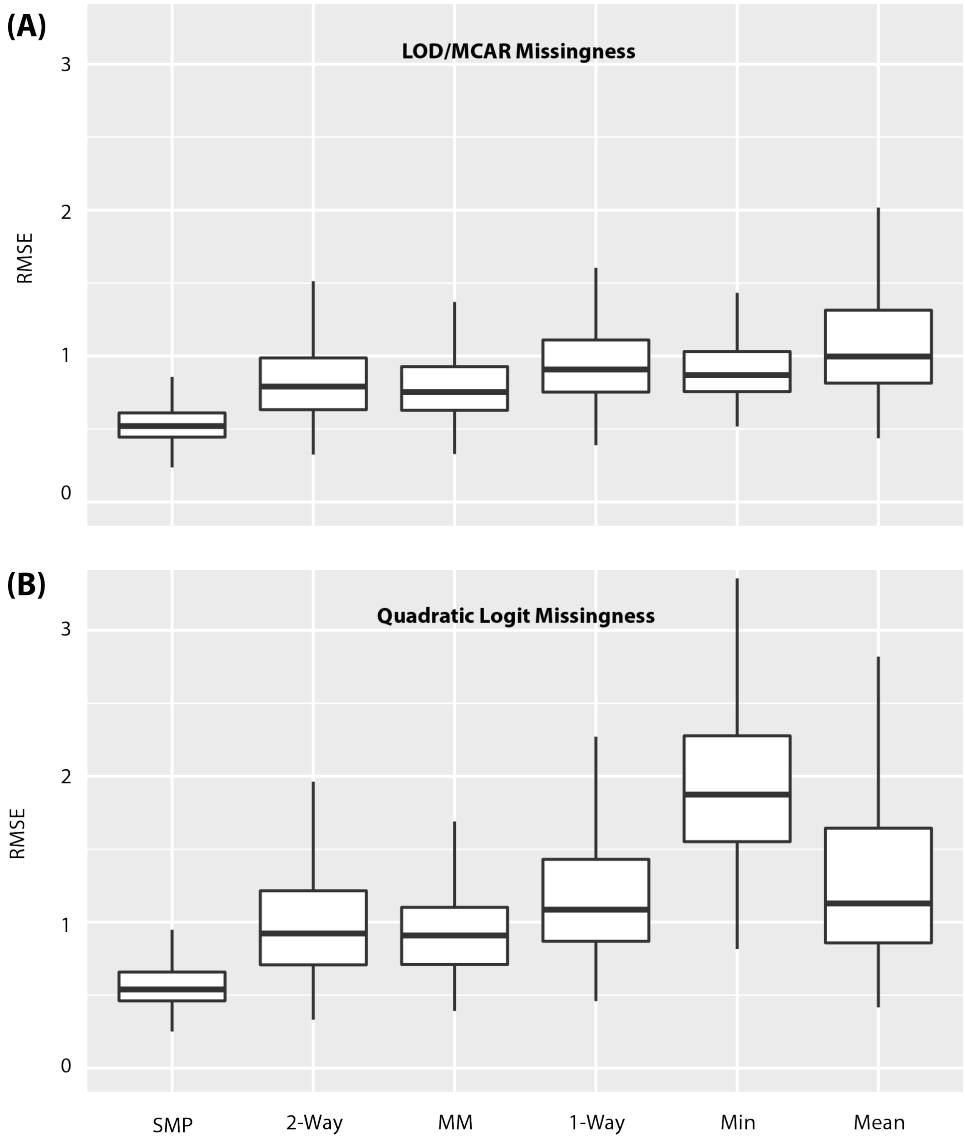

FIG 1. *Root mean squared error (RMSE) of log base 2 fold-changes from 500 simulated data sets where missing values were simulated from a two separate missing data mechanisms; a quadratic logit function (A) and values missing completely at random mixed with random Limits of Detection (LOD) (B). Only estimable contrasts are included in the plot.*

mm The mixed model is defined separately for each protein with the model $y_{jk} = \beta_0 + a_j + \theta_k + \epsilon_{jk}$, where $\hat{\theta}_1 = 0$, $a_j \sim N(0, \tau_{pep})$ is independent of ϵ_{jk} $N(0, \sigma)$. This model is reduced to a fixed effects One-Way ANOVA whenever $J(i) = 1$. Models were fit using the LME4 package (Bates et al., 2015). Bootstrap confidence intervals were obtained with the confint(method = "boot") function. This method was only implemented on estimable contrasts.

cMin This imputation method replaces every missing value with the minimum observed value in a column (condition). Estimates and intervals are obtained on the completed data using the two-way ANOVA defined above.

mean This imputation method replaces every missing value with the mean observed value in a column (condition). Estimates and intervals are obtained on the completed data using the two-way ANOVA defined above.

pMin This method is similar to the cMin imputation however it requires more data. Instead of imputing a minimum column value, the minimum observed value for each peptide sequence is used. Estimates and intervals are obtained on the completed data using the two-way ANOVA defined above.

KNN K-Nearest Neighbors imputation was performed using the "impute" package (Troyanskaya et al., 2001). Estimates and intervals are obtained on the completed data using the two-way ANOVA defined above.

SVD Imputation based on the Singular value decomposition was performed using the "bcv" package (Owen and Perry, 2009). Estimates and intervals are obtained on the completed data using the two-way ANOVA defined above.

SMP The formulation of the selection model for proteomics is primarily described in the main text. The model definition is completed with the prior and posterior distributions in Table 1. Priors for mean parameters were intended to be weakly-informative based on experience with the reasonable range of observed values in a mass spectrometry experiment. Priors for the variance component hyperparameters were

TABLE 1
The prior and posterior distributions for hyperparameters for the Bayesian model. IG is used as shorthand to denote the inverse gamma distribution. θ_i represents one of the Gaussian parameters such that $\theta_i \sim N(\beta, \tau), i = 1, ..., m$

Parameter	Prior	Posterior
τ^2	$IG(.001, .001)$	$IG(.001 + m/2, .001 + \frac{1}{2}\sum_i (\theta_i - \beta))$
β	$N(0, 10000)$	$N(\sum \theta_j/\tau^2, (\frac{1}{10000} + \frac{m}{\tau_1^2}))$
a	$N(0, 10000)$	Probit Regression Estimation
b	$N(0, 10000)$	Probit Regression Estimation

selected from default values used in BUGS (Lunn et al., 2000). The posterior distribution of (a, b) can be estimated by fitting the probit regression model

$$\Phi^{-1}(E[R_{ijk}|y_{ijk}]) = a + by_{ijk}$$

The posterior distribution is then approximated as

$$\begin{pmatrix} a \\ b \end{pmatrix} \sim N(\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}, \hat{\Sigma}) \tag{1}$$

Where $\hat{a}, \hat{b}$ and $\hat{\Sigma}$ are the parameter estimates from the probit regression and their corresponding covariance estimate, respectively. The bivariate normal distribution used here approximates the posterior distribution as a consequence of Bayesian large sample theory (Gelman et al., 2004, chap. 4).
The SMP model was fit specifying 2,000 draws and a burn-in of 500. Estimates are taken as the posterior mean for each protein contrast parameter and intervals are estimated by taking quantiles from the posterior.

4. Supplementary Text C: Determining Estimability. When dealing with missing data problems and potential imputation solutions it becomes very important to know what parameters are estimable. In the absence of missing data this is rarely a problem as standard software will simply not report estimates for inestimable parameters. After an imputation these parameters all become estimable, but as we have shown that may not be desirable. Finding out which parameters are estimable requires some thought. Simply fitting fixed effects models and checking whether or not estimates were obtained will not work. This is because standard software may set more than one parameter equal to zero in order to satisfy estimability constraints. Consequently, the parameter interpretations are subject

to change. In order to be sure that the parameters of interest are estimable, a systematic approach is required. The following algorithm was used to categorize all of the parameters in this paper. Let X denote the design matrix for a single protein and let C be an indicator vector defining a parameter of interest. For example in the two-way ANOVA model with 3 peptides and 2 conditions, there is one parameter of interest and $C = (0, 0, 0, 1)$.

1. Create the design matrix of the observed values, X_s, by removing every row of X that corresponds to a missing value.

2. For each parameter of interest create the corresponding indicator vector C.

3. Regress C onto the transpose of X_s to create regression coefficients β.

4. If the fit is perfect, i.e. $X_s^T \beta = C$, then mark the parameter estimable. If not then mark it as inestimable.

5. Repeat for each parameter of interest.

This algorithm is built into the SMP software and the results are automatically included in the output.

5. Supplementary Text D: Breast Cancer Dataset. To test model performance on real data we analyzed label-free proteomics data from two patient-derived xenograft tumors. Tissues labeled WHIM2 and WHIM16 are from basal and luminal A breast cancer tumors respectively. Sample analysis was performed via reversed phase LC-MS/MS using a Proxeon 1000 nano LC system coupled to a Q Exactive mass spectrometer (Thermo Scientific, San Jose, CA). Mass spectra were processed, and peptide identification was performed using the Andromeda search engine found in MaxQuant software ver. 2.2.1. (Cox et al., 2014). All protein database searches were performed against the uniprot human and mouse protein sequence database downloaded from the Clinical Proteomic Tumor Analysis Consortium Data Portal (https://cptac-data-portal.georgetown.edu). Peptide level data was exported from MaxQuant and the full data is provided in the supplementary tables.

This data was used for two separate analyses. First we analyzed the complete data with the SMP algorithm providing motivation for the structure of our simulation study. Posterior means for the three variance components in the model were 2.95 for peptide effects, 1.38 for protein contrasts and 0.77 for experimental error. The simulations were designed to generally mimic this structure. Complete results for the analysis are also available in the supplementary tables.

The second analysis involved removing peptides that were not present in both samples and then simulating missing values to observe the effect on estimates from various methods. LOD/MCAR Missingness, as described above, was used to create datasets with approximately 1, 5, 10, 20, 30, 40 and 50 percent missing values. The MCAR probability was always set to be one-tenth of the overall missingness percentage, e.g. when we wanted 20% missing data, we set the MCAR percentage to 20/10. The LOD standard deviation was held constant at 1, while the mean was altered. Values of 23.3, 24.4, 25.05, 25.875, 26.475, 27.05 and 27.65 gave the desired results. Each dataset was then analyzed with the same set of methods used in the simulation study. The only exception is that in the SMP model, rather than using hierarchical variance components we used a fixed prior variance of 100. This was done to avoid any shrinkage in the estimates so that every method would be guaranteed to start at the same baseline. RMSE was calculated by first taking the squared deviation of each estimate from the two-way ANOVA estimate generated on the complete data. The squared deviations were then averaged and taking the square root provided the values plotted in Figure 2.

Data and results are provided in the supplementary tables.

6. Web Appendix C: The Dilution Experiment.

6.1. *Cell growth and harvesting.* HEK (human embryonic kidney) (Graham et al., 1977), HeLa (Scherer and Syverton, 1952), and SH-SY5Y (Biedler, Helson and Spengler, 1973) were the three cell types used. Methods of cell growth and propagation followed previously utilized techniques (Paulo et al., 2011a,b). In brief, cells were propagated in DMEM supplemented with 10% FBS. Upon achieving 80% confluency, the growth media was aspirated and the cells were washed 3 times with ice-cold phosphate-buffered saline (PBS). Confluent cells were dislodged with a non-enzymatic reagent, harvested by trituration following the addition of 10 mL PBS, pelleted by centrifugation at 3,000 x g for 5 min at 4°C, and the supernatant was removed. One milliliter of HBSp (50 mM HEPES, 50 mM NaCl, pH 8.0 supplemented with 1X Roche Complete protease inhibitors), and 2% SDS were added per each 10 cm cell culture dish.

6.2. *Cell lysis and protein digestion.* Cells were homogenized by 10 passes through a 21 gauge (1.25 inches long) needle and incubated at 4°C with gentle agitation for 30 min. The homogenate was sedimented by centrifugation at 21,000 x g for 5 min and the supernatant was transferred to a new tube. Protein concentrations were determined using the bicinchoninic acid (BCA) assay (ThermoFisher Scientific). Proteins were subjected to disulfide bond reduction with 5 mM tris (2-carboxyethyl) phosphine (room temperature, 30 min) and alkylation with 10 mM iodoacetamide (room temperature, 30 min in the dark). Excess iodoacetamide was quenched with 10 mM dithiotreitol (room temperature, 15 min in the dark). Methanol-chloroform precipitation was performed prior to protease digestion. In brief, 4 parts of neat methanol were added to each sample and vortexed, 1 part chloroform was added to the sample and vortexed, and 3 parts water was added to the sample and vortexed. The sample was centrifuged at 14,000 RPM for 2 min at room temperature and subsequently washed twice with 100% methanol. Samples were resuspended in 50 mM HEPES, pH 8.5 and digested at room temperature for 13 h with LysC protease at a 100:1 protein-to-protease ratio. Trypsin was then added at a 100:1 protein-to-protease ratio and the reaction was incubated for 6 h at 37°C. Samples were subsequently acidified with 1% formic acid and vacuum centrifuged to near dryness. Each sample was desalted via StageTip, dried again via vacuum centrifugation, and reconstituted in 5% acetonitrile, 5% formic acid for LC-MS/MS processing.

6.3. *Liquid chromatography and tandem mass spectrometry.* Our mass spectrometry data were collected using a Q Exactive mass spectrometer

(Thermo Fisher Scientific, San Jose, CA) coupled with a Famos Autosampler (LC Packings) and an Accela600 liquid chromatography (LC) pump (Thermo Fisher Scientific). Peptides were separated on a 100 μm inner diameter microcapillary column packed with 20 cm of Accucore C18 resin (2.6 μm, 150 Å, Thermo Fisher Scientific). For each analysis, we loaded 1 μg, 0.25 μg, 0.0625 μg, and 0.01 μg onto the column. Peptides were separated using a 1 hr gradient of 5 to 25% acetonitrile in 0.125% formic acid with a flow rate of 300 nL/min. The scan sequence began with an Orbitrap MS1 spectrum with the following parameters: resolution 70,000, scan range 300-1500 Th, automatic gain control (AGC) target 1×105, maximum injection time 250 ms, and centroid spectrum data type. We selected the top twenty precursors for MS2 analysis which consisted of HCD high-energy collision dissociation with the following parameters: resolution 17,500, AGC 1×105, maximum injection time 60 ms, isolation window 2 Th, normalized collision energy (NCE) 25, and centroid spectrum data type. The underfill ratio was set at 2%. In addition, unassigned and singly charged species were excluded from MS2 analysis and dynamic exclusion was set to automatic.

6.4. *Data preparation.* Mass spectra were processed using MaxQuant (Cox et al., 2014) and peptide level intensities were exported for data analysis. Prior to data analysis columns were normalized and artificial treatment groups were generated. Column normalization is usually done based on the premise that the average intensity in each column should be the same. This is clearly not the case in the dilution experiment. Accordingly, multiplicative factors were used to guarantee that averages of the columns corresponded to the known dilution ratios. Another problem with the design of the dilution experiment is that the expected values are the same for all proteins in the same column. Combining this design with a model that estimates a column mean can artificially reduce the difficulty of the estimation problem. To avoid this artifact, we randomly permuted the columns for each protein while keeping track of the actual dilution level for each protein treatment combination. The same permutation is used for all peptides belonging to the same protein group. Both the exported data from MaxQuant and the prepared data are available in the supplementary tables.

Proteins were assumed to be nested within treatment groups so that the same protein from different cell lines did not share any parameters from the mean model. The data was analyzed with seven methodologies, described above, and performance was assessed in terms of root mean squared error and interval coverage.

For the SMP model, convergence was assessed by re-running the analysis

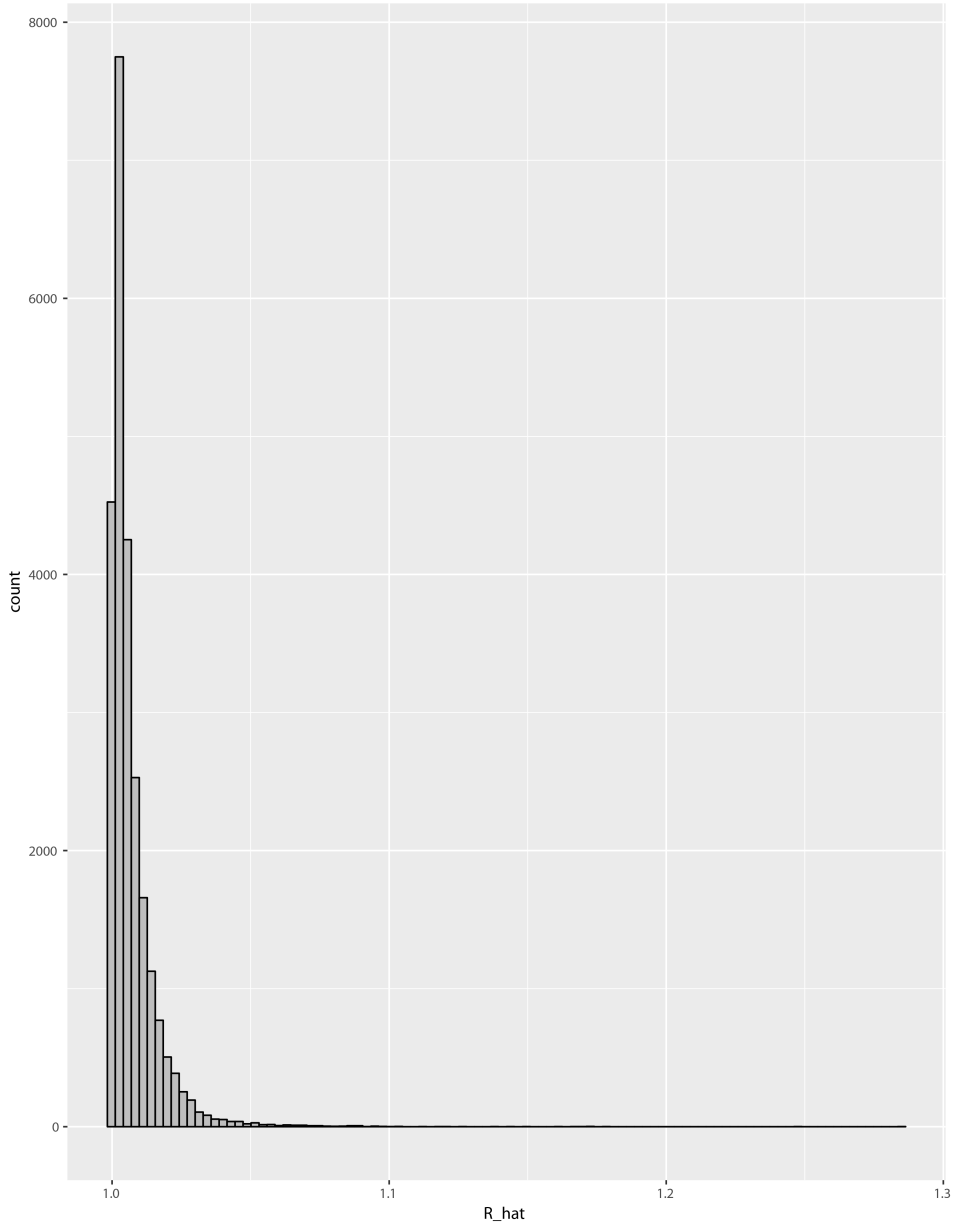

FIG 2. *Histogram of R-hat statistics from all parameters in the SMP fit of the dilution experiment.*

with different random seeds and computing R-hat statistics for every parameter (Gelman and Hill, 2006, pp 251-276). When the model has converged these values should be equal to one. A histogram of R-hat values from the dilution experiment is shown in Figure 2.

References.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67** 1–48.

Biedler, J. L., Helson, L. and Spengler, B. A. (1973). Morphology and growth, tumorigenicity, and cytogenetics of human neuroblastoma cells in continuous culture. *Cancer Research* **33** 2643–2652.

Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N. and Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & cellular proteomics : MCP* **13** 2513–26.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis Second Edition* **1**. Chapman & Hall/CRC.

Graham, F. L., Smiley, J., Russell, W. C. and Nairn, R. (1977). Characteristics of a human cell line transformed by DNA from human adenovirus type 5. *The Journal of general virology* **36** 59–74.

Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* **10** 325–337.

Owen, A. B. and Perry, P. O. (2009). Bi-cross-validation of the SVD and the nonnegative matrix factorization. *The Annals of Applied Statistics* **3** 564–594.

Paulo, J. A., Urrutia, R., Banks, P. A., Conwell, D. L. and Steen, H. (2011a). Proteomic analysis of a rat pancreatic stellate cell line using liquid chromatography tandem mass spectrometry (LC-MS/MS). *Journal of Proteomics* **75** 708–717.

Paulo, J. A., Urrutia, R., Banks, P. A., Conwell, D. L. and Steen, H. (2011b). Proteomic analysis of an immortalized mouse pancreatic stellate cell line identifies differentially-expressed proteins in activated vs nonproliferating cell states. *Journal of Proteome Research* **10** 4835–4844.

Scherer, W. F. and Syverton, J. T. (1952). Studies on the propagation in vitro of poliomyelitis viruses. III. The propagation of poliomyelitis viruses in tissue cultures devoid of nerve cells. *The Journal of experimental medicine* **96** 389–400.

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** 520–525.

Department of Cell Biology
Harvard Medical School
240 Longwood Ave
Boston, MA, 02115
USA
E-mail: obrienj@hms.harvard.edu

Department of Biostatistics
University of North Carolina at Chapel Hill
135 Dauer Drive
3101 McGavran-Greenberg Hall, CB 7420
Chapel Hill, NC 27599
USA

Department of Biochemistry and Biophysics
University of North Carolina at Chapel Hill
120 Mason Farm Rd, Campus Box 7260
Chapel Hill, NC 27599
USA