research articles Journal of proteome research

Protein Quantification in Label-Free LC-MS Experiments

Timothy Clough,[†,#] Melissa Key,[†,#] Ilka Ott,[‡] Susanne Ragg,[§] Gunther Schadow,[||] and Olga Vitek*[,†,⊥]

Department of Statistics, Purdue University, West Lafayette, Indiana 47907, Department of Medicine, Technical University Munich, Germany, School of Medicine, Indiana University, Indianapolis, Indiana 46202, School of Informatics, Indiana University, Indianapolis, Indiana 46202, and Department of Computer Science, Purdue University, West Lafayette, Indiana 47907

Received July 11, 2009

The goal of many LC-MS proteomic investigations is to quantify and compare the abundance of proteins in complex biological mixtures. However, the output of an LC-MS experiment is not a list of proteins, but a list of quantified spectral features. To make protein-level conclusions, researchers typically apply *ad hoc* rules, or take an average of feature abundance to obtain a single protein-level quantity for each sample. We argue that these two approaches are inadequate. We discuss two statistical models, namely, fixed and mixed effects Analysis of Variance (ANOVA), which views individual features as replicate measurements of a protein's abundance, and explicitly account for this redundancy. We demonstrate, using a spike-in and a clinical data set, that the proposed models improve the sensitivity and specificity of testing, improve the accuracy of patient-specific protein quantifications, and are more robust in the presence of missing data.

Keywords: Quantitative proteomics • Protein quantification • LC-MS • Analysis of Variance • Mixed models • Missing data

1. Introduction

Liquid chromatography coupled with mass spectrometry (LC-MS) is a method of choice for identification and quantification of proteins and peptides in complex biological mixtures.[1] A widespread use of LC-MS is within the label-free shotgun ("bottom up") workflow, which enables nontargeted analysis for discovery-oriented research. With this workflow, proteins in a sample are enzymatically digested into peptides, and subjected to chromatographic separation, ionization, and mass analysis. The intensity of the resulting LC-MS features is used for relative quantification of peptides and proteins. Tandem mass spectra are typically acquired simultaneously, and are used to determine the sequence identity of peptides and proteins underlying a subset of the features. A variety of software tools have been proposed to extract and quantify LC-MS features from the acquired spectra, annotate the features with sequence identity, and align the features across runs.[2] Overall, the workflow is now capable of accurately identifying and quantifying thousands of peptides simultaneously.

Information from LC-MS experiments is subsequently used as input to statistical and machine learning analysis steps, which either compare protein abundance between sample groups (e.g., test proteins for differential abundance between patients with a disease and healthy controls), or analyze protein abundance of individual biological subjects (e.g., perform unsupervised clustering or supervised classification of individual patients, based on their quantitative protein profiles). In most cases, a desirable unit of analysis is a protein. However, the currency of LC-MS experiments are spectral features that correspond to peptide ions, and multiple such features can be observed for a protein of moderate and high abundance. Although features from the same protein are expected to present a consistent quantitative profile across all samples and runs, random variation and experimental and biological artifacts can distort the profiles of some features, and deriving protein-level conclusions in these cases is not straightforward.

As a partial solution to this problem, researchers frequently continue statistical and machine learning analysis on the per-feature basis.[3] With this approach, group comparison requires an *ad hoc* decision rule, for example, a protein is considered differentially abundant if a predefined number of features has statistically significant changes in the same direction. For protein quantification of individual biological subjects, such as patients, the per-feature analysis provides no clear solution.

An alternative approach is to summarize intensities of all features of the protein using a quantitative summary, for example, by averaging feature intensities on the log scale.[4,5] This yields a single number per protein per run that can be compared across groups of subjects, or used to quantify the protein abundance in individual subjects.

Here we argue that, from the statistical methodology perspective, both per-feature and averaging approaches are inadequate for protein quantification. We examine the statistical

* To whom correspondence should be addressed at 250 N. University Street, West Lafayette, IN 47907; e-mail, ovitek@stat.purdue.edu.
[†] Department of Statistics, Purdue University.
[#] These authors share equal contribution.
[‡] Technical University Munich.
[§] School of Medicine, Indiana University.
[||] School of Informatics, Indiana University.
[⊥] Department of Computer Science, Purdue University.

properties of these methods and note that they suffer from multiple drawbacks. We further note that the averaging approach is a special instance of an Analysis of Variance (ANOVA) model.[6,7] This approach relies upon a series of very specific assumptions which are implicitly made, but rarely explicitly verified. We verify the plausibility of the assumptions, and extend the approach with a more general family of ANOVA models where some of the assumptions are relaxed. We show that instances of this family of models improve the sensitivity and the specificity of finding true changes in abundance, and have a more accurate subject quantification.

In the following, we focus on comparative experiments where distinct individuals are selected in each group. We do not consider experiments which involve repeated measurements, that is, experiments where groups correspond to time points and samples are collected from the same set of individuals across time. We further assume that the input LC-MS features are reliably identified and quantified, and are unambiguously grouped into proteins. In other words, we omit from consideration LC-MS features with no reliable identification, and peptide sequences with ambiguous protein memberships. We also assume that feature abundance values have been appropriately normalized across runs.

We illustrate our discussion using two data sets: one a controlled spike-in data set[2] where we evaluate the methods by their ability to detect true fold changes, and the other a clinical data set of patients with cardiovascular disease,[8] where the methods are evaluated by comparing LC-MS quantifications to clinical assay measurements on the same proteins and patients. Throughout the text, we give examples from clinical research; however, the discussion can be applied to any label-free quantitative proteomic experiment.

2. Methods

2.1. Statistical Analysis of LC-MS Data: Goals and Procedures. Many proteomic investigations involve large groups (also called *populations*) of distinct biological individuals, such as patients, animals, or plants. While it is impossible to collect samples from each individual in the population, a good experimental design includes samples from multiple individuals selected from the populations. The individuals, which are also called *subjects* or *biological replicates* in statistical literature, help assess the natural biological variation of protein abundance, and distinguish the natural variation from systematic deviations that are due to treatments or stresses.[9] Experimental designs can also specify replicate mass spectrometry runs on the same biological sample. The runs, called *technical replicates*, help assess the additional variation introduced into the measurements by the experimental procedure.

A typical goal of group comparison is to determine whether the average protein abundance is different between two or more populations of interest, for example, between diseased patients and healthy controls. A statistical model is necessary to characterize the biological and the experimental variation, and to expand the scope of conclusions from the selected individuals to the entire populations. A summary quantity (called test statistic) is derived based on the model, and the strength of evidence in favor of differential abundance is quantified by a *p*-value. Differential abundance is concluded when the *p*-value of the comparison is below a predefined cutoff. When tests are performed for multiple proteins in parallel, of interest is the False Discovery Rate (FDR), defined as the expected proportion of falsely detected changes in the list of differentially abundant proteins. A multiple comparisons procedure, such as the one proposed by Benjamini and Hochberg,[10] adjusts the *p*-values to control for the FDR. A list of proteins with the adjusted *p*-values below a cutoff has the FDR of at most the cutoff.

