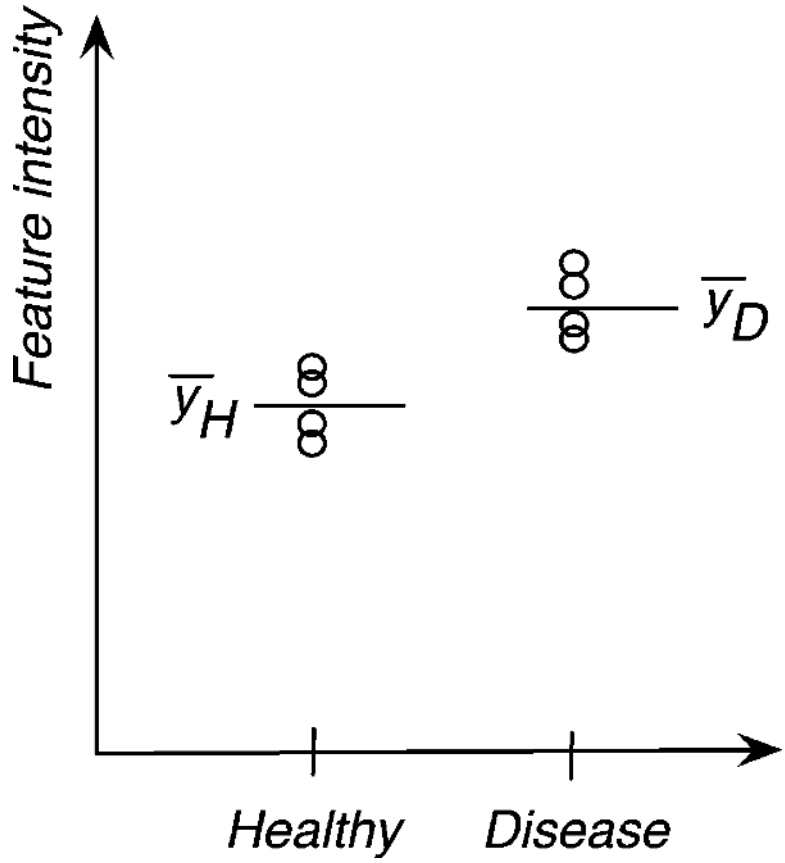
Journal of proteome research

Reviews

Statistical Design of Quantitative Mass Spectrometry-Based Proteomic Experiments

Ann L. Oberg, and Olga Vitek

J. Proteome Res., **Article ASAP** • DOI: 10.1021/pr8010099 • Publication Date (Web): 17 February 2009



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Access to high resolution figures

Journal of Proteome Research is published by the American Chemical Society.
1155 Sixteenth Street N.W., Washington, DC 20036

reviews Journal of proteome research

Statistical Design of Quantitative Mass Spectrometry-Based Proteomic Experiments

Ann L. Oberg[†] and Olga Vitek*[,‡]

Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, Minnesota 55905, and Department of Statistics and Department of Computer Science, Purdue University, 250 North University Street, West Lafayette, Indiana 47907

Received November 21, 2008

We review the fundamental principles of statistical experimental design, and their application to quantitative mass spectrometry-based proteomics. We focus on class comparison using Analysis of Variance (ANOVA), and discuss how randomization, replication and blocking help avoid systematic biases due to the experimental procedure, and help optimize our ability to detect true quantitative changes between groups. We also discuss the issues of pooling multiple biological specimens for a single mass analysis, and calculation of the number of replicates in a future study. When applicable, we emphasize the parallels between designing quantitative proteomic experiments and experiments with gene expression microarrays, and give examples from that area of research. We illustrate the discussion using theoretical considerations, and using real-data examples of profiling of disease.

Keywords: Quantitative proteomics • Statistical design of experiments • Analysis of Variance • Mixed models • Randomization • Replication • Blocking • Pooling • Sample size

1. Introduction

Quantitative proteomics monitors patterns of protein abundance in biological samples under various conditions and states. It plays an important role in understanding the functioning of living organisms, and in search of biomarkers for early detection, diagnosis and prognosis of disease. Tremendous progress in performance of mass spectrometers, as well as in experimental workflows that these instruments support, make mass spectrometry-based investigations particularly suitable for quantitative proteomics. It is now possible to measure with high sensitivity the abundance of peptides obtained from protein digestion by liquid chromatography coupled with online mass spectrometry analysis (LC-MS), or based on stable isotope labeling of proteins where samples are labeled chemically (e.g., in isotope coded affinity tag, ICAT; or isobaric tags for relative and absolute quantification, iTRAQ) or metabolically (e.g., in stable isotope labeling with amino acids in cell culture SILAC) mixed together.[1−3] These analyses can be performed in both global (hypothesis-free) or targeted (hypothesis-driven) mode.[4]

