Supplemental Material

Protein profiles. We define the *expected protein expression profile* for a given condition as the collection of proteins and their corresponding amounts in a representative sample (e.g. serum, saliva, tissue supernatant) of uniform size (volume or mass) from a *population* of interest. We use the term *expected* throughout the manuscript in a statistical sense to represent the true, but unknown, average state of nature for the universe of subjects (or *population*) to whom scientific interest is directed (e.g. all adults with a given disease). Furthermore, we define the *complete proteome* as the set of all proteins encoded in an organism's genome, and the *protein profile* as the list of numbers giving the concentrations for each protein in that proteome (unexpressed proteins have a protein profile concentration of zero). While the complete proteome is fixed, the protein profile can change across conditions, individuals, and even samples from the same individual. Since we are constrained by finite sample sizes, the expected protein profile is a theoretical concept and impossible to determine exactly. It is important, however, in our statistical model as it provides the foundation for the relationships that follow.

Let $\left[P_i\right]=\left[P_1,\ldots,P_I\right]$ and $\left[Q_i\right]=\left[Q_1,\ldots,Q_I\right]$, $i=1,\ldots,I$, be the expected protein expression profiles for the control and treated conditions, respectively. For the proteome containing *I* proteins, we write the concentration of the *i*th protein as P_i and the entire set of numbers giving the expression levels of each of those proteins as $\left[P_i\right]$. The concentrations for the same set of proteins in the treated condition are given similarly using the notation $\left[Q_i\right]$. Let $P=\sum_i P_i$ and $Q=\sum_i Q_i$ represent the total amounts of protein in the aforementioned representative samples from the control and treated conditions, respectively. We relate the treated and control expected protein profiles by

defining ratios R and R_i that give the relative amount of total protein and the relative amounts of protein *i* in the proteome. The expected protein profile from the treated condition can then be expressed in terms of the profile of the control condition as $[Q_i] = [P_i \cdot R_i \cdot R]$, where $R = Q/P$ is the ratio of expected total protein for the treated and control conditions, respectively, and $R_i = (Q_i/Q)/(P_i/P)$ is the ratio of the expected amount of protein *i* for the treated relative to the control condition, after adjusting for differences in total protein. One can verify algebraically the relationship between Q_i and P_i, but it is satisfying that the terms in the relationship have an intuitive interpretation as illustrated in the following example. Suppose that a given protein (protein 1, say) is present in the control and treated conditions at expected concentrations of $P_1 = 4$ and $Q_1 = 2$ pg/ml, respectively. Furthermore, suppose the total amount of protein in the control and treated conditions are $P = 12$ and $Q = 10$ μg/ml, respectively, so that the expected fold-change comparing the treated to control condition for this protein is $R_1 = (2/10)/(4/12) = 0.6$. Then $P_1 \cdot R_1 = 4 \text{ pg/ml} \times 0.6 = 2.4 \text{ pg/ml}$ is the expected amount of protein 1 in the treated condition, assuming the amounts of total protein are equivalent. Multiplication by $R = 10/12$ simply corrects that product based on differences in total expected protein between the two conditions. Thus, $P_1 \cdot R_1 \cdot R = 2.4 \text{ pg/ml} \times (10/12) = 2 \text{ pg/ml} = Q_1$. We include both R and R_i in the model to indicate explicitly that, when compared to the control group, treatment may result in change across all proteins as well as changes in individual proteins. We acknowledge, however, that iTRAQ experimental design generally precludes the possibility of

estimating changes in total protein expression due to iTRAQ protocol that stipulates loading equal amounts of total protein in each iTRAQ channel.

Biological variation. Biological variation has been shown to contribute substantially to the variability observed in iTRAQ data.[10] As such, we include terms in our model that express the variation in the protein profile attributable to each subject. Let k and k' index subjects from the control and treated conditions, respectively. For subjects k and k', define $D_{k,i}$ and $D_{k',i}$ as the ratios of the amount of protein i to the expected amount of protein i for that subject's condition. Specifically, let $S_{k,i}$ and $T_{k',i}$ be the amounts of protein *i* for subjects k and k', respectively. Then $D_{k,i} = S_{k,i} / P_i$ and $D_{k',i} = T_{k',i} / Q_i$. The protein expression profile for control subject k is given by

$$\left[P_i \cdot D_{k,i} \right],$$

where the expression level for each protein is given by the product of the expression level from the expected profile and a subject-specific factor. Similarly, the profile for treated subject k' is given by

$$\left[P_i \cdot R_i \cdot R \cdot D_{k',i} \right],$$

where the product also includes the proteome-wide and protein-specific factors for the treated condition.