When the true status of the proteins is known, approaches to group comparison can be compared, for a fixed FDR, according to their sensitivity (or statistical power, i.e., the ability to detect true changes in abundance) and specificity (i.e., the number of false positive discoveries). In addition, the models can be evaluated with respect to bias and variance of estimates of fold changes of protein abundance between groups. Bias is a systematic deviation of an estimated quantity from the true value.[9] When quantifying proteins from mass spectra, bias can be introduced by features with missing intensities, which tend to appear more frequently in features with low signal. Variance refers to the uncertainty in the estimated abundance. A preferred summarization procedure is the one that avoids the bias and reduces the variance.

In addition to comparing average protein abundance in the underlying populations, it is often of interest to separately quantify protein abundance for each individual (equivalently, for each subject or biological replicate) in the study from multiple features, and on a continuous scale that is comparable across runs. We call this procedure *subject quantification*. Results of subject quantification are used as input for subsequent analysis steps such as clustering or classification. The importance of statistical modeling for accurate subject quantification is often overlooked. However, a statistical model helps viewing the procedure as an estimation of the "true" abundance of the protein in the individual from noisy measurements, and appropriately deriving the summary. The performance of model-based subject quantification can be evaluated, for example, by comparing it to independent external measurements on the same protein.

True biological signals are widely viewed as multiplicative in nature; therefore, in the following, we consider abundance of the LC-MS features on the logarithmic scale. Application of the logarithm has multiple advantages as compared to working with ratios of abundance. In particular, it provides a natural way for modeling the replicate structure of the data within the Analysis of Variance (ANOVA) framework. The use of ANOVA models in proteomic applications has recently been introduced.[9,11−14] Here, we systematically present and contrast several such models, and compare their properties for group testing and subject quantification.

In the following, we assume that the experiment consists of g disease groups with n distinct individuals in each group. Each protein is represented by f peptide features. We will also assume for simplicity that the data set has no technical replicates (i.e., each run corresponds to a different individual); however, extensions to experiments with technical replicates are straightforward. When a data set contains an equal number of subjects in each disease group, and when peptide features have no missing observations, the data set is called *balanced*. The balance is desirable because it optimizes the precision of our conclusions as compared to the unbalanced case, and many quantities of interest can be estimated by averages. Table 1 summaries the formulas used for group comparison and subject quantification for the models that we consider in the balanced case, and extensive additional information on these models is provided in Supporting Information.

Table 1. Model-Based Quantities Used for Group Comparison and Subject Quantification[a]

Approach	Extension	Group Testing $H_0: \mu_i = \mu_{i'}$ **vs** $H_a: \mu_i \neq \mu_{i'}$	Subject Quantification
Per-feature	basic	$\frac{\bar{y}_{ij\cdot} - \bar{y}_{i'j\cdot}}{\sqrt{\frac{2}{n}\hat{\sigma}_j^2}} \sim Student_{g(n-1)}$	none
	Empirical Bayes	$\frac{\bar{y}_{ij\cdot} - \bar{y}_{i'j\cdot}}{\sqrt{\frac{2}{n}\frac{g(n-1)\hat{\sigma}_j^2 + d_0\hat{\sigma}_0^2}{g(n-1)+d_0}}} \sim Student_{g(n-1)+d_0}$	none
"Average"	basic	$\frac{\bar{y}_{i\cdot\cdot} - \bar{y}_{i'\cdot\cdot}}{\sqrt{\frac{2}{n}\hat{\sigma}^2}} \sim Student_{g(n-1)}$	$\bar{y}_{i\cdot k}$
	median polish	$\frac{\tilde{y}_{i\cdot\cdot} - \tilde{y}_{i'\cdot\cdot}}{\sqrt{\frac{2}{n}\hat{\sigma}^2}} \sim Student_{g(n-1)}$	$\tilde{y}_{i\cdot k}$
	Empirical Bayes	$\frac{\bar{y}_{i\cdot\cdot} - \bar{y}_{i'\cdot\cdot}}{\sqrt{\frac{2}{n}\frac{g(n-1)\hat{\sigma}^2 + d_0\hat{\sigma}_0^2}{g(n-1)+d_0}}} \sim Student_{g(n-1)+d_0}$	$\bar{y}_{i\cdot k}$
Fixed	basic	$\frac{\bar{y}_{i\cdot\cdot} - \bar{y}_{i'\cdot\cdot}}{\sqrt{\frac{2}{nf}\hat{\sigma}^2}} \sim Student_{g(f-1)(n-1)}$	$\bar{y}_{i\cdot k}$
	feature-specific variances	$\frac{\frac{1}{f}\sum_j \frac{1}{\hat{\sigma}_j^2}y_{ij\cdot} - \frac{1}{f}\sum_j \frac{1}{\hat{\sigma}_j^2}y_{i'j\cdot}}{\sqrt{\frac{2}{nf}(\frac{1}{f}\sum_j \hat{\sigma}_j^2)}} \sim Student_{g(f-1)(n-1)}$	$\frac{1}{f}\sum_j \frac{1}{\hat{\sigma}_j^2}y_{ijk}$
	Empirical Bayes	$\frac{\bar{y}_{i\cdot\cdot} - \bar{y}_{i'\cdot\cdot}}{\sqrt{\frac{2}{nf}\frac{g(f-1)(n-1)\hat{\sigma}^2 + d_0\hat{\sigma}_0^2}{g(f-1)(n-1)+d_0}}} \sim Student_{g(f-1)(n-1)+d_0}$	$\bar{y}_{i\cdot k}$
Mixed	basic	$\frac{\bar{y}_{i\cdot\cdot} - \bar{y}_{i'\cdot\cdot}}{\sqrt{2\{\frac{1}{nf}\hat{\sigma}^2 + \frac{1}{n}\hat{\sigma}_S^2\}}} \sim Student_{g(n-1)}$	$\bar{y}_{i\cdot k}$
	feature-specific variances	$\frac{\frac{1}{f}\sum_j \frac{1}{\hat{\sigma}_j^2}y_{ij\cdot} - \frac{1}{f}\sum_j \frac{1}{\hat{\sigma}_j^2}y_{i'j\cdot}}{\sqrt{2\{\frac{1}{nf}(\frac{1}{f}\sum_j \hat{\sigma}_j^2) + \frac{1}{n}\hat{\sigma}_S^2\}}} \sim Student_{g(n-1)}$	$\frac{1}{f}\sum_j \frac{1}{\hat{\sigma}_j^2}y_{ijk}$
	Empirical Bayes	$\frac{\bar{y}_{i\cdot\cdot} - \bar{y}_{i'\cdot\cdot}}{\sqrt{\frac{2}{nf}\frac{g(n-1)(f\hat{\sigma}_S^2 + \hat{\sigma}^2) + d_0\hat{\sigma}_0^2}{g(n-1)+d_0}}} \sim Student_{g(n-1)+d_0}$	$\bar{y}_{i\cdot k}$

[a] The quantities are derived for the case of a balanced design, i.e., for situations where each group has an equal number of subjects, and peptide features have no missing observations. See Table 2 for notation.