Once mass spectra are collected, a typical computational analysis of these experiments involves extraction and quantification of spectral features, which can be peaks in the initial MS1 spectra (characterized by their ratio of mass over charge and retention time) in label-free workflows, peaks in MS2 and MS3 spectra for labeling workflows, or transitions from targeted SRM experiments. A variety of computational tools for feature extraction and quantification have recently been developed and optimized.[5] The list of quantified features is subsequently subjected to statistical and machine learning analysis steps which aim at class comparison (i.e., hypothesis testing), class discovery (i.e., unsupervised clustering) and class prediction (i.e., supervised classification).[6,7]

Despite the progress, all quantitative investigations fail to deliver reproducible and accurate results if proper attention is not devoted to the experimental design. A comparison of two populations, such as disease patients and controls, will result in systematic mistakes if biological samples were selected or handled in different ways not intended by the purposes of the experiment. No amount of improvement in instrument sensitivity, sophisticated statistical analysis or increase in number of biological replicates will be able to correct these mistakes. At the same time, an unbiased experimental design that fails to account for known sources of technical variation will result in inefficient comparisons, and hamper the ability to find true quantitative changes. The issues of bias and efficiency in proteomic research have recently received a lot of attention, and sources of bias[8] and of experimental variation have been widely discussed.[9−11] These discussions emphasized the need for applying principles of statistical experimental design in quantitative proteomics.

This paper reviews the fundamental principles of statistical design of experiments,[12−14] and provides guidelines for their applications in quantitative mass spectrometry-based proteomics. Successful application of these principles will help increase reproducibility of the conclusions by avoiding bias, and maximize the sensitivity of the statistical analysis procedures. We will illustrate the discussion by examples in disease-related research; however, the same principles apply in other

* To whom correspondence should be addressed. Department of Statistics and Department of Computer Science, Purdue University, 250 North University Street, West Lafayette, Indiana 47907. E-mail: ovitek@stat.purdue.edu.
[†] Department of Health Sciences Research, Mayo Clinic.
[‡] Purdue University.

10.1021/pr8010099 CCC: $40.75

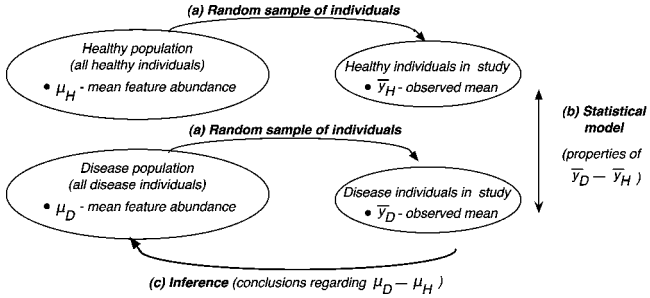


Figure 1. A schematic representation of the statistical inference procedure. (a) Random sampling ensures that the individuals in the study are representative of the population of interest. $\bar{y}_H$ and $\bar{y}_D$ are data-derived estimates of the population means. (b) A statistical model provides a mathematical description of the observed data. (c) The data and the model are used to make conclusions regarding the entire populations.

experiments, for example, when working with model organisms. We will focus on studies aiming at class comparison, that is, at finding spectral features that significantly differ in abundance between disease groups. We will assume that biological samples are already selected from the underlying populations and are independent, and will avoid more complex designs such as repeated measurements on a given individual in time.

We will examine in detail the choices of allocating experimental resources in the context of label-free and labeling workflows. It is not our goal to judge the relative performance of these workflows. Instead we will give examples of resource allocations in each case, while emphasizing properties of the corresponding designs from the statistical perspective. We will assume that spectral features are appropriately extracted, and quantified on a continuous scale. We also assume that an appropriate normalization procedure that removes systematic differences in intensities of individual MS runs, labeling agents, and so on is already applied. We will not consider feature quantification based on spectral counting, or protein quantification on the basis of multiple peptides.