Peptide profile. iTRAQ measurements are made at the peptide level and therefore the model must reflect the relationships between peptide and associated protein expression levels. This association may be ambiguous as some tryptic peptides can be associated with more than one protein. These peptides are said to be degenerate and are often eliminated prior to analysis. We therefore assume that observed peptides are uniquely

assigned to a protein. Accordingly, we use function notation in our peptide subscripts, *j*(*i*), to indicate the *j*th peptide is uniquely derived from the *i*th protein. In statistical parlance, we say that peptides are *nested* within proteins.

Post-translational modifications and/or splice variants can affect individual peptides within a protein in a condition-dependent manner. As such, our model includes terms capturing the effect of condition at the peptide level in addition to the condition-specific protein-level effects discussed previously. Accordingly, define $V_{j(i)}$ and $W_{j(i)}$ as the expected amount of the *j*th peptide in the control and treated conditions, respectively. For control subjects, define $F_{j(i)} = V_{j(i)} / P_i$ as the ratio of the expected amount of the *j*th peptide to the expected amount of the *i*th protein. Further define $G_{j(i)} = \left(W_{j(i)} / Q_i\right) / \left(V_{j(i)} / P_i\right)$ as the ratio of the expected amount of the *j*th peptide in the treated relative to the control condition, adjusting for the expected amount of the *i*th protein present in each condition. The quantity $G_{j(i)}$ compares the peptide per protein ratio in the treated condition to the peptide per protein ratio in the control condition. These relationships allow us to write expressions for the peptide profiles for subjects k and k' in the control and treated conditions as

$$\left[P_i \cdot D_{k,i} \cdot F_{j(i)} \right]$$

and

$$\left[P_i \cdot R_i \cdot R \cdot D_{k',i} \cdot F_{j(i)} \cdot G_{j(i)} \right],$$

respectively.

iTRAQ labeling and mixing. The assignment of samples to iTRAQ reagents (channels) is an experimental design issue and many configurations are possible. To

facilitate further development of this model, we assume a specific design in which samples for two subjects from each of the two conditions are labeled with the four iTRAQ reagents. The mass spectral analysis of these four labeled samples comprises a single iTRAQ experiment. Specifically, assume samples for control subjects k_1 and k_2 are labeled using the 114 and 115 tags, and samples for treated subjects k_1' and k_2' are labeled using the 116 and 117 tags. iTRAQ protocol stipulates loading equal amounts of total protein in each iTRAQ channel (e.g. 100 μg recommended by ABI). Accordingly, let I_{114}, I_{115}, I_{116} and I_{117} be the proportion of the respective samples loaded into each iTRAQ channel. Only labeled peptides contribute to the reporter ion cluster and, thus, each contributing peptide must have been derivatized in the iTRAQ labeling reaction. Define Z_{114}, Z_{115}, Z_{116} and Z_{117} as the labeling efficiencies (values between 0 and 1) of the four reagents indicating the fraction of the peptides in the each of the four samples successfully derivatized in the labeling step. Perfect efficiency is achieved when $Z_{114} = Z_{115} = Z_{116} = Z_{117} = 1$ and every peptide in the mixture is labeled with the appropriate tag. The labeled iTRAQ samples are mixed together and the peptides typically are separated in two dimensions using reverse phase and strong cation-exchange chromatography. Let M_{114}, M_{115}, M_{116} and M_{117} be the relative amounts of each iTRAQ sample added to this mixture. Ideally, samples are mixed in equal proportions with $M_{114} = M_{115} = M_{116} = M_{117}$. However, these proportions may differ as a result of pipetting errors. The labeled peptide profiles for the control and treated samples are