Table 2. Descriptions of Basic Quantities Used in Table 1[a]

notation	description	approach used in
$\bar{y}_{ij\cdot}$	average intensity of feature j in group i	per-feature
$\bar{y}_{i\cdot\cdot}$	average intensity of group i	"average", fixed, and mixed
$\tilde{y}_{i\cdot\cdot}$	median polish-based average intensity of group i	median polish extension to "average"
$\hat{\sigma}_j^2$	estimated variance of measurement error for feature j	per-feature; feature-specific variance extensions
$\hat{\sigma}^2$	estimated variance of measurement error	"average", fixed, and mixed
$\hat{\sigma}_S^2$	estimated between-subject variance	mixed
$\hat{\sigma}_0^2$	prior estimate of measurement error variance	Empirical Bayes in per-feature, "average", and fixed models
$\tilde{\sigma}_0^2$	prior estimate of overall variance ($f\sigma_S^2 + \sigma^2$) in mixed model	Empirical Bayes in mixed model
d_0	degrees of freedom associated with prior estimate of variance	Empirical Bayes models

[a] Where $i = 1, ..., g$ is a group, $j = 1, ..., f$ is a feature, and $k = 1, ..., n$ is a subject.

Observed feature intensity	=	Overall feature mean	+	Systematic deviation due to disease group	+	Random deviation due to non-systematic sources of variation
y_{ijk}	=	μ_j	+	G_{ij} $\sum_{i=1}^{g} G_{ij} = 0$	+	ϵ_{ijk} $\sim N\left(0, \sigma_j^2\right)$

Figure 1. Per-feature ANOVA. i is the index of a disease group, j the index of a feature, and k the index of a subject. $\sum_{i=1}^{g} G_{ij} = 0$ is an identifiability constraint that is required for estimation of the model parameters. σ_j^2 is the variance of the measurement error associated with feature j. All random deviations are independent.

2.2. Per-Feature Model. 2.2.1. Basic Model. The statistical model underlying this approach is shown in Figure 1. The model considers separately each feature, and decomposes its abundance across runs into a systematic contribution G_{ij} of disease group i to feature j, and a random contribution of biological and technical variation ε_{ijk} which is normally distributed with feature-specific variance σ_j^2. As such, the model is a multigroup generalization of the model underlying a two-group t test. The approach has a practical advantage in that it is simple, and a technical advantage in that it specifies a separate variance of the error for each feature. The latter is particularly appropriate for LC-MS data, where variation typically depends upon intensity, and differs across features.

The main drawback of using this model for proteins is the lack of formal protein inference for both group testing and subject quantification. For testing, an *ad hoc* rule requires a predefined number of differentially abundant features to claim the differential abundance of the protein as a whole, and this

Average feature intensity per subject	=	Overall feature mean	+	Systematic deviation due to disease group	+	Random deviation due to non-systematic sources of variation
$\bar{y}_{i\cdot k}$	=	μ	+	G_i $\sum_{i=1}^{g} G_i = 0$	+	ϵ_{ik} $\sim N\left(0, \sigma^2\right)$

Figure 2. Average ANOVA. *i* is the index of a disease group, and *k* the index of a subject. $\sum_{i=1}^{g} G_i = 0$ is an identifiability constraint that is required for estimation of the model parameters. σ^2 is the variance of the measurement error. All random deviations are independent.

is undesirable since the statistical properties of the overall procedure, such as the FDR in the resulting protein list, are unknown. There is no clear subject-level protein quantification.

In addition, the approach suffers from technical drawbacks. First, the model for each feature is based on a smaller number of data points than a model which combines the intensities from all features. Second, the per-feature approach results in an overall larger number of tests, and therefore, requires a stronger adjustment for multiple testing than when working with proteins. As we show in section 4, these two facts undermine the sensitivity and specificity of detecting differences in abundance at the level of features. Finally, peptide abundances within a protein are strongly correlated. The correlation undermines the calculation of the FDR, and the error rate in the list of differentially abundant features can deviate substantially from the stated level.

2.2.2. Extensions. When the number of individuals per group is small, estimates of error variances σ_j^2 in Figure 1 can be unreliable. This can be remedied somewhat by means of an Empirical Bayes procedure proposed in the context of gene expression microarrays.[3,15] In addition to the model in Figure 1, we assume that $(1/\sigma_j^2) \sim [(1/d_0 s_0^2)\chi_{d_0}^2]$, where s_0^2 and d_0 are constants that we estimate empirically from the entire collection of peptide features in the data set. In other words, the joint analysis of peptide features provides additional information on the variance, which is equivalent to a prior data set with estimated variance s_0^2 based on d_0 degrees of freedom.

The Empirical Bayes approach provides a more accurate estimation of feature-specific variation, modifies the testing procedure as shown in Table 1, and has been shown to improve the sensitivity and the specificity of the individual tests. However, this extension does not address the other drawbacks of the per-feature model. The Empirical Bayes procedure is applicable to the general class of linear models, and in the following sections we apply it to all the models that we consider.

2.3. Averaging of Intensities. 2.3.1. Basic Model. An alternative to the per-feature analysis is to average all feature intensities of a protein within a run, and use this summary for the subsequent analyses. The model used in this case, that we call "average model", decomposes the summaries into the systematic contribution G_i of disease group *i*, and a random deviation ε_{ik} which is normally distributed with variance σ^2 as shown in Figure 2. The resulting model is quite simple. Since only one model is fit for each protein, it provides protein-level conclusions of testing. Averaging of feature intensities in each run yields an obvious subject quantification. The model also produces a smaller number of tests and requires a less stringent adjustment for multiple comparisons.

However, the approach has several disadvantages. First, averaging reduces the nominal number of data points, and loses the information from individual features. Second, the "average model" treats signals from each peptide feature equally, and does not account for inconsistencies or uncertainties in the profiles of some features. Finally, averaging does not appropriately account for missing data or outliers. As we show in section 4, these factors can result in a reduction of sensitivity and specificity of group comparisons, and can produce biased subject quantification.

2.3.2. Extensions. The Empirical Bayes approach can be used to extend the "average model" similarly as in section 2.2. Another possible extension is based on Tukey median polish,[16] a robust alternative to averaging used, for example, by the RMA normalization procedure for gene expression microarrays.[17] Briefly, the procedure can be viewed as fitting an ANOVA model to a two-way table, where rows denote runs and columns denote peptide features. The approach proceeds by iteratively subtracting the medians of the rows or columns from feature intensities, until the median values are close to zero. The fitted intensities are then obtained by adding the remaining values in the table to the original intensities. The row averages of the fitted intensities are subject quantifications, and are used as abundance measures in Figure 2. The approach produces subject quantification which is more robust to outliers than a simple average; however, it does not address the other drawbacks of the averaging.