Most concepts and issues of experimental design in quantitative proteomics hold in the broader class of high-throughput experiments in molecular biology. In particular, analysis of data from gene expression microarrays has similar objectives,[15] and experimental designs for microarrays have attracted much research. A series of specialized designs has been proposed,[16−18] and validity and reproducibility of microarray-based clinical research has also been discussed.[19] We will emphasize the parallels between designing quantitative proteomic and microarray experiments, and will give examples from this area when appilcable to proteomics. Our conclusions are applicable to other proteomic technologies such as selected reaction monitoring (SRM, also known as multiple reaction monitoring (MRM)) and in-gel proteomics, as well as to the profiling experiments outside of the proteomic research such as NMR- and MS-based metabolomics.

2. Fundamental Principles of Statistical Design of Experiments

Statistical analysis is used in studies of large groups of individuals given the information collected on a smaller subset. An example of statistical analysis is illustrated in Figure 1 where one compares two large groups of individuals, such as patients with a disease and healthy controls. We are interested in comparing the abundance of a spectral feature between the two populations. To this end, we select subsets of patients from each population. *Statistical inference* uses feature abundance measured on the selected individuals to make conclusions regarding the entire underlying populations.

Experimental design is a protocol that defines the populations of interest, selects the individuals for the study from the populations and/or allocates them to treatment groups, and arranges the experimental material in space and time. Experimental design has two goals. First, it ensures that statistical inference avoids *bias*, that is, it avoids systematic errors in our conclusions regarding the populations. Second, it ensures that the experiment is *efficient*, that is, it minimizes the random variation for a given amount of cost. The formal mathematical methodology for the statistical design of experiments was introduced by R. A. Fisher,[20] who considered three fundamental principles: replication, randomization and blocking.

Replication serves two purposes. First, it allows one to assess whether the observed difference in a measurement is likely to occur by random chance. For example, the experimental design in Figure 2a involves a single individual from both healthy and disease groups. The observed difference can represent the true difference between the populations, but can also be an artifact of selecting these specific individuals, or of the measurement error. An experimental design with replication, such as in Figures 2b and c, allows us to distinguish these situations. In Figure 2b, experimental variation is small, and the observed difference is likely to be informative of the disease. In Figure 2c, the experimental variation is large, and the observed difference is more likely to be due to chance.

The second purpose of replication is to ensure the reliability of our conclusions drawn from the observed data. Increasing the number of replicates results in a more precise inference regarding differences between the groups. In the example of Figure 2c, a larger number of replicates may allow us to demonstrate that, despite large experimental variation, the observed difference is indeed systematic.

Randomization also serves two purposes. First, it guards us against biases caused by undesirable, and potentially unknown experimental artifacts. Figure 3 illustrates an experiment in the presence of an undesirable instrumental drift in time. In the nonrandomized (sequential) experimental design in Figure 3a, spectra from healthy individuals were acquired in days 1−2, and from disease patients in days 3−4. This design creates a *confounding effect*, that is, it introduces two convoluted reasons for the difference between the groups, and biases our conclusions regarding the disease. Randomly selecting individuals from the underlying population, and randomizing the order of sample processing and spectral acquisition, avoids such a situation. The artifacts will be roughly equally distributed across groups as shown in Figure 3b, thereby eliminating the bias.

The second purpose of randomization is to allow the observed measurements to be regarded as random samples from underlying populations as shown in Figure 1. This is necessary since most statistical analysis techniques are based on the random sampling assumption.

Blocking helps reduce the bias and variance due to known sources of experimental variation. A completely randomized design in Figure 3b has two drawbacks. First, randomization of the order of spectral acquisition can potentially produce unequal allocations, for example, in assigning more disease patients toward days 3 and 4. Second, the variability within each group in Figure 3b is inflated by a combination of the