$$\left[P_i \cdot D_{k_1,i} \cdot F_{j(i)} \cdot I_{114} \cdot Z_{114} \cdot M_{114} \right],$$

$$\left[P_i \cdot D_{k_2,i} \cdot F_{j(i)} \cdot I_{115} \cdot Z_{115} \cdot M_{115} \right],$$

$$\left[P_i \cdot R_i \cdot R \cdot D_{k_1',i} \cdot F_{j(i)} \cdot G_{j(i)} \cdot I_{116} \cdot Z_{116} \cdot M_{116} \right],$$

and

$$\left[P_i \cdot R_i \cdot R \cdot D_{k_2',i} \cdot F_{j(i)} \cdot G_{j(i)} \cdot I_{117} \cdot Z_{117} \cdot M_{117} \right].$$

The expressions above give the labeled peptide profiles present in the mixture prior to separation, one profile for each of the four iTRAQ reagents.

Mass spectrometry. The mixture of labeled peptides is then separated using two stages of chromatography and fractions are subjected to mass spectrometry. Peptide peaks are selected from the mass spectra of each fraction and subjected to tandem mass spectrometry (MS/MS). Four reporter ion peaks appear in a small cluster in the low mass range (m/z 114 – 117). These four peak intensities (areas under the peaks corrected for isotopic impurity) are assumed to be proportional to the amount of the given peptide labeled with the appropriate tag. We define this constant of proportionality as *B* and include it as a factor in each of the peptide profiles to yield the expected reporter ion peak area profiles.

Measurement noise. We relate the expected reporter ion peak areas to the observed reporter ion peak areas through a term that represents the measurement noise associated with each observed peak. We assume that the measurement noise, represented by the random quantity *E*, is distributed such that the mean of the logarithm of *E* is zero. Under this condition, the error contributes no bias to the reporter ion peak area measurement scale. We assume further that the biological and measurement errors are uncorrelated. The resulting profiles of observed ion currents are

$$\left[P_i \cdot D_{k_1,i} \cdot F_{j(i)} \cdot I_{114} \cdot Z_{114} \cdot M_{114} \cdot B \cdot E_{s,j(i),114} \right],$$

$$\left[P_i \cdot D_{k_2,i} \cdot F_{j(i)} \cdot I_{115} \cdot Z_{115} \cdot M_{115} \cdot B \cdot E_{s,j(i),115}\right],$$

$$\left[P_i \cdot R_i \cdot R \cdot D_{k_1',i} \cdot F_{j(i)} \cdot G_{j(i)} \cdot I_{116} \cdot Z_{116} \cdot M_{116} \cdot B \cdot E_{s,j(i),116}\right],$$

and

$$\left[P_i \cdot R_i \cdot R \cdot D_{k_2',i} \cdot F_{j(i)} \cdot G_{j(i)} \cdot I_{117} \cdot Z_{117} \cdot M_{117} \cdot B \cdot E_{s,j(i),117}\right].$$

Here, *s* indexes the spectrum indicating that the error contribution varies across spectra from the same peptide. The expressions above now provide a model that relates the relative changes in protein and peptide expression due to condition (given by R, R_i and $G_{j(i)}$) to the observed reporter ion peak areas while explicitly capturing biological variation ($D_{k,i}$), tagging inefficiencies ($Z_{114}, Z_{115}, Z_{116}, Z_{117}$), pipetting errors ($I_{114}, I_{115}, I_{116}, I_{117}$ and $M_{114}, M_{115}, M_{116}, M_{117}$), peptide-to-protein relationships ($F_{j(i)}$), reporter ion peak area to peptide relationships (B), and measurement noise ($E_{s,j(i),114}, \ldots, E_{s,j(i),117}$).