2.4. Fixed Effects Analysis of Variance. 2.4.1. Basic Model. The fixed effects Analysis of Variance (ANOVA) model,[6,7] shown in Figure 3, views features mapped to the same protein as replicate measurements of protein abundance, and explicitly describes the structure of this replication. Intensity of each peak is decomposed into the contribution of disease group G_i (expressing the systematic difference in protein abundance between groups), and the contribution of the LC-MS feature F_j that produced this value (expressing the fact that some peptides have a systematically higher ionization efficiency than the others). The model also contains a statistical interaction term $(G \times F)_{ij}$, which describes potential deviations of individual LC-MS features from the average profile, which can be due to both biological interferences and experimental artifacts. In cases where feature profiles are perfectly consistent, the interaction term is not necessary and can be dropped.

The model also specifies the deviation of each individual (i.e., biological replicate, or subject) $S(G)_{k(i)}$ from the overall group mean, and expresses the fact that some individuals have a higher natural abundance of the protein than others. Notation $S(G)_{k(i)}$ is read as "subject within a group", and reflects the property of comparative experiments where each individual can only belong to a single group. A different model, and a different notation, is necessary, for example, in experiments with repeated measurements where groups correspond to time points, and samples from the same individual are collected across time. Such situation is beyond the scope of this discussion. The last term in the model ε_{ijk} describes the

			Systematic deviation due to						
Observed intensity	=	Overall mean	+ disease group	+ feature	+ interaction	+ subject	+	Random deviation	
y_{ijk}	=	μ	+ G_i	+ F_j	+ $(G\times F)_{ij}$	+ $S(G)_{k(i)}$	+	ϵ_{ijk} $\sim N\left(0,\sigma^2\right)$	

$$\sum_{i=1}^{g} G_i = \sum_{j=1}^{f} F_j = \sum_{i=1}^{g} (G\times F)_{ij} = \sum_{j=1}^{f} (G\times F)_{ij} = \sum_{k=1}^{n} S(G)_{k(i)} = 0$$

Figure 3. Fixed effects ANOVA. *i* is the index of a disease group, *j* the index of a feature, and *k* the index of a subject. $\sum_{i=1}^{g} G_i = \sum_{j=1}^{f} F_j = \sum_{i=1}^{g} (G \times F)_{ij} = \sum_{j=1}^{f} (G \times F)_{ij} = 0$ and $\sum_{k=1}^{n} S(G)_{k(i)} = 0$ are identifiability constraints that are required for estimation of the model parameters. σ^2 is the variance of the measurement error. All random deviations are independent.

			Systematic deviation due to			Random deviation due to		
Observed intensity	=	Overall mean	+ disease group	+ feature	+ interaction	+ subject	+	error
y_{ijk}	=	μ	+ G_i	+ F_j	+ $(G\times F)_{ij}$	+ $S(G)_{k(i)}$ $\sim N\left(0,\sigma_S^2\right)$	+	ϵ_{ijk} $\sim N\left(0,\sigma^2\right)$

$$\sum_{i=1}^{g} G_i = \sum_{j=1}^{f} F_j = \sum_{i=1}^{g} (G\times F)_{ij} = \sum_{j=1}^{f} (G\times F)_{ij} = 0$$

Figure 4. Mixed effects ANOVA. *i* is the index of a disease group, *j* the index of a feature, and *k* the index of a subject. $\sum_{i=1}^{g} G_i = \sum_{j=1}^{f} F_j = \sum_{i=1}^{g} (G \times F)_{ij} = \sum_{j=1}^{f} (G \times F)_{ij} = 0$ are the identifiability constraints that are required for estimation of the model paramaters. σ_S^2 is the between-subjects variance, and σ^2 is the variance of the measurement error. All random deviations are independent.

remaining deviations of peak intensity, and is viewed as nonsystematic replicates of a normally distributed measurement error with constant variance σ^2.

The terms in the model are estimated from the data using procedures such as least-squares or maximum likelihood,[6,7] and results of the estimation can be used for both group testing and subject quantification. Testing in this model involves comparing the estimated abundances of groups $\mu + G_i$, and making a decision on whether the differences exceed the measurement error. Subject quantification amounts to reporting the estimated contributions $\mu + G_i + S(G)_{k(i)}$. In the case of a balanced design, these model-based quantities are based on sample averages, and are shown in Table 1. When a balanced experiment cannot be achieved, the model-based summaries differ from averages, and have been shown to provide a closer to optimal solution. Overall, the joint modeling of all peaks of the protein using ANOVA makes the best use of the available data. We show in section 4 that it increases the sensitivity and the specificity of testing, and improves the accuracy of subject quantification.

The "fixed effects" in the model name refers to the fact that the individuals selected for the study are considered fixed, and the model limits the scope of our conclusions to these specific individuals. For example, for testing, this implies that we are interested in differences in protein abundances in the specific individuals selected for experiment, and the only randomness associated with the measurements is the experimental noise. While this type of inference may be appropriate for an initial screening, it is inappropriate for a validation experiment where we would like to expand the scope of the conclusions to a larger population of patients, and individuals in the study are viewed as random instances from this larger population. This is further discussed in section 2.5.

2.4.2. Extensions. The main disadvantage of the model is that it assumes a common variance of error for all features, and this is not always realistic for LC-MS data. We can specify an alternative version of the model where we replace σ^2 in Figure 3 with feature-specific variances σ_j^2. In this case, the contribution of each feature to the model-based quantities is inversely proportional to its variation. As before, the model in Figure 3 can also be extended with an Empirical Bayes approach (Table 1).

2.5. Mixed Effects Analysis of Variance. Figure 4 displays an alternative mixed effects ANOVA model. In contrast to Figure 3, the model views the individuals in the study as a random selection from larger underlying populations. Here the goal of testing is not to compare means of protein abundance between the individuals selected for the study, but to compare the average abundances in the entire underlying populations. The model reflects the change in the scope of our conclusions by expressing the contributions of individuals $S(G)_{k(i)}$ as random nonsystematic quantities, which have a normal distribution with variance σ_s^2, and are independent from the measurement error. The name "mixed effects" emphasizes the fact that the model contains both fixed and random terms in addition to the measurement error.