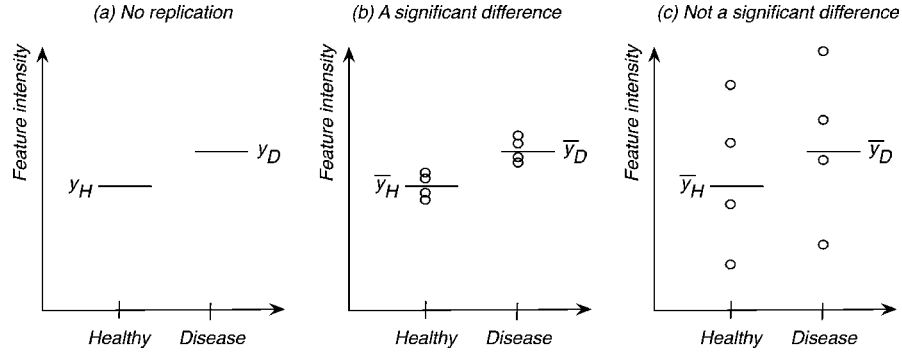


Figure 2. (a) Experimental design with no replication; horizontal lines are measurements from the single individual in each group. In absence of replication, one cannot determine whether the observed difference is systematic, or due to random chance. (b) Experimental design with replication. Circles represent measurements from patients, and horizontal lines are average measurements in each group. Small variation indicates that the difference in group means is unlikely to occur by random chance alone. (c) Large variation indicates that the difference in group means is likely to occur by random chance.

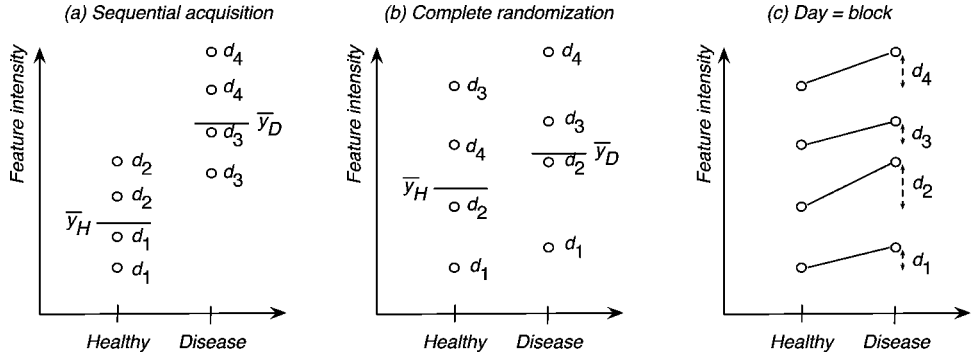


Figure 3. (a) Sequential acquisition creates a confounding effect: the difference in group means can be due to both differences between groups and differences between days. (b) Complete randomization removes the confounding effect. The variance within each group is now a combination of the biological difference and of the day-to-day variation. (c) Paired design uses day as a block of size 2. The design allows one to compare differences between individuals from two groups within a block.

biological and of the day-to-day variation, and it may be difficult to detect the true differences between the groups.

Block-randomized designs improve upon these two aspects. In the example of Figure 3c, a block-randomized design pairs two randomly selected individuals (one from the healthy group and one from the disease group) within each day, and randomly assigns the order of spectral acquisition within the pair. This enforces a *balanced* allocation of groups between days and prevents the bias. We then consider differences between the individuals in each pair. In Figure 3c, one can see that, although there is large within-group variation, pair-specific differences are consistent and point in the same direction. Thus, the block-randomized design can be more efficient in uncovering the true differences between the groups.

Even the most careful experimental design is wasted if the subsequent statistical analysis fails to take into account the structure of the data. For example, if in the block-randomized design in Figure 3c we ignore the paired allocation of individuals within a day and compare average feature intensities in each group, we do not improve our ability to detect the true differences. In general, a *statistical model* is necessary to mathematically describe the structure of the data (such as replication, randomization and blocking), and the assumptions regarding the characteristics of the experimental noise. The model will allow us to formally characterize the properties of data-derived quantities, and select the most efficient experimental design and method of statistical inference.

In proteomic experiments, we have options of working with feature intensities on the original scale, with ratios of feature intensities of samples from different disease groups, or with log-transformed intensities. It is generally believed that the true biological effects underlying these experiments are multiplicative in nature, and the use of ratios reflects this assumption. However, the ratios do not provide a natural framework for handling replication.[21] In contrast, analysis of differences of the log-transformed intensities fits naturally into the framework of *analysis of variance* (ANOVA), and there are numerous examples of successful application of ANOVA in proteomic research[22−24] as well as in the context of gene expression microarrays.[25−28]

On the logarithmic scale, the quantity of interest is $\Delta = \mu_H - \mu_D$, which represents the logarithm of the *fold change* of feature abundance, that is, the logarithm of the ratio of intensities between the groups. In the examples of Figures 1 and 3, analysis of variance will allow us to characterize the bias of $\bar{y}_H - \bar{y}_D$ as its systematic deviation from $\mu_H - \mu_D$, and precision as $Var(\bar{y}_H - \bar{y}_D)$. In the following, we will evaluate experimental designs in proteomic research according to their ability to eliminate the bias and reduce $Var(\bar{y}_H - \bar{y}_D)$ in spectral feature quantification.

3. Application of the Principles of Experimental Design

3.1. Randomization: Avoiding Bias from Sources of Undesirable Variation. Bias in experimental design occurs when healthy individuals and disease patients are selected or handled in systematically different ways, not intended by the purpose of the study. One can distinguish two sources of bias.[8] The first source is due to selection of individuals from the corresponding

populations, for example, when disease patients differ from heathy patients in age, gender, race or some other (known or unknown) important characteristics. The second source is due to systematic differences in protocols of specimen collection or spectral acquisition between the groups. For example, Banks et al.[29] studied the influence of sample processing, such as differences in anticoagulant reagent, serum collection tubes, or sample storage time. Hu et al.[30] discuss how a change of experimental protocol in the middle of a study, sample degradation and differences in spectral acquisition time biased the results.

Biases due to these sources of variation cannot be removed by increasing the sample size, or by demonstrating the reproducibility of the results in a repeated instance of the same workflow. Instead, randomization should be incorporated whenever possible into the experimental design during both selection of individuals and data acquisition.

Randomized selection of individuals is easiest in the case of a designed experiment where the investigator has full control of group membership of each individual. For example, in an experiment where rats are artificially induced with a disease, we can define a single population of rats, and randomly assign individual rats to disease or control. When assigning individuals to groups, it is important to use a random number generator. (One can imagine that, if instead of using random allocations, the experimenter assigns the first rats removed from the cage to the treatment group, rats that are slower and easier to handle will be more likely to be assigned to the treatment, thereby biasing the results.)

In observational studies that are typical in clinical research, the experimenter has no control on the disease status of an individual. Groups of patients may differ in complex ways that bias the comparison of disease versus controls. A vast literature exists on selecting patients for observational studies, and coverage of this topic is beyond the scope of this review. See, for example, Mann et al.[31] for a summary of these methods.

Allocation of the experimental material in space and time can be designed to avoid bias from sample handling and spectral acquisition. One design option is to fix known sources of experimental variation that are under the experimenter's control throughout the study. While this approach removes bias due to these sources and reduces the variation, it limits the scope of inference to this particular laboratory setting, and reproducibility of the experiment in other workflows should be independently verified.[8,9] In addition, known sources of variation should be consistently reassessed in light of constantly changing technology.

The second design option is to randomly assign the order and location of sample storage, processing and spectral acquisition to all samples in the study. The unknown sources of variation will be roughly equally distributed across groups, eliminating the bias and providing the foundation for inference in multiple laboratory settings.

When a source of variation has a systematic trend in space and time, for example, as related to location of a sample on a plate or to instrumental drift, it results in a correlation between adjacently located samples. Randomization will not eliminate the correlation, but will reduce its overall extent with increasing replications. Thus, for studies with a fairly large number of individuals, randomization allows us to apply methods of statistical inference and to assume that the measurements are nearly independent. Occasionally, randomization may provide an undesirable pattern, for example, scheduling spectral acquisition of most healthy individuals to the beginning of the experiment, and of disease individuals toward the end. This case is unlikely for studies of moderate and large size. For small studies, formal solutions have been proposed to avoid such situations,[32] but there is no final answer. In practice, the experimenter will typically discard an undesirable allocation, and select another randomization sequence.[12] As before, it is necessary to use a random number generator when assigning all random allocations.

Finally, it is possible to develop multistage experimental designs where one limits the scope of inference to control the variation in the initial stages of the investigation, and increases the scope of inference in subsequent stages of the study. This approach is increasingly used in proteomic research.[11]

Observed feature intensity	=	Systematic mean signal of disease group	+	Random deviation due to all sources of variation
y_{ij}	=	$Group\ mean_i$	+	$Error_{j(i)} \sim N(0, \sigma^2)$

Figure 4. Statistical model for a completely randomized design with a single mass spectrum replicate per patient. *i* indicates the index of a disease group, and *j*(*i*) the index of a patient within the group. All $Error_{j(i)}$ are assumed independent.

3.2. Replication: Selection of Replicate Types To Maximize Efficiency. Proteomