Published in final edited form as:
Stat Biosci. 2009 November ; 1(2): 228–245. doi:10.1007/s12561-009-9013-2.

Bayesian Analysis of iTRAQ Data with Nonrandom Missingness: Identification of Differentially Expressed Proteins

Ruiyan Luo,
Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, USA

Christopher M. Colangelo,
W.M. Keck Foundation, Biotechnology Resource Laboratory, Yale University School of Medicine, New Haven, CT 06511, USA

William C. Sessa, and
Department of Pharmacology, Yale University School of Medicine, New Haven, CT 06510, USA

Hongyu Zhao
Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, USA

Ruiyan Luo: ruiyan.luo@yale.edu

Abstract

iTRAQ (isobaric Tags for Relative and Absolute Quantitation) is a technique that allows simultaneous quantitation of proteins in multiple samples. In this paper, we describe a Bayesian hierarchical model-based method to infer the relative protein expression levels and hence to identify differentially expressed proteins from iTRAQ data. Our model assumes that the measured peptide intensities are affected by both protein expression levels and peptide specific effects. The values of these two effects across experiments are modeled as random effects. The nonrandom missingness of peptide data is modeled with a logistic regression which relates the missingness probability for a peptide with the expression level of the protein that produces this peptide. We propose a Markov chain Monte Carlo method for the inference of model parameters, including the relative expression levels across samples. Our simulation results suggest that the estimates of relative protein expression levels based on the MCMC samples have smaller bias than those estimated from ANOVA models or fold changes. We apply our method to an iTRAQ dataset studying the roles of Caveolae for postnatal cardiovascular function.

Keywords

Bayesian hierarchical model; iTRAQ; Mixed-effects model; Nonignorable missing; Protein quantitation

1 Introduction

One main objective of proteomic research is to detect and quantify all proteins present in a biological sample. iTRAQ, a shotgun technique using Isobaric Tags for Relative and Absolute Quantitation, has become commonly used because of its improved quantitative

Correspondence to: Ruiyan Luo, ruiyan.luo@yale.edu.

reproducibility and higher quantification sensitivity [16] compared to other methods such as 2DE [9], ICAT [3], and DIGE [4, 10]. Using four or eight isobaric tags, iTRAQ can simultaneously analyze up to eight biological samples [2, 12]. Peptides digested from different samples of protein mixtures are labeled with different tags independently, mixed together, separated, and studied by MS (mass spectrometry) and MS/MS (tandem mass spectrometry). The resulting collection of mass spectra provides information on peptide identification and quantification, which can be utilized to identify and quantify relative protein expression levels.

We use the data from an iTRAQ experiment with four isobaric tags (114, 115, 116, and 117) as an example to illustrate the iTRAQ data format in Table 1. Each row represents a specific peptide identified from a software, such as MASCOT [11], which searches a protein sequence database to identify the peptide corresponding to a specific peak in the mass spectrum. The peptides thus identified are given in the second column. The peak areas for different samples labeled with different tags are shown in the last four columns, and their values can be used to calculate the relative abundance of a given peptide across samples. Each peptide may arise from different spectra and hence have multiple observations in an experiment. For example, the first three rows in Table 1 correspond to the same peptide across all the spectra. Missing peptides is a common phenomenon in iTRAQ data. That is, a peptide may be only observed in some of the samples, or some spectra, or some experiments. For example, the seventh row in Table 1 shows that peptide "DVDEIEAWISEK" is only observed in the samples labeled with 114 and 117. The fifth row indicates that in one spectrum, the intensities of the peptide "DLASVQALLR" are missing in all the samples. When multiple experiments are conducted, a peptide may be found to be missing in one experiment but observed in some other experiments (not shown in Table 1).

As seen above, the basic unit of iTRAQ data is the peptide. Each peptide has an associated intensity level. Several factors can affect the observed peptide intensities, i.e., the area columns in Table 1. The most obvious factor is the level of the protein in the sample that generates the peptide. Peptide specific features, such as ionization and fragmentation efficiency, affect the intensity levels for different peptides derived from the same protein. This is easily seen in Table 1, where all peptides are derived from the same protein. In addition, other factors such as sample preparation and experimental variation also contribute to the variabilities in the observed iTRAQ data. Hill et al. [5] illustrate in detail the possible sources of variations in iTRAQ data.

Another commonly encountered issue in iTRAQ data analysis is data missingness. Due to the nature of the technology, overlap in protein and peptide identifications between replicate experiments is less than ideal, and certain peptides are only observed for some samples in some spectra, leading to a large amount of missing data. Table 2 gives the number of proteins and peptides that are identified in only one, only two, or all three experiments when iTRAQ is performed three times on the same biological sample. It can be seen that only about 1/3 of the proteins we identified in all three experiments, whereas only about 1/4 of the peptides produced by these proteins we observed in all experiments. Liu et al. [6] and Wang et al. [15] suggested that the probability that a protein is missing is not random, but rather related to its abundance. Less abundant peptides are harder to detect due to the data-dependent acquisition of the analysis process, hence more likely to be missing. This is a nonignorable missing data problem. Ignoring the nonrandom missing pattern in statistical analysis may introduce significant bias in statistical inference and scientific conclusions.

To identify differentially expressed proteins, one common approach is to calculate the ratio of the observed peptide intensities (the area columns in Table 1) between two samples and to

compare the calculated ratios against prespecified upper and lower bounds. However, the criterion for threshold selection is subjective. For example, Seshi [14] considered iTRAQ ratios >5/4 or <4/5 as significant, whereas Salim et al. [13] used thresholds 1.20 and 0.83. These thresholds fail to consider the variability in data and are not statistically based. Oberg et al. [8] and Hill et al. [5] applied ANOVA models to incorporate the variability sources in inferring differentially expressed proteins. But they do not consider the nonrandom missingness, potentially biasing their results. In this paper, we introduce a novel approach to inferring the relative protein expression levels and hence to identify differentially expressed proteins. We model the measured peptide intensities as the results of both protein expression levels and peptide specific effects. For iTRAQ data from multiple experiments, we utilize a Bayesian hierarchical model in the sense that the model has an observation component that models the observed peptide intensities as random effects whose conditional distribution depends on the expected protein expression levels and peptide effects, and a second (hierarchical) component that defines the distributions of these expected values. If a sample is labeled with multiple tags in a single experiment, the variations across different tags are modeled as random effects. In this paper, we also describe a model for iTRAQ data from a single experiment. As for the nonrandom missingness, we use a logistic regression to model the missingness probability as a function of the protein expression level. Based on this model set-up, we infer differentially expressed proteins through posterior inferences.