As in the case of fixed effects, testing involves comparing the estimated abundances of groups $\mu + G_i$, and subject quantification involves reporting the estimated contributions $\mu + G_i + S(G)_{k(i)}$. However, the introduction of the random term has multiple implications. First, it implies that the variance of peak intensities is now assumed to be a combination of σ_s^2 and σ^2. Second, it implies that peaks intensities from the same protein and the same individual are assumed to be correlated,[6,7] that is, their abundance tends to change from subject to subject stochastically, but in the same direction. The structure of the correlation is shown in Supporting Information. In the special case of balanced data sets, the terms of the model can be estimated from the data using sample averages summarized in Table 1. We show in Supporting Information that in the

Table 3. The Latin Square Design Used to Create the Spike-in Data Set[a]

protein	mixture 1	2	3	4	5	6
Myoglobin (horse)	800	25	50	100	200	400
Carbonic anhydrase (bovine)	400	800	25	50	100	200
Cytochrome C (horse)	200	400	800	25	50	100
Lysozyme (chicken)	100	200	400	800	25	50
Alcohol dehydrogenase (yeast)	50	100	200	400	800	25
Adolase A (rabbit)	25	50	100	200	400	800

[a] Each entry is the amount of injected protein per sample (fmol).

balanced case the approach is equivalent to the "average model" for the purpose of group testing. When a balanced experiment cannot be achieved, the model often requires more sophisticated estimation procedures such as Restricted Maximum Likelihood (REML). Similarly to the fixed effects ANOVA, the model-based quantities differ from the averages. We illustrate in section 4 that mixed effects ANOVA outperforms the "average model" in unbalanced cases.

Overall, extending the scope of conclusions to a larger group of individuals affects decisions of differential abundance and subject quantification. The model frequently yields more conservative decisions of differential abundance, and has a lower sensitivity as compared to its fixed effects counterpart.

2.5.1. Extensions. As in the case of the fixed effects, the model can be extended to incorporate feature-specific variances, replacing σ^2 with σ_f^2 in Figure 4. Furthermore, the approach can also be extended with the Empirical Bayes approach, as shown in Table 1.

3. Data Sets

We evaluate the proposed models using two data sets. The first is a controlled mixture, where 6 proteins were spiked into human serum in known concentrations according to the latin square design in Table 3. Each mixture was acquired in triplicates on the Thermo Electron Fourier transformed-LTQ mass spectrometer, and LC-MS features were quantified, aligned, and annotated with peptide and protein identities using the Superhirn software.[2] Between 2 and 95 peptide features were reliably identified for each spiked protein; however, some features contained up to 33% of missing values.

We use the data set to evaluate protein quantification approaches in terms of sensitivity (i.e., the ability to detect true fold changes) and specificity (i.e., the ability to prevent discovery of false changes). The layout of the latin square design enables us to make this evaluation for multiple fold changes, starting from multiple abundance baselines, and for several proteins and runs, thereby averaging out potential protein- or run-specific artifacts.

The second data set came from a clinical investigation of 246 patients with cardiovascular disease, comparing patients with acute coronary syndrome (ST segment elevation myocardial infarction and non-ST segment myocardial infarction) to patients with unstable angina, stable angina, and control patients without coronary artery disease. Albumin was depleted prior to tryptic digest, and each sample was analyzed with a single replicate using Thermo-Finnigan linear ion-trap mass spectrometry. LC-MS features were quantified, aligned, and annotated with peptide and protein identities as described previously.[18] When a feature in a run could not be detected, the procedure reported the background signal, and therefore, the data set contains no missing values. To reduce the number of peptides with ambiguous mappings to protein isoforms, each peptide was mapped to the underlying ENTREZ GeneID. Only peptides with unambiguous gene IDs, and genes with at least two peptides, were retained for the subsequent analysis. Overall, the procedure identified 77 protein groups, with 2−169 peptides per group. For protein groups with over 74 peptides, 74 peptides were randomly selected for further analysis. In addition to the LC-MS data, 11 of the 77 proteins were measured on the BN Pro Spec Nephelometer (Siemens Healthcare Diagnostics).

We evaluate the sensitivity of group testing on this data set by comparing the number of differentially abundant proteins in all pairwise comparisons of disease groups for a fixed FDR. Furthermore, we evaluate subject-level protein quantification by examining the correlation of protein abundance estimated for the patients by each model with the nephelometry measurements.

4. Results

4.1. Model Fitting. Figure 5 shows experimental measurements for an example protein, for both the spike-in and clinical data sets. For the spike-in data set, the fold changes in concentrations of the spiked proteins are fairly large, and therefore, changes in feature intensities are clearly visible, and are as expected. However, missing intensities tend to appear in mixtures where the protein was spiked at lower concentrations, and for features with a lower overall signal. For the clinical example, the fold changes between disease groups are more subtle, and some features produce a somewhat contradictory evidence of differential abundance.

For each data set, we fit the models discussed in section 2, jointly to all groups and separately for each protein, using the Restricted Maximum Likelihood (REML) method implemented in the mixed procedure of the SAS software system (SAS Institute Inc., Cary, NC). Supporting Information contains examples of the SAS code corresponding to the models. Since the models assume normally distributed random quantities and a constant variance of measurement errors across groups, we verified these assumptions in a series of randomly selected proteins, using normal quantile-quantile plots and likelihood-ratio tests of constant variance. The results (not shown for lack of space) do not indicate gross departures from the assumptions.

The feature x group interaction in the ANOVA models represents deviations of some features from the consensus protein profile. For proteins with consistent patterns over all features, the model that contains feature-specific variances may overfit the data, and create problems with convergence of the REML procedure. The removal of interaction terms, and/or manual removal of some perfectly collinear features may be necessary. In our data sets, this was the case for 17 background serum proteins in the spike-in data set, and 18 proteins in the clinical data set.

The models were used to test each protein for differential abundance in all pairwise comparisons of groups, and to report the *p*-values of the tests. The *p*-values were then adjusted for multiple comparisons using the approach by Benjamini and Hochberg,[10] separately for each pairwise comparison. Since the per-feature model requires an *ad hoc* rule to make protein-level conclusions, we take an "aggressive" approach, and report the feature with the strongest evidence in favor of differential abundance for a protein over all features. Subject quantifications were derived as described in Table 1.

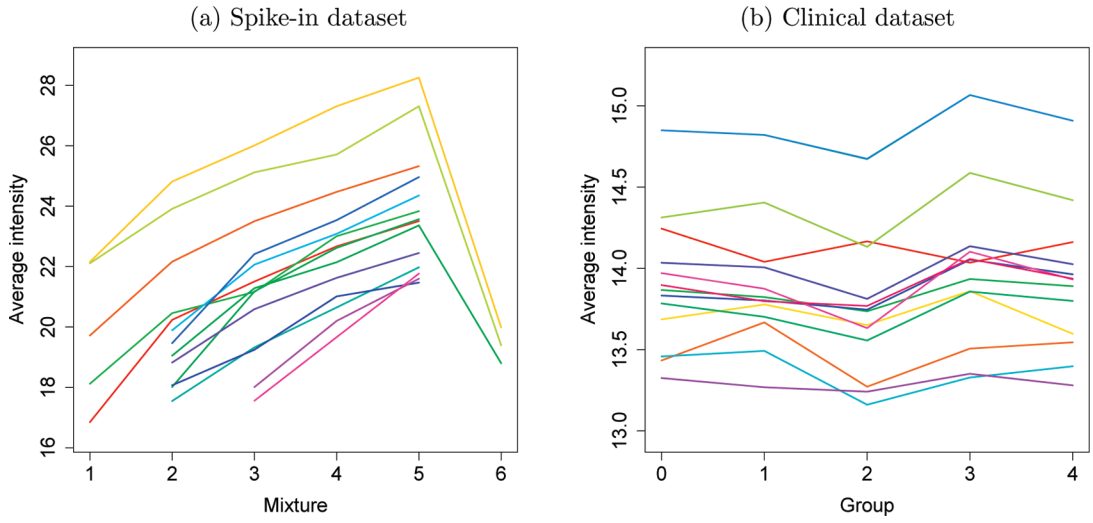

Figure 5. Representative quantitative protein profiles in the experimental data sets. *X*-axis: mixture type or disease group. *Y*-axis: average log-intensity of a feature, on average over the replicates. Each line represents a feature. (a) Spike-in data set: Alcohol dehydrogenase. (b) Clinical data set: protein 116844.