Multiple experiments. Current protein quantitation software for iTRAQ data limits analysis to that of a single iTRAQ experiment. Nonetheless, we increase the power of our study by increasing the number of biological replicates, so it is desirable to include in a single analysis data collected across multiple experiments. Accordingly, we extend the model to incorporate multiple iTRAQ experiments. We assume the process leading to data from additional iTRAQ experiments follows the described model to the iTRAQ labeling step, but the effects due to loading, labeling, mixing, mass spectrometry, and measurement noise are likely to vary across experiments. For the *q*th iTRAQ experiment, we include an additional subscript for terms *I*, *Z*, *M*, *B*, and *E*, indicating that these terms

vary across experiments. The observed reporter ion peak areas for the qth experiment are therefore

$$\left[P_i \cdot D_{k_1,i} \cdot F_{j(i)} \cdot I_{q,114} \cdot Z_{q,114} \cdot M_{q,114} \cdot B_q \cdot E_{q,s,j(i),114}\right],$$

$$\left[P_i \cdot D_{k_2,i} \cdot F_{j(i)} \cdot I_{q,115} \cdot Z_{q,115} \cdot M_{q,115} \cdot B_q \cdot E_{q,s,j(i),115}\right],$$

$$\left[P_i \cdot R_i \cdot R \cdot D_{k_1',i} \cdot F_{j(i)} \cdot G_{j(i)} \cdot I_{q,116} \cdot Z_{q,116} \cdot M_{q,116} \cdot B_q \cdot E_{q,s,j(i),116}\right], \quad (1)$$

and

$$\left[P_i \cdot R_i \cdot R \cdot D_{k_2',i} \cdot F_{j(i)} \cdot G_{j(i)} \cdot I_{q,117} \cdot Z_{q,117} \cdot M_{q,117} \cdot B_q \cdot E_{q,s,j(i),117}\right].$$

Translation to a Statistical Model.

Our purpose is to use Model 1 to address the scientific question of interest, namely to identify differentially expressed proteins across conditions. We focus now on translating the conceptual model described in Model 1 to a statistical model. We begin with two simplifications, the first of which combines the biological error, D, and measurement noise, E, into a single error term, H. We also drop the subject-level subscript from the combined error term, since the subject is uniquely identified from the combination of experiment (q) and tag (114, 115, 116, or 117). The second simplification combines factors associated with sample loading (I), iTRAQ labeling (Z), and sample mixing (M) into a single factor, V. The latter is necessary since these factors are *confounded*, that is to say their contributions can not be estimated separately. Practically, this means that if we detect a uniform shift (bias) in the collection of reporter ion peak areas associated with one label relative to another within a given iTRAQ experiment, that shift could be attributed to inaccuracies in sample loading, differences in labeling efficiencies, or to

pipetting or other technical errors. Regardless, there is no way to disentangle these effects. This simplified version of Model 1 becomes

$$\left[P_i \cdot F_{j(i)} \cdot V_{q,114} \cdot B_q \cdot H_{i,j(i),q,s,114}\right],$$

$$\left[P_i \cdot F_{j(i)} \cdot V_{q,115} \cdot B_q \cdot H_{i,j(i),q,s,115}\right],$$

$$\left[P_i \cdot R_i \cdot R \cdot F_{j(i)} \cdot G_{j(i)} \cdot V_{q,116} \cdot B_q \cdot H_{i,j(i),q,s,116}\right], \tag{2}$$

and

$$\left[P_i \cdot R_i \cdot R \cdot F_{j(i)} \cdot G_{j(i)} \cdot V_{q,117} \cdot B_q \cdot H_{i,j(i),q,s,117}\right].$$

The collection of observed reporter ion peak areas can be described jointly using a single model if we introduce subscripts for condition, c, and iTRAQ tag, ℓ, as follows:

$$\left[P_i \cdot R_{i,c} \cdot R_c \cdot F_{j(i)} \cdot G_{j(i),c} \cdot V_{q,\ell} \cdot B_q \cdot H_{i,j(i),c,q,s,\ell}\right]. \tag{3}$$

In words, the collection of observed reporter ion peak areas is described by a product of factors capturing effects due to: protein (P_i); protein by condition ($R_{i,c}$); condition (R_c); peptide ($F_{j(i)}$); peptide by condition ($G_{j(i),c}$); loading, labeling and mixing differences across iTRAQ experiments ($V_{q,\ell}$); iTRAQ experiment (B_q); and the biological and experimental error not captured by the remaining terms ($H_{i,j(i),c,q,s,\ell}$). We note that by defining the effects for protein by condition, condition, and peptide by condition to equal one for the control samples in Model 3, we recover all four tag-specific collections of reporter ion peak areas specified in Model 2.