The paper is organized as follows. Section 2 develops the hierarchical model and details the inferential procedure. Section 3 reports a simulation study comparing our method with ANOVA methods and ratio estimates, and studies the robustness of our method. Section 4 reports the analysis of a mouse caveolin-1 experiment, and discussion follows in Sect. 5. We describe the detailed MCMC scheme in Appendix A, and a model for iTRAQ data from a single experiment in Appendix B.

2 Model

We first describe the model for iTRAQ data from multiple experiments and estimate the relative expressions of proteins that are present in all experiments. We assume that the labeling effects have been removed by normalization methods such as quantile normalization [1]. Throughout the paper, we consider log-transformed peptide intensities and protein expressions. We assume that there are S (≥ 2) biological samples studied in K (≥ 2) experiments. Multiple isobaric tags may label the same sample in one experiment. We use $L_s \geq 1$ to denote the number of tags labeling the sth sample. Then $\sum_s L_s = M$ is the number of isobaric tags used in one experiment, which is 4 when we use 4-plex isobaric reagents and $M = 8$ in the 8-plex version. Assume that there are I proteins in the sample and there are J_i peptides for the ith protein. For the lth label of the sth sample in the kth experiment, let y_{kijsln} denote the observed intensity for the jth peptide of the ith protein from the nth spectrum. Note that j should be more appropriately denoted as $j(i)$ to explicitly indicate that peptides are nested within proteins, and l should be denoted as $l(s)$ to indicate the lth labeled tag of the sth sample. For notational simplification, we omit the parentheses. The measured intensity of a peptide depends on the protein expression level and the peptide effect. Let x_{kisl} denote the expression level of the ith protein of the sth sample with the lth labeling tag in the kth experiment. Let z_{kij} denote the peptide effect for the jth peptide of the ith protein in the kth experiment. We consider an additive model for y_{kijsln} ($k = 1, \ldots, K$; $i = 1, \ldots, I$; $j = 1, \ldots, J_i$; $s = 1, \ldots, S$; $l = 1, \ldots, L_s$; $n = 1, \ldots, N_{kijsl}$):

$$y_{kijsln} = x_{kisl} + z_{kij} + \varepsilon_{kijsln}, \quad (1)$$

which corresponds to a multiplicative model in the original scale. In (1), we assume $\varepsilon_{kijsln} \sim N(0, \sigma_\varepsilon^2)$ independently, where $N(0, \sigma_\varepsilon^2)$ denotes a Normal distribution with mean 0 and variance σ_ε^2.

In addition to the additive model in (1), we also consider the multiplicative model $y_{kijsln} = x_{kisl} \times z_{kij} + \varepsilon_{kijsln}$ on a small dataset with one protein and 11 peptides observed in three caveolin-1 experiments. The inferences from both models are quite close in terms of the magnitudes of the residual standard deviation (0.58 for the additive model vs. 0.60 for the multiplicative model) and the ratio of sum of squares of predicted values and sum of squares of original data R^2 (0.73 for the additive model vs. 0.69 for the multiplicative model). The residuals vs. fitted values plots are also similar (not shown). This is also true when we apply both models to the data in the original scale. Since the multiplicative model in the logarithm scale and the additive model in the original scale do not greatly improve the inference (or even do worse), we use model (1) which is also easy to interpret.

Missing Data Mechanism

Peptide missingness presents a challenge even when we focus on proteins that are detected in all experiments. It is known that the probability of peptide missingness depends on the intensity of the peptide: lower intensity peptides are harder to detect. So there is a nonignorable missing data problem. To motivate a statistical model for missing peptide probability, we study the proportion of peptides observed in one experiment but missing in another experiment. As shown in Fig. 1, there is a negative correlation between missingness probability and peptide intensity. Furthermore, there is an approximate linear relationship between the peptide missingness probability and the observed intensity at the logit scale. Therefore, we model the missingness probability through a simple logistic regression,

$$\text{logit}\left(P\left(I_{kijsln}=0 \mid y_{kijsln}, a, b\right)\right) = a + b \times y_{kijsln}, \tag{2}$$

where $I_{kijsln} = 0$ indicates that the jth peptide of the ith protein is missed in the kth experiment, the lth replicate of the sth sample and the nth spectrum. Formula (2) implies that the logit of the probability of peptide missingness is linearly dependent on its intensity. We expect $b < 0$, because peptides with lower intensities are more likely to be missing.

Priors

Noting the hierarchical structure of the iTRAQ data and taking into account the variability across experiments and samples, we utilize a Bayesian hierarchical framework to model the data. We assume that x_{kisl} and z_{kij} are independently normally distributed across different experiments, i.e.,

$$x_{kisl} \sim N\left(x_{isl}, \sigma_x^2\right), \tag{3}$$

$$z_{kij} \sim N\left(z_{ij}, \sigma_z^2\right), \tag{4}$$

where x_{isl} and z_{ij} denote the protein and peptide effects averaged over multiple experiments, respectively. The protein expression levels in different replicates (labeled with different tags) of the same sample are also assumed to be normally distributed:

$$x_{isl} \sim N\left(x_{is}, \sigma_{\delta}^2\right), \tag{5}$$

where x_{is} denotes the expression level of the ith protein in the sth sample. Assumptions (3)–(5) lead to an equivalent form of (1):

$$y_{kijsln} = x_{is} + z_{ij} + e^t_{isl} + e^x_{kisl} + e^z_{kij} + \varepsilon_{kijsln}, \tag{6}$$

where $e^x_{kisl} \sim N(0, \sigma_x^2)$ and $e^z_{kij} \sim N(0, \sigma_z^2)$ denote the random effects across experiments, and $e^t_{isl} \sim N(0, \sigma_{\delta}^2)$ denotes the variation among multiple replicates of the same sample. Formula (6) is a mixed-effects model. To ensure the identifiability of the model, we restrict $x_{i1} = 0$. Then x_{is} denotes the expression level of the ith protein in the sth sample relative to the first sample.

The second level of priors are normal distributions for x_{is} and z_{ij}:

$$x_{is} \sim N\left(0, \tau_x^2\right) \quad \text{for } s>1, \tag{7}$$

$$z_{ij} \sim N\left(0, \tau_z^2\right). \tag{8}$$

When we further assume hyperpriors for the hyperparameters, we finish the hierarchical model (Fig. 2) and can infer the posterior distributions of relevant parameters, x_{is}, by MCMC simulations. Appendix A describes other hyperpriors and the MCMC updates in detail. Hence we can summarize the simulated posterior distributions with statistics such as posterior means, standard deviations and quantiles, and identify differentially expressed proteins.