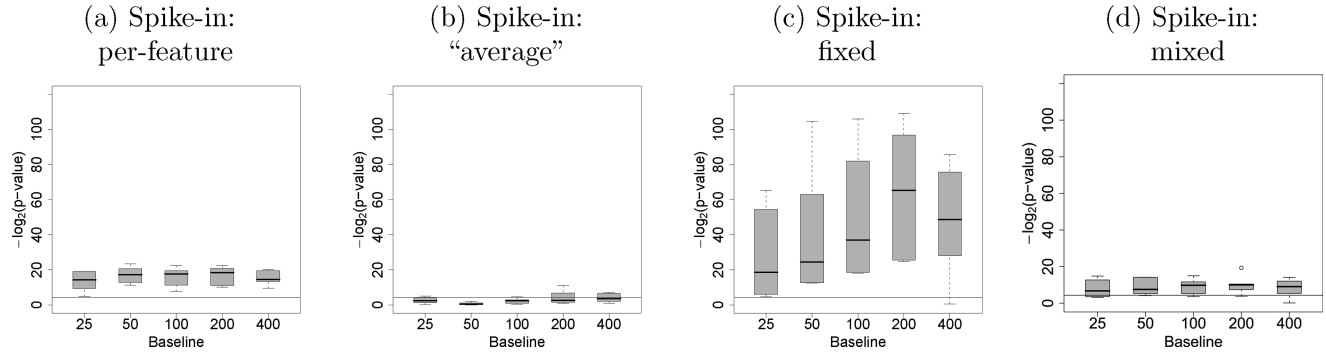

Figure 6. Sensitivity of the four basic models at detecting a 2-fold change in the spike-in data set, starting from five abundance baselines. *X*-axis: abundance baseline. *Y*-axis: $-\log_2$ (FDR-adjusted *p*-value). Each box contains the middle 50% of the proteins; the line within the box is the median. The higher the value on the *y*-axis, the stronger the evidence for differential abundance. The horizontal line corresponds to the FDR cutoff of 0.05.

Proteomic experiments often have a relatively small sample size. To make the discussion relevant to a typical experimental situation, results from the clinical data set were obtained with a reduced version where 10 subjects were randomly selected from each disease group, unless stated otherwise. Since there are no missing values, the reduced data set is balanced.

4.2. Group Testing: Sensitivity. Figure 6 shows the *p*-values of pairwise comparisons of abundance that correspond to 2-fold changes of the spiked concentrations, summarized over the 6 spiked proteins, and starting from a series of baselines. Several conclusions can be made from the figure.

First, the "average model" results in the lowest sensitivity. This is due to a smaller number of data points as compared to the fixed and mixed ANOVA, and to the presence of missing values in the data set. As discussed in section 2.5, averaging of the intensities is not an optimal summary in the presence of missing data. Second, the per-feature model has a better sensitivity than the "average model". This result is an artifact of our decision rule where the feature with the smallest *p*-value is selected to describe the results for the whole protein. However, even with this strategy, the increase in sensitivity is small. The reasons for this are a smaller number of data points per feature, and the fact that the per-feature approach involves a larger number of tests, and requires a more conservative adjustment for multiple comparisons. Moreover, as we will see in section 4.3, the relative increase of sensitivity with the per-feature approach has the price of a reduced specificity.

Third, the fixed effects ANOVA model outperforms the per-feature model and the "average model" in terms of sensitivity. The sensitivity increases with the increasing abundance baseline for all but the highest baseline. The highest baseline reaches the upper limit of the dynamic range, and makes it more difficult to detect the 2-fold change. Finally, the fixed effects model produces stronger evidence in favor of differential abundance than the mixed effects model. This result is typical for this type of ANOVA models, and is due to the fact that results from the mixed effects model refer to a larger scope of individuals, the individuals from the larger population, than the fixed effects model.

Figure 7a,b shows the number of differentially abundant proteins detected in all pairwise comparisons of disease groups in both the reduced and full clinical data sets, at the estimated FDR of 0.05. As in the case of the spike-in data set, the fixed effects ANOVA is the most sensitive. Since the reduced data set is fully balanced, and the full data set is nearly balanced, the results of the mixed ANOVA are identical to the results of the "average" model. The full data set yields a larger number of detected differences than the reduced data set due to an increased sample size; however, the fixed effects model maintains the highest sensitivity.

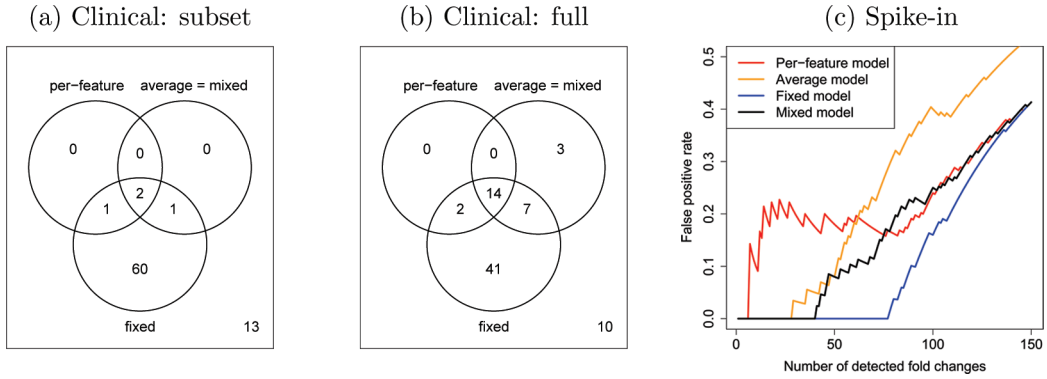

Figure 7. (a and b) Sensitivity of the four basic models at detecting changes in abundance in the clinical data set. Each circle shows the number of proteins with at least one detected pairwise difference between disease groups after the FDR cutoff 0.05. (a) Reduced data set with 50 subjects. (b) Full data set with 246 subjects. (c) Specificity of the four basic models at detecting changes in abundance in the spike-in data set. *X*-axis: number of differences in a comparison between two mixtures. *Y*-axis: false positive rate in all pairwise comparisons of mixtures. More specific models produce lower curves.

Results from extensions to the models are shown in Supporting Information. The extensions do not improve the sensitivity in these specific data sets. However, the results are dependent on the proteins under consideration and on the extent of changes in abundance in the biological system, and therefore, the extensions can be helpful in other situations. In particular, Supporting Information shows the results of tests of equal variance between features in the cardiovascular data set for the fixed effects and mixed effects models, which indicate that allowing for feature-specific variances in these ANOVA models is appropriate.

4.3. Group Testing: Specificity. In the spike-in data set, only 6 proteins changed in concentration between mixtures, and the concentration of the remaining proteins was kept fixed. Therefore, the models can be compared by their ability to prevent false positive discoveries of differential abundance. For each pairwise comparison of mixtures, the models were used to (1) test all proteins for differential abundance, (2) rank the proteins according to the resulting adjusted *p*-values, and (3) record the proportion of false positive discoveries while varying differential abundance cutoffs. Figure 7c displays the resulting proportion of false positives over all pairwise comparisons, as a function of the total number of discoveries for different cutoffs.

As can be seen from the figure, the per-feature approach is the least specific, and produces the largest proportion of false positive discoveries as compared to the other models. This is expected since a single feature is more likely to appear differentially abundant by random chance than a consensus summary of all the data for the protein. The "average model" has an intermediate specificity, and the fixed ANOVA model is the most specific and the most robust to experimental artifacts. Supporting Information shows that extensions to the basic models do not improve the specificity of finding differences in this data set.

4.4. Subject Quantification. The clinical data set contains biological replicates, and can be used to investigate subject quantification. We consider 11 proteins for which independent nephelometry measurements were made for each patient, and the reduced version of the data set. Subject quantifications were derived from the "average model" and from fixed and mixed ANOVA models as described in Table 1. Since the per-feature model has no subject quantification, we omit it from consideration.

We anticipate that more accurate models produce higher Pearson correlations with the nephelometry measurements. Figure 8a illustrates the improvements in correlations of the quantifications derived from the ANOVA models, as compared to the "average model". The correlations are roughly similar between the models for all proteins, which is a typical conclusion for fully balanced data sets. The highest improvement was obtained by applying the ANOVA models to the proteins where the "average model" performed poorly (i.e., where correlations with the nephelometry measurements were less than 0.4). Extensions to the basic models, in particular the median polish, did not improve the results substantially in this data set.

4.5. Impact of Missing Intensities: Group Testing. When estimating differences in protein abundance between groups, missing intensities of LC-MS features can introduce a bias, that is, a systematic deviation of the estimation from the true value. Here we evaluate the models according to their ability to limit the bias. To this end, we consider 10 randomly selected proteins in the reduced clinical data set, and artificially remove 30% of feature intensities for each of the proteins. We consider two scenarios for the removal. The first mimics noninformative missing data, and removes features with an equal probability. In the second scenario, we remove features with a probability that is inversely proportional to the abundance, so that low-abundant features are more likely to have missing values. The artificial data sets from the two scenarios were then used to fit all the models, and to derive model-based estimates of changes in abundance for all pairwise comparisons of groups. We calculate the bias as the deviation of the estimates in the presence of missing values from the corresponding estimates in the complete data set. For the per-feature model, the bias was calculated by computing the average deviation across the features.

Figure 8b summarizes the extent of the resulting bias, and points to two conclusions. First, when missing values appear more frequently at lower feature abundances, this increases the bias as compared to the case of noninformative missing. This is expected when nonmissing intensities are biased toward higher abundances. Second, the fixed and mixed ANOVA models are more robust to the missing data, and result in a smaller bias as compared to the per-feature and the "average model". The improved performance of the ANOVA is due to the fact that, in the presence of missing values, the estimates deviate from sample averages, and are more refined and more optimal consensus patterns over all features. Therefore, ANOVA-

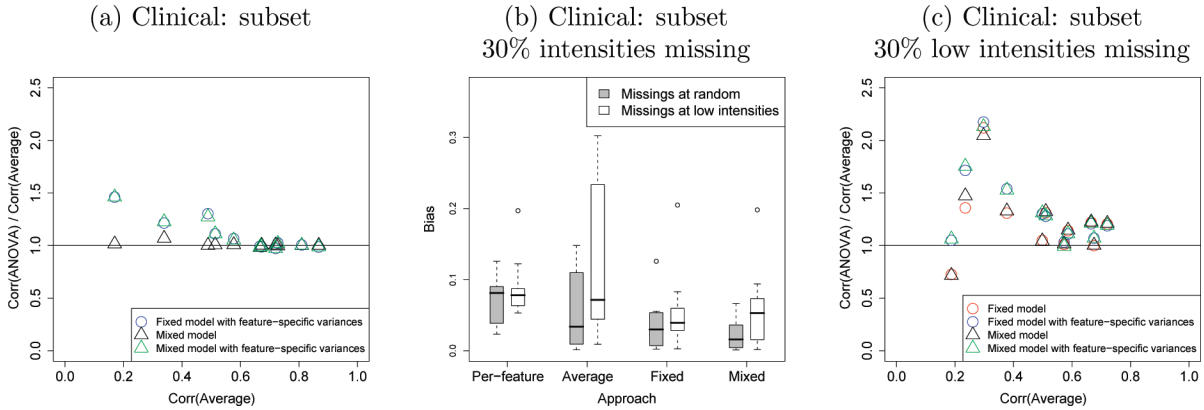

Figure 8. (a) Accuracy of subject quantification in the clinical data set. *X*-axis: correlation of protein abundances derived from the "average model" and the nephelometry measurements. *Y*-axis: ratio of ANOVA-based correlations and the "average model"-based correlations. Points above the horizontal line indicate improved quantification. (b) Effect of missing intensities on estimates of differences between disease groups. *X*-axis: model type. *Y*-axis: absolute value of bias in estimation of the differences after removing 30% of intensities. Lower boxes correspond to more robust models. (c) Same as panel a, but after removing 30% of feature intensities, with higher probability of removal for low-abundance features.

based models are particularly beneficial in the presence of missing data, and for other unbalanced experimental designs.

4.6. Impact of Missing Intensities: Subject Quantification. The two data sets with artificially removed observations were also used to illustrate the robustness of the approaches to missing data with respect to subject quantification. We expect that more accurate approaches produce higher Pearson correlations with the nephelometry measurements, despite the missing data. Figure 8c summarizes the improvements in the correlations of the subject quantifications derived from the ANOVA models as compared to the "average model" on these data sets.

First, as can be seen from the *x*-axis of the figure, the correlations derived from the "average model" are slightly lower than in the absence of missing intensities in Figure 8a, reflecting the extra uncertainty resulting from missing observations. Further, the ANOVA models show a greater relative improvement, up to 2-fold, in correlations over the "average model" than in Figure 8a. Models with a feature-specific variance component show the most improvement for these proteins.