When a sample is labeled with a unique isobaric tag in an experiment, there is no replicate variation component within a sample. We note that it is easy to modify the model and the MCMC updates for statistical inference in this scenario. We will not discuss it further in this paper.

Single Experiment

When the iTRAQ data is from one experiment, we can similarly model the observed peptide intensities as the result of both protein expression levels and peptide effects, and model the nonrandom missingness through a logistic regression. We can further apply normal distributions as priors for protein expressions and peptide effects. The difference from the case of multiple experiments is that the experimental variability cannot be modeled. Appendix B describes this model and MCMC updates in more details.

Comparison to ANOVA Model

The most important difference between our Bayesian model and the ANOVA model proposed by Hill et al. [5] and Oberg et al. [8] is that we clearly model the nonignorable missingness in iTRAQ data. Oberg et al. [8] remarked at the end of their paper that using a censoring mechanism to fit the model would be a natural next step. Instead of censoring the data at an unknown threshold value, we model a higher probability of peptide missingness

for lower peptide intensities. Our Bayesian model also differs from the ANOVA model in the sources of variations included in the model. In addition to the terms in our model, the ANOVA analysis also considers the labeling effect, the interaction between labeling and experimental effect, and variable peptide effects under different conditions (we talk about this in Discussion). The experimental effect and the replicative effect (when multiple tags label a sample) are considered constants for all proteins in the ANOVA model. In contrast, we model them as random effects that are specific to peptides and (or) proteins.

3 Simulation Study

We simulate data from a 4-plex version of iTRAQ on one protein containing ten peptides across three replicate experiments. Each sample is labeled with a distinct isobaric tag. In this case, there is no need to model the replicate effects specified by prior (5). We assume $x = (0, -0.04, -0.48, -0.66)$ to be the true relative protein expression levels in log scale compared to the first sample. Under different parameter values for σ_x, σ_z, and σ_δ, we simulate data as follows: (1) sample $x_{ks} \sim N(x_s, \sigma_x^2)$ and $z_{kj} \sim N(z_j, \sigma_z^2)$, where $z_j \sim N(0, 1)$; here we dismiss subscripts i and l since there is only one protein and only one isobaric tag for a sample in an experiment; (2) sample $y_{ksjn} \sim N(x_{ks}+z_{kj}, \sigma_\varepsilon^2)$, calculate the missing data probability $P(I_{ksjn} = 0)$, and determine the missing pattern I_{ksjn}. We take $a = -0.16$ and $b = -1.03$ in the simulation, based on the posterior inference of a small subset of a real data.

We analyze the simulated data with our Bayesian method and infer the relative protein expression levels through the MCMC samples. For comparison, we also analyze the data with the ANOVA model proposed by Hill et al. [5] and Oberg et al. [8], and calculate the means of the log ratios of peptide intensities. For each parameter setting, we simulate ten data sets and summarize the results from one data set in Table 3. The Bayesian method and the ANOVA analysis provide measures of the uncertainties of estimates. We either obtain the 95% credible intervals of the posterior distributions or the 95% confidence intervals for the estimates from the ANOVA analysis. When performing the ANOVA analysis, we consider two models. “ANOVA 1” includes the sample effect, peptide effect, experimental effect, and the interaction of sample effect and peptide effect. “ANOVA 2” removes the interaction term from “ANOVA 1.” From Table 3 we observe that all but one credible interval cover the true values when using our Bayesian method to analyze the data. But about 1/3 of the confidence intervals from ANOVA analysis fail to cover the true values, including the case where Bayesian analysis fails (estimate x_3 for the simulated data when $\sigma_x^2=0.01, \sigma_z^2=1,$ and $\sigma_\varepsilon^2=1.5$). Comparing the estimates to the true values, we find that our Bayesian estimates have smaller bias than those from ANOVA analysis. Figure 3 draws the boxplots of the biases of the estimates using different methods for all six parameter settings. It is clear that the Bayesian method leads to the smallest bias. The better coverage and smaller bias of the Bayesian method are consistently observed in the analyses of the other nine simulated data sets. In the 60 analysis (10 data sets for each parameter setting), the 95% credible intervals from our Bayesian method fail to cover the true values 3% of the time, but the 95% confidence intervals from the ANOVA method fail in 1/3 of the cases. The means of the biases for estimates of x from the Bayesian analysis are at least 1/2 smaller than those from the ANOVA method. The lengths of the credible intervals and confidence intervals are specific to a data set or the parameter setting. Neither is consistently smaller than the other.

In the above results, we simulated data according to our model, which may favor our approach. To study the robustness of our approach, we also consider a different missing mechanism. For each experiment, we first simulate whether each peptide is present from a Bernoulli distribution with probability p, which determines the potential frequency r_j of the presence of peptide j in $K = 3$ experiments ($r_j = 0$, 1, 2, or 3). Given r_j, we sample the

peptide effect $z_j \mid r_j \sim \text{logGamma}\ (l_{rj}, sh_{rj}, sc_{rj})$ for $r_j > 0$. The density function of a log-gamma distribution with shape $a > 0$, scale $b > 0$, and location c is

$$\text{logGamma}\ (x \mid a, b, c) = \frac{1}{b\Gamma(a)} \exp\left(\frac{a(x-c)}{b} - \exp\left(\frac{x-c}{b}\right)\right). \tag{9}$$

Peptides with $r_j = 0$ are missed. Then we simulate the variabilities across experiments: $x_{ks} \sim N(x_s, \sigma_x^2), z_{kj} \sim N(z_j, \sigma_z^2)$. Finally we follow the second step in the previous study to simulate y_{ksjn} and I_{ksjn}. This mechanism differs from our model in two ways: (1) the distribution for z_j differs; and (2) the missing data mechanism differs since the simulation of possible presences of peptides from the Bernoulli distribution will also cause peptides missed. The resulting peptide frequency may be less than r_j. We study how our method performs for the data simulated under this missing mechanism. We consider different values for the success probability in the Binomial distribution (0.9 and 0.2). For each case, we simulate ten data sets and analyze them with our method. Table 4 gives the means and standard deviations of the ten estimations. We find that the means are close to the true values and the inference is not sensitive to the new mechanism of peptide missing. Compared to the results obtained from the ANOVA analysis which contains the main effects of protein, peptide and experiments, the estimates from our Bayesian analysis are closer to the true values and have less variability (except x_4 for Bi(3,0.2)) in the estimations.

In previous simulations, we fix the number of observations for each peptide as the same. When a peptide is not observed in an experiment, we assume that only one spectrum is missing and impute the values for all samples in only one observation. To study the effect of varying number of observations for different peptides on our inference, we randomly sample these numbers from a Poisson distribution. The rate of the distribution is randomly picked from a set of values. We also apply the missing mechanism described in the previous paragraph with $p = 0.5$. Under this scheme, we simulate ten data sets and analyze them with our method and the ANOVA model. From the calculated means and standard deviations in Table 4 we see that the distribution of the number of observations does not have great effect on the inference, and the estimates from our method have less variability than those obtained from ANOVA analysis.

4 Case Study

We apply our method to an iTRAQ dataset which aims to identify proteins affected by caveolin-1. Caveolin-1 is essential to the formation of caveolae, while the functional perturbations in the caveolae and the caveolae coat proteins may cause a wide range of diseases from cancer to a rare form of muscular dystrophy. Recent studies from mice suggest that they may be important for postnatal cardiovascular function [7]. Comparing the protein profiles from wild-type (WT) mice and knock-out Cav-1 (KO) mice using iTRAQ, we can explore the physiological and pathophysiological roles of caveolins for postnatal cardiovascular function. Samples from three KO mice and three WT mice were labeled with iTRAQ reagents as shown in Table 5. Among the 424 proteins identified in the study, a total of 138 common proteins were identified in all three comparisons of the WT/KO mice from iTRAQ analysis (Table 2). Focusing on the 4765 peptides of these 138 common proteins, we found that 2124 of them were observed in all three experiments.

We first perform quantile normalization with each protein in the two replicates of each sample. Then we do two iterations of quantile normalization on each pair of samples to remove the systematic bias in the data. Applying our method to the log transformed value of the normalized data, we conduct 101000 iterations of MCMC updates and take the first 6000

as burn-in. The simulation takes 138 hours on the caveolin data with 4765 peptides and 200684 observations. Sampling every tenth iteration, we get 9500 samples, based on which we infer the posterior distributions of protein expression levels.

We illustrate the inferred posterior means of the relative protein expression levels in Fig. 4. We also depict the upper and lower 2.5% posterior quantiles in the figure. From these posterior inferences, we can further identify differentially expressed proteins. For example, if we require the 2.5% quantile above zero or the 97.5% quantile below zero, there are 19 up-regulated and 7 down-regulated proteins. We summarize the posterior inferences of other parameters in Table 6. For this normalized data, the randomness of peptide effects across experiments contributes the most significant source of variation (313.141). The replicate variation within a sample is almost ignorable (0.002). We infer the slope parameter in (2) to be negative (−0.217), implying that peptides with lower intensities are more susceptible to be missing.

To make a comparison with other methods, we also apply the ANOVA method to the data. Since there are 138 identified proteins and 4765 identified peptides, it is difficult to estimate all of the parameters in the ANOVA model simultaneously using current software and computers. Applying the stagewise regression idea in Oberg et al. [8], we first estimate the effects of experiments, proteins, and peptides (the first two groups of model (1) in Oberg et al. [8]), and then we take the residuals as responses for estimating the effects of samples, interactions between samples and proteins, peptides. The sample-related parameters are estimated for each protein individually, assuming that each protein has a different variance parameter, rather than a global variance parameter. Regarding the proteins as differentially expressed where the 95% confidence intervals do not cover zero, we find 60 up-regulated and 26 down-regulated proteins. They contain all the differentially expressed proteins inferred from our Bayesian model. Focusing on the proteins that are only found by ANOVA, we study their missing patterns and compare the estimates from both methods. We find that for 35 of the 41 (= 60 − 19) up-regulated and 15 of the 19 (= 26 − 7) down-regulated proteins, the differences of estimates of expression levels from both models may be due to missingness. Another reason that ANOVA identifies more proteins is likely due to the fact that protein-by-protein estimation leads to smaller variances than the global variance under our Bayesian approach. So the credible intervals from Bayesian analysis have wider, and more appropriate, ranges than the confidence intervals from ANOVA model.

5 Discussion

We have developed a novel Bayesian model to analyze iTRAQ data from multiple experiments or a single experiment. In our model, the observed peptide intensities are influenced by both the protein expression levels and the peptide effects. For data from multiple experiments, these two effects across experiments are modeled as random effects. If a sample is labeled with multiple isobaric tags, our model also allows random effects across replicates. We explicitly model the nonignorable missingness for peptides, which is a common phenomenon in iTRAQ data. The logit probability of peptide missingness is assumed to be linearly dependent on its intensity. We implement an MCMC approach to simulate the posterior distributions of relative protein expression levels. The MCMC samples provide both estimates of the expressions and measure of uncertainty for the estimates. Compared to the estimates from the ANOVA analysis and the simple log ratio calculation, we find that the estimates from the MCMC samples greatly reduces the bias due to missing data.

In our model, we assume that the logit of the missingness probability is linearly dependent on the peptide intensity y_{kijsln} (2), and the later depends on the protein expression levels,

peptide effects, and several variation terms (6). For a particular peptide *j*, in addition to the variation (ε_{kijsln}) across multiple spectra in an experiment, experimental variations are modeled at both the protein (e^{x}_{kisl}) and peptide (e^{z}_{kij}) levels. A small peptide effect specific to a particular experiment (z_{kij}) may cause the missingness of the peptide in that experiment (*k*). When both protein effect and peptide effect are large in an experiment *k*, this peptide will be observed in experiment *k*, but an extremely small value of ε_{kijsln} can lead to the missingness of this peptide in spectrum *n* of experiment *k*. So this model explains the peptide missingness both at the experiment level and at the spectrum level.

We have performed simulation studies to check whether our analysis is sensitive to this assumption of missingness. We simulate data sets from different missing mechanisms and analyze them with our Bayesian method. The estimated values are close to the true values and have smaller bias than the results from the ANOVA analysis. Furthermore, we also check how variable number of spectra for peptides affects our analysis, since when a peptide is missing in an experiment, we impute the values in only one MS spectrum. It is found that when we sample the number of MS spectra for a peptide from a Poisson distribution, our analysis leads to estimates close to the true values. This implies that our method is robust to these model violations.

Labeling effect is an issue that is not directly addressed in our model. The ability of peptides' linkage to isobaric reagents may vary, implying peptide-tag specific labeling effect. Modeling all such labeling effects increases the number of model parameters dramatically. If we treat the labeling effects as constants for all peptides, this amounts to adding a constant specific to each tag in model (1). Due to the limitation of the data in caveolin study, the labeling effect is confounded with signals. In this paper, we first perform normalization to remove the labeling effect and systematic bias, and then apply our method to infer the relative protein expressions. In practical studies, we suggest to randomize the isobaric tags applied to samples when multiple experiments are conducted.

The fast convergence requirement is a challenge to our Bayesian approach. For a larger scale study, more MCMC iterations and hence longer time are needed to ensure the convergence. Although the Bayesian method is slower than the ANOVA method, the latter cannot fit all the involved parameters simultaneously using current software and computers. Oberg et al. [8] suggest to use the stagewise regression and then to infer the sample effects based on protein-by-protein estimation. But to get correct answers from the stagewise approach, it is necessary that the portions of the linear model design matrix corresponding to the multiple stages be orthogonal, which is not necessarily true.

In this study, we assume that all of the peptide-based observations accurately reflect the intact proteins. As a result, we ignore the possibility of homologous genes resulting in two or more proteins that share identical and nonidentical peptides as well as the possibility of post-transcriptional modifications. In addition to ignoring labeling effects, we do not include the interactions between peptide effects and sample conditions comparing to the ANOVA model. This corresponds to the assumption that certain proteins will have differential expressions under different conditions, but that any change in protein expression will affect all of the peptides for that protein equally. We expect this to be the common case, except for certain biological conditions: for example, a post-translational modification that involved a peptide substitution [8]. Despite these limitations, our method explicitly models the nonrandom missingness of iTRAQ data and provides a great improvement in estimating the relative expressions of proteins.

Acknowledgments

The work was supported in part by NIH grants HV28286, DA018343, GM59507 and NSF grant DMS 0714817. The work was also supported in part by "Yale University Biomedical High Performance Computing Center" and NIH grant RR19895, which funded the instrumentation.

Appendix A: MCMC Updates

We assume inverse gamma distributions as priors for the hyperparameters of variance: $\sigma_x^{-2} \sim \text{Gamma}\ (\gamma_1, \gamma_2)$, $\sigma_z^{-2} \sim \text{Gamma}\ (\gamma_3, \gamma_4)$, $\sigma_\delta^{-2} \sim \text{Gamma}\ (\gamma_5, \gamma_6)$, and $\sigma_\varepsilon^{-2} \sim \text{Gamma}\ (\gamma_7, \gamma_8)$, , where γ_1 and γ_2 denote the shape and scale parameters of a gamma distribution, respectively. We assume $a \sim N(0, v^2)$ and $b \sim N(0, v^2)$. The joint distribution of the model is

$$
\begin{aligned}
f(y, I_y, x, z, \phi) = & \prod_{kijsl} \{\text{MVN}(\mathbf{y}_{kijsl} \mid (x_{kisl} + z_{kij})\mathbf{1}, \sigma_\varepsilon^2 \mathbf{I}) p(\mathbf{I}_{kijsl} \mid \mathbf{y}_{kijsl}, a, b)\} \times \prod_{kisl} N(x_{kisl} \mid x_{isl}, \\
& \sigma_x^2) \\
& \times \prod_{isl} N(x_{isl} \mid x_{is}, \\
& \sigma_\delta^2) \times \prod_{i,s>1} N(x_{is} \mid 0, \tau_x^2) \times \prod_{kij} N(z_{kij} \mid z_{ij}, \sigma_z^2) \times \prod_{ij} N(z_{ij} \mid 0, \tau_z^2) \times N(a \mid 0, v^2) \times N(b \mid 0, v^2) \times \text{invGamma}\ (\sigma_\varepsilon^2) \\
& \times \text{invGamma}\ (\sigma_x^2) \\
& \times \text{invGamma}\ (\sigma_z^2) \times \text{invGamma}(\sigma_\delta^2),
\end{aligned} \tag{10}
$$

where MVN(. | μ, Σ) denotes a multivariate normal distribution with mean vector μ and covariance matrix Σ, invGamma(.) denotes an inverse gamma distribution, and $p(\mathbf{I}_{kijsl} \mid \mathbf{y}_{kijsl}, a, b)$ can be determined by formula (2). The full conditional distributions for involved parameters are given below.

1. Protein and peptide effects: x_{kisl}, z_{kij}, x_{isl}, x_{is}, and z_{ij}.

$$
x_{kisl} \sim N\left(\frac{\sum_j \frac{\bar{y}_{kijsl} - z_{kij}}{\sigma_\varepsilon^2 / N_{kijsl}} + \frac{x_{isl}}{\sigma_x^2}}{\sum_j \frac{N_{kijsl}}{\sigma_\varepsilon^2} + \frac{1}{\sigma_x^2}}, \frac{1}{\sum_j \frac{N_{kijsl}}{\sigma_\varepsilon^2} + \frac{1}{\sigma_x^2}} \right), \tag{11}
$$

$$
z_{kij} \sim N\left(\frac{\sum_m \frac{\bar{y}_{kijsl} - x_{kisl}}{\sigma_\varepsilon^2 / N_{kijsl}} + \frac{z_{ij}}{\sigma_z^2}}{\sum_m \frac{N_{kijsl}}{\sigma_\varepsilon^2} + \frac{1}{\sigma_z^2}}, \frac{1}{\sum_m \frac{N_{kijsl}}{\sigma_\varepsilon^2} + \frac{1}{\sigma_z^2}} \right), \tag{12}
$$

$$
x_{isl} \sim N\left(\frac{\sum_k x_{kisl} / \sigma_x^2 + x_{is} / \sigma_\delta^2}{K / \sigma_x^2 + 1 / \sigma_\delta^2}, \frac{1}{K / \sigma_x^2 + 1 / \sigma_\delta^2} \right), \tag{13}
$$

$$
x_{is} \sim N\left(\frac{\sum_l x_{isl} / \sigma_\delta^2}{L_s / \sigma_\delta^2 + 1 / \tau_x^2}, \frac{1}{L_s / \sigma_\delta^2 + 1 / \tau_x^2} \right) \quad \text{for } s>1, \tag{14}
$$

$$z_{ij} \sim N\left(\frac{\sum_k z_{kij}/\sigma_z^2}{K/\sigma_z^2+1/\tau_z^2}, \frac{1}{K/\sigma_z^2+1/\tau_z^2}\right). \tag{15}$$

When we take $\tau_x = \tau_z = \infty$, i.e., noninformative prior for x_{is} and z_{ij}, $x_{is} \sim N(\sum_l x_{isl}/L_s, \sigma_\delta^2/L_s)$ and $z_{ij} \sim N(\sum_k z_{kij}/K, \sigma_z^2/K)$.

2. Missing value y_{kijsln}. Let $\mu_{kijsl} = x_{kisl} + z_{kij}$; then

$$f(y_{kijsln} \mid \ldots) \propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(y_{kijsln} - \mu_{kijsl})^2\right\} \times \frac{1}{1+\exp(a+by_{kijsln})}. \tag{16}$$

Note that $f(y_{kijsln} \mid \ldots)$ is log-concave. We can use Adaptive Rejection Sampling (ARS) method.

3. Parameters in the logistic model for missing mechanism: a and b. Since

$$f(a, b \mid \ldots) \propto \frac{\prod_{kijsln:I_{kijsln}=1} \exp(a+by_{kijsln})}{\prod_{kijsln}(1+\exp(a+by_{kijsln}))} \times N(a \mid 0, v^2) \times N(b \mid 0, v^2) \tag{17}$$

is log-concave, we can use ARS.

4. Variances σ_ε, σ_x, and σ_z.

$$\sigma_x^{-2} \sim \text{Gamma}\left(\gamma_1 + \frac{KI\sum_s L_s}{2}, \left[\frac{1}{\gamma_2} + \frac{1}{2}\sum_{kisl}(x_{kisl} - x_{mi})^2\right]^{-1}\right), \tag{18}$$

$$\sigma_z^{-2} \sim \text{Gamma}\left(\gamma_3 + \frac{K\sum_i J_i}{2}, \left[\frac{1}{\gamma_4} + \frac{1}{2}\sum_{kij}(z_{kij} - z_{ij})^2\right]^{-1}\right), \tag{19}$$

$$\sigma_\delta^{-2} \sim \text{Gamma}\left(\gamma_5 + \frac{I\sum_s L_s}{2}, \left[\frac{1}{\gamma_6} + \frac{1}{2}\sum_{isl}(x_{isl} - x_{is})^2\right]^{-1}\right), \tag{20}$$

$$\sigma_\varepsilon^{-2} \sim \text{Gamma}\left(\gamma_7 + \frac{\sum_{kijsl} N_{kijsl}}{2}, \left[\frac{1}{\gamma_8} + \frac{1}{2}\sum_{kijsln}(y_{kijsln} - x_{kisl} - z_{kij})^2\right]^{-1}\right). \tag{21}$$

Appendix B: iTRAQ Data from One Experiment

We illustrate the model when each sample is labeled differently, or we treat the samples with distinct isobaric tags as different samples. It is easy to modify this model to take the

replicates of samples into account. For the *m*th marker (or sample) and the *i*th protein, let y_{ijmn} denote the log value of the *n*th measured intensity for the *j*th peptide, and let x_{mi} denote the (log) protein expression level in the experiment. Let z_{ij} be the peptide effect for the *j*th peptide of the *i*th protein. We consider the additive model for y_{ijmn} ($m = 1, \ldots, M$; $i = 1, \ldots, I$; $j = 1, \ldots, J_i$; $n = 1, \ldots, N_{ijm}$) and missing mechanism:

$$y_{ijmn}=x_{mi}+z_{ij}+\varepsilon_{ijmn}, \tag{22}$$

$$\text{logit}\,(P(I_{ijmn}\mid y_{ijmn},a,b))=a+b\times y_{ijmn}, \tag{23}$$

and restrict $x_{i1} = 0$. We take normal distributions as priors for x_{mi} and z_{ij}:

$$x_{mi}\sim N(0,\tau_x^2)\quad \text{for } m>1, \tag{24}$$

$$z_{ij}\sim N(0,\tau_z^2). \tag{25}$$

Priors for other parameters are the same as those in Sect. 2. Then the joint distribution of the model is

$$\begin{aligned} f(y,I_y,x,z,a,b,\sigma_\varepsilon)=&\prod_{ijm}\{\text{MVN}(\mathbf{y}_{ijm}\mid(x_{mi}+z_{ij})\mathbf{1},\sigma_\varepsilon^2\mathbf{I})f(\mathbf{I}_{ijm}\mid\mathbf{y}_{ijm},a,b)\}\times\prod_{m>1,i}N(x_{mi}\mid 0,\tau_x^2)\times\prod_{ij}N(z_{ij}\mid 0,\tau_z^2)\times N(a\mid 0,\nu^2)\\ &\times N(b\mid 0,\nu^2)\\ &\times \text{invGamma}(\sigma_\varepsilon^2). \end{aligned} \tag{26}$$

The full conditional distributions for missing y_{ijmn}, a and b are the same as those in multiple experiments. For x_{mi}, z_{ij}, and σ_ε, their full conditional distributions are given below:

$$x_{mi}\sim N\left(\frac{\sum_j\frac{\bar{y}_{ijm}-z_{ij}}{\sigma_\varepsilon^2/N_{ijm}}}{\sum_j\frac{N_{ijm}}{\sigma_\varepsilon^2}+\frac{1}{\tau_x^2}},\frac{1}{\sum_j\frac{N_{ijm}}{\sigma_\varepsilon^2}+\frac{1}{\tau_x^2}}\right) \tag{27}$$

$$z_{ij}\sim N\left(\frac{\sum_m\frac{\bar{y}_{ijm}-x_{mi}}{\sigma_\varepsilon^2/N_{kijsl}}}{\sum_m\frac{N_{ijm}}{\sigma_\varepsilon^2}+\frac{1}{\tau_z^2}},\frac{1}{\sum_m\frac{N_{ijm}}{\sigma_\varepsilon^2}+\frac{1}{\tau_z^2}}\right) \tag{28}$$

$$\sigma_\varepsilon^{-2}\sim\text{Gamma}\left(\gamma_5+\frac{\sum_{ijm}N_{ijm}}{2},\frac{2+\gamma_6\sum_{ijmn}(y_{ijmn}-x_{mi}-z_{ij})^2}{2\gamma_6}\right). \tag{29}$$

References

1. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003; 19(2):185–193. [PubMed: 12538238]
2. Choe L, D'Ascenzo M, Relkin NR, Pappin D, Ross P, Williamson B, Guertin S, Pribil P, Lee KH. 8-plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease. Proteomics. 2007; 7:3651–3660. [PubMed: 17880003]
3. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol. 1999; 17:994–999. [PubMed: 10504701]
4. Hamdan M, Righetti PG. Modern strategies for protein quantification in proteome analysis: advantages and limitations. Mass Spectrom Rev. 2002; 21:287–302. [PubMed: 12533801]
5. Hill EG, Schwacke JH, Comte-Walters S, Slate EH, Oberg AL, Eckel-Passow JE, Therneau TM, Schey KL. A statistical model for iTRAQ data analysis. J Proteome Res. 2008; 7:3091–3101. [PubMed: 18578521]
6. Liu H, Sadygov RG, Yates JR. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem. 2004; 76:4193–4201. [PubMed: 15253663]
7. Marx J. Caveolae: a once-elusive structure gets some respect. Science. 2001; 294:1862–1865. [PubMed: 11729300]
8. Oberg A, Mahoney D, Eckel-Passow J, Malone C, Wolfinger R, Hill E, Cooper L, Onuma O, Spiro C, Therneau T, Bergen H. Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. J Proteome Res. 2008; 7:225–233. [PubMed: 18173221]
9. O'Farrell PH. High resolution two-dimensional electrophoresis of proteins. J Biol Chem. 1975; 250:4007–4012. [PubMed: 236308]
10. Patton WF. Detection technologies in proteome analysis. J Chromatogr B, Anal Technol Biomed Life Sci. 2002; 771:3–31.
11. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999; 20:3551–3567. [PubMed: 10612281]
12. Ross PL, Huang YN, Marchese JN, 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, Pappin DJ. Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics. 2004; 3:1154–1169. [PubMed: 15385600]
13. Salim K, Kehoe L, Minkoff MS, Bilsland JG, Munoz-Sanjuan I, Guest PC. Identification of differentiating neural progenitor cell markers using shotgun isobaric tagging mass spectrometry. Stem Cells Dev. 2006; 15:461–470. [PubMed: 16846381]
14. Seshi B. An integrated approach to mapping the proteome of the human bone marrow stromal cell. Proteomics. 2006; 6:5169–5182. [PubMed: 16947122]
15. Wang P, Tang H, Zhang H, Whiteaker J, Paulovich AG, Mcintosh M. Normalization regarding non-random missing values in high-throughput mass spectrometry data. Pac Symp Biocomput. 2006; 11:315–326. [PubMed: 17094249]
16. Wu WW, Wang G, Baek SJ, Shen R-F. Comparative study of three proteomic quantitative methods, DIGE, cICAT, and iTRAQ, using 2D Gel- or LC-MALDI TOF/TOF. J Proteome Res. 2006; 5:651–658. [PubMed: 16512681]

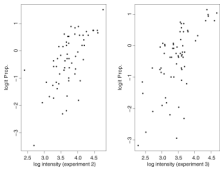

Fig. 1.

Relationship between the logit peptide missingness probability in one experiment and the observed peptide intensity in another experiment. *X*-axis: the log median intensity of 100 observed peptides binned based on their intensities in experiments 2 (*left panel*) and 3 (*right panel*). *Y*-axis: the logit proportion of the 100 peptides in each bin that are observed in the first experiment

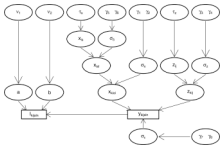

Fig. 2.
The hierarchical structure of our model. The unknown parameters are in *circles*, and the observations are in *rectangles*

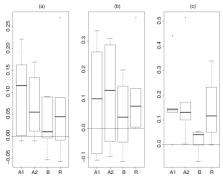

Fig. 3.
Boxplots of the biases of the estimates from Table 3 using different methods (A1: ANOVA 1; A2: ANOVA 2; B: Bayesian; R: log-ratio). **a** $\hat{x_2}$; **b** $\hat{x_3}$; **c** $\hat{x_4}$

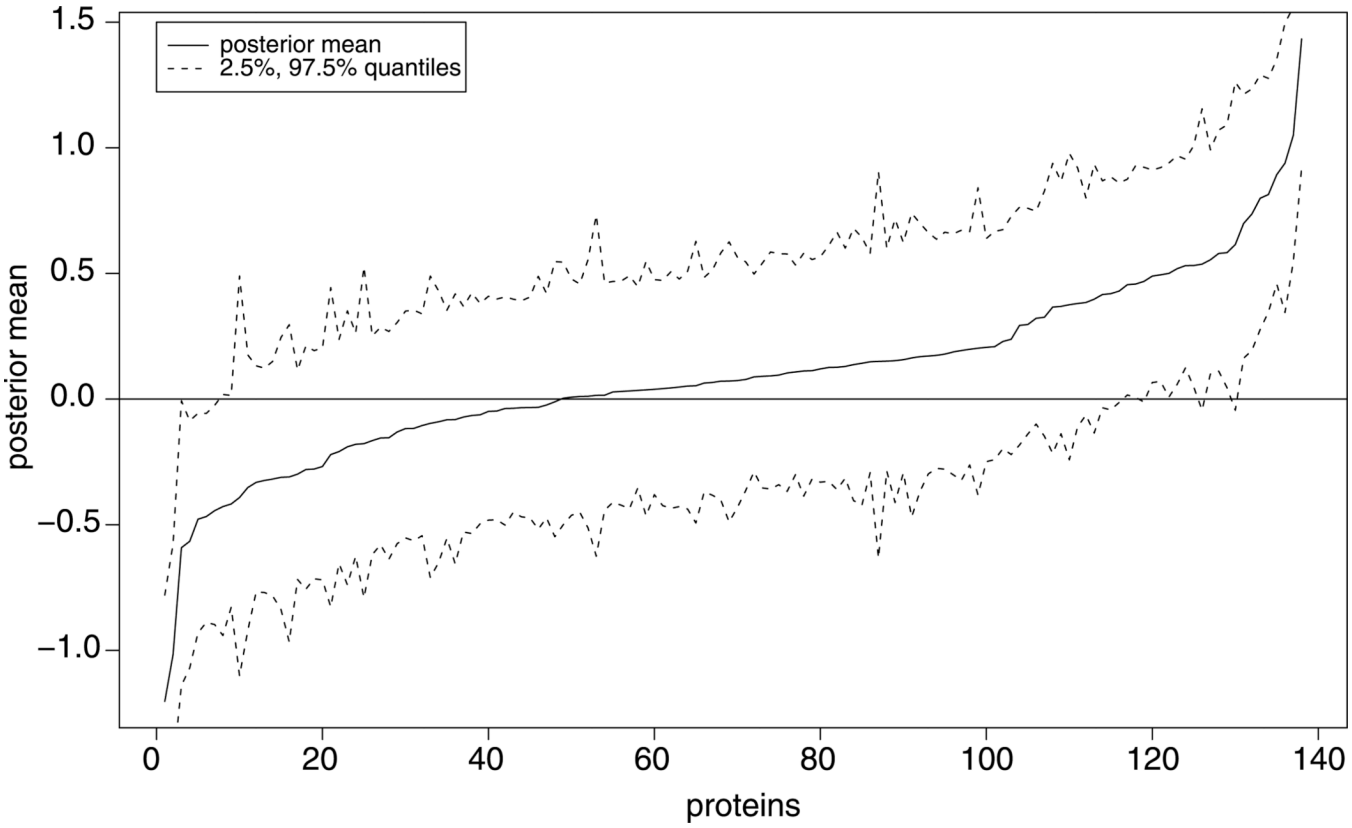

Fig. 4.
Posterior means of the proteins sorted in increasing order

Table 1

An example of the iTRAQ data when four isobaric tags are used. Each row corresponds to a specific peak in the mass spectrum. The first column gives the protein ID, the second column gives the peptide sequence that corresponds to this peak, and the last four columns give the peak areas for the four tags

Protein accessions	Peptide sequence	Area114	Area115	Area116	Area117
IPI00798592.1	ADVVESWIGEK	22.03	29.88	29.08	36.89
IPI00798592.1	ADVVESWIGEK	6.32	6.91	6.8	8.13
IPI00798592.1	ADVVESWIGEK	5.3	3.84	3.66	10.26
IPI00798592.1	DLASVQALLR	31.36	33.68	59.77	41.93
IPI00798592.1	DLASVQALLR	NA	NA	NA	NA
IPI00798592.1	DLASVQALLR	54.56	64.83	114.21	86.9
IPI00798592.1	DVDEIEAWISEK	NA	7.11	13.6	NA
IPI00798592.1	DVDETIGWIK	15.33	32.09	75.23	33.78
IPI00798592.1	…	…	…	…	…

Table 2

Missing pattern in Caveolin-1 study. 8045 is the total number of peptides observed in the study for the 138 common proteins. Only about 1/3 of proteins and 1/4 of peptides are observed in all three experiments

		Number of experiments protein/peptide is present		
	Counts	**1**	**2**	**3**
proteins	424	192 (45.3%)	94 (22.2%)	138 (32.5%)
peptides	8045	4765 (59.2%)	1156 (14.4%)	2124 (26.4%)

Table 3

Comparison of the inference results for the relative protein expression levels under different settings of variations. The true values are $x_2 = -0.04$, $x_3 = -0.48$, and $x_4 = -0.66$. "Bayesian" refers to our method. "ANOVA 1" considers the sample effect, peptide effect, experimental effect, and the interaction of sample and peptide effect. "ANOVA 2" removes the interaction term from the analysis of "ANOVA 1." "Log-ratio" calculates the means of peptide log-ratios

			Bayesian			**ANOVA 1**		
σ_x^2	σ_z^2	σ_ε^2	$\hat{x_2}$	$\hat{x_3}$	$\hat{x_4}$	$\hat{x_2}$	$\hat{x_3}$	$\hat{x_4}$
10^{-4}	10^{-4}	10^{-4}	−0.037 (−0.06, −0.01)	−0.480 (−0.50, −0.46)	−0.662 (−0.69, −0.64)	−0.051 (−0.053, −0.048)	−0.484 (−0.487, −0.481)	−0.660 (−0.663, −0.657)
10^{-4}	1	1	−0.044 (−0.23, 0.14)	−0.526 (−0.73, −0.33)	−0.613 (−0.80, −0.43)	0.100 (−0.13, 0.33)	−0.567 (−0.78, −0.35)	−0.529 (−0.75, −0.30)
10^{-2}	0.1	1.5	−0.094 (−0.28, 0.09)	−0.593 (−0.79, −0.40)	−0.729 (−0.93, −0.52)	0.052 (−0.12, 0.22)	−0.589 (−0.77, −0.41)	−0.516 (−0.71, −0.32)
10^{-2}	1	1.5	0.051 (−0.13, 0.23)	−0.279 (−0.46, −0.10)	−0.610 (−0.80, −0.42)	0.123 (−0.05, 0.30)	−0.218 (−0.40, −0.03)	−0.514 (−0.70, −0.33)
10^{-2}	10	1.5	−0.023 (−0.24, 0.19)	−0.336 (−0.55, −0.12)	−0.608 (−0.83, −0.39)	−0.039 (−0.31, 0.23)	−0.272 (−0.54, −0.01)	−0.218 (−0.52, 0.08)
10^{-2}	50	1.5	0.050 (−0.15, 0.25)	−0.403 (−0.60, −0.20)	−0.626 (−0.83, −0.42)	0.182 (−0.07, 0.43)	−0.146 (−0.40, 0.11)	−0.520 (−0.77, −0.27)
			ANOVA 2			Log-ratio		
10^{-4}	10^{-4}	10^{-4}	−0.051 (−0.053, −0.049)	−0.484 (−0.486, −0.482)	−0.658 (−0.661, −0.656)	−0.036	−0.480	−0.662
10^{-4}	1	1	0.005 (−0.17, 0.18)	−0.575 (−0.75, −0.40)	−0.560 (−0.74, −0.38)	−0.098	−0.475	−0.533
10^{-2}	0.1	1.5	0.025 (−0.14, 0.19)	−0.525 (−0.70, −0.35)	−0.506 (−0.68, −0.33)	0.046	−0.429	−0.427
10^{-2}	1	1.5	0.094 (−0.08, 0.27)	−0.216 (−0.39, −0.04)	−0.488 (−0.67, −0.31)	−0.049	−0.343	−0.555
10^{-2}	10	1.5	−0.028 (−0.28, 0.22)	−0.194 (−0.45, 0.06)	−0.144 (−0.41, 0.12)	0.232	−0.100	−0.321
10^{-2}	50	1.5	0.130 (−0.10, 0.36)	−0.172 (−0.41, 0.07)	−0.553 (−0.79, −0.32)	0.048	−0.379	−0.610

Table 4

Sensitivity of inference to a different missing mechanism and random number of observations. The true values are $x_2 = -0.04$, $x_3 = -0.48$, and $x_4 = -0.66$. Values in parentheses are standard deviations of the estimates. The ANOVA model involves the sample effect, peptide effect, and experimental effect

	Bayesian			**ANOVA 1**		
	$\hat{x_2}$	$\hat{x_3}$	$\hat{x_4}$	$\hat{x_2}$	$\hat{x_3}$	$\hat{x_4}$
Bi(3,0.9)	−0.027 (0.064)	−0.486 (0.069)	−0.666 (0.091)	−0.058 (0.150)	−0.442 (0.073)	−0.587 (0.181)
Bi(3,0.2)	−0.055 (0.056)	−0.500 (0.069)	−0.651 (0.048)	−0.054 (0.086)	−0.512 (0.079)	−0.630 (0.033)
Poisson	−0.036 (0.101)	−0.473 (0.136)	−0.634 (0.131)	−0.036 (0.108)	−0.469 (0.182)	−0.619 (0.212)

Table 5

Experimental design of the caveolin-1 study

	tag			
Experimental run order	**114**	**115**	**116**	**117**
1	WT	WT	KO	KO
2	WT	WT	KO	KO
3	WT	WT	KO	KO

Table 6

Posterior means and standard deviations of other parameters

	σ_ε^2	σ_x^2	σ_δ^2	σ_z^2	a	b
mean	1.351	0.141	0.002	313.141	−0.442	−0.217
sd	0.004	0.008	0.002	23.160	0.021	0.005