5. Discussion

As we have seen, fixed and mixed effects ANOVA are more sensitive and specific for group testing, and more accurate for subject quantification than the per-feature and averaging approaches, in particular in the presence of abundance-dependent missing data. We therefore recommend the general ANOVA framework for protein-level quantification. The difference between the fixed and mixed effects lies primarily in the scope of inference for group testing. The fixed effects model restricts the scope of conclusions to the individuals in the study, and results in a higher power of tests. We recommend this model in the initial screening stage. The mixed effects model extends the scope of the inference to the entire underlying populations of individuals, and is more appropriate for investigations at the validation stage. Even though the mixed effects model loses power as compared to the fixed effects, it outperforms both per-feature and averaging approaches, in particular in the presence of missing data.

The superior performance of the ANOVA models over the per-feature analysis and averaging is due to the fact that they capitalize on the redundancies provided by multiple LC-MS features, and explicitly model the replicate structure of the data. This underscores the importance of working with the observed feature intensities directly, as opposed to using sums, averages, or ratios of intensities from different samples as is sometimes done. The log transformation translates a multiplicative signal into an additive signal, and facilitates the development of such models.

The fixed and mixed models that we discussed are not the only possible ANOVA models for LC-MS data. The ANOVA framework is general and flexible, and can accommodate a variety of experimental situations and settings. In fact, changes in experimental design such as repeated measurements, the introduction of technical replication, or the presence of other important sources of variation will typically require changes or additions of model terms, which in turn will affect group testing and subject quantification. Conversely, some sources of variation, such as deviations of features from the overall abundance profiles, or natural variation between individuals in the study, may be relatively small. The corresponding terms can then be removed from the model, as has been done, for example, in ref 14. However, the assumption of negligible variation needs to be verified explicitly and anew in each experimental setting. Development of a model that appropriately represents the specific experimental situation can be a difficult task, and we recommend consulting with a statistician whenever possible.

Extensions examined in this work, namely, feature-specific variances, Empirical Bayes ANOVA, and Tukey median polish, do not substantially improve the sensitivity and specificity of the original models in these particular data sets. However, they may be helpful in other problems. For example, the spike-in data set contains no biological replicates, and is therefore less variable than a typical biological data set. Empirical Bayes ANOVA is geared toward improving performance specifically in cases of high variability and small sample size. Furthermore, tests of equal variance across features performed on these two data sets rejected the null hypothesis for many proteins, indicating that feature-specific variance can improve the performance of the models in other problems.

The ANOVA framework supports numerous additional extensions. For example, throughout the paper, we assumed that

peptide features are unambiguously mapped. It is possible to develop extensions to these procedures that would simultaneously deconvolute ambiguous protein memberships and provide quantification.[19] Furthermore, the models can be extended to protein quantification from other workflows, such as labeling workflows and selected reaction monitoring (SRM). While these extensions may also require changes in the specific model terms, as well as in the quantities used for group testing and subject quantification, the general underlying principle will remain the same.

Supporting Information Available: Supplemental materials containing model-based testing quantities, results and discussion of extensions to the quantification models, and SAS code. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198–207.
(2) Mueller, L. N.; Rinner, O.; Schmidt, A.; Letarte, S.; Bodenmiller, B.; Brusniak, M.; Vitek, O.; Aebersold, R.; Muller, M. *Proteomics* **2007**, *7*, 3470–3480.
(3) Brusniak, M.; Bodenmiller, B.; Campbell, D.; Cooke, K.; Eddes, J.; Garbutt, A.; Lau, H.; Letarte, S.; Mueller, L.; Sharma, V.; Vitek, O.; Zhang, N.; Aebersold, R.; Watts, J. *BMC Bioinf.* **2008**, 9.
(4) Higgs, R. E.; Knierman, M. D.; Gelfanova, V.; Butler, J. P.; Hale, J. E. *J. Proteome Res.* **2005**, *4*, 1442–1450.
(5) Patil, S. T.; Higgs, R. E.; Brandt, J. E.; Knierman, M. D.; Gelfanova, V.; Butler, J. P.; Downing, A. M.; Dorocke, J.; Dean, R. A.; Potter, W. Z.; Michelson, D.; Pan, A. X.; Jhee, S. S.; Hale, J. E. *J. Proteome Res.* **2007**, *6*, 955–966.
(6) Kutner, M.; Nachtsheim, C.; Neter, J.; Li, W. *Applied Linear Statistical Models*, Fifth ed.; McGraw-Hill/Irwin: New York, 2004.
(7) Montgomery, D. C. *Design and Analysis of Experiments*, Fifth ed.; John Wiley and Sons: New York, 2000.
(8) Ragg, S.; Fokin, V.; Podgorski, K.; Schadow, G.; Vitek, O.; Kastrati, A.; Schmig, A.; Lorenz-Braun, S.; Ott, I. *Circulation* **2007**, *116*, 575.
(9) Oberg, A. L.; Vitek, O. *J. Proteome Res.* **2009**, *8*, 2144–2156.
(10) Benjamini, Y.; Hochberg, Y. *J. R. Stat. Soc.* **1995**, *57*, 289–300.
(11) Daly, D. S.; Anderson, K. K.; Panisko, E. A.; Purvine, S. O.; Fang, R.; Monroe, M. E.; Baker, S. E. *J. Proteome Res.* **2008**, *7*, 1209–1217.
(12) 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.; Bergen, I. H. R. *J. Proteome Res.* **2008**, *7*, 225–233.
(13) Bukhman, Y. V.; Dharsee, M.; Ewing, R.; Chu, P.; Topaloglou, T.; Le Bihan, T.; Goh, T.; Duewel, H.; Stewart, I. I.; Wisniewski, J. R.; Ng, N. F. *J. Bioinf. Comput. Biol.* **2008**, *6*, 107–123.
(14) Karpievitch, Y.; Stanley, J.; Taverner, T.; Huang, J.; Adkins, J. N.; Ansong, C.; Heffron, F.; Metz, T. O.; Qian, W.-J.; Yoon, H.; Smith, R. D.; Dabney, A. R. *Bioinformatics* **2009**, 1–7.
(15) Smyth, G. K. *Stat. Appl. Genetics Mol. Biol.* **2004**, *3* (1), Article 3.
(16) Mosteller, F.; Tukey, J. *Data Analysis and Regression*; Addison-Wesley: Reading, MA, 1977.
(17) Bolstad, B. M.; Irizarry, R. A.; Astrand, M.; Speed, T. P. *Bioinformatics* **2003**, *19*, 185–193.
(18) Higgs, R. E.; Knierman, M. D.; Gelfanova, V.; Butler, J. P.; Hale, J. E. *Methods Mol. Biol.* **2008**, *428*, 209–230.
(19) Dost, B.; Bafna, V.; Bandeira, N.; Li, X.; Shen, Z.; Briggs, S. Shared Peptides in Mass Spectrometry Based Protein Quantification. In *Proceedings of the International Conference on Research in Computational Molecular Biology (RECOMB)*; Springer: New York, 2009.

PR900610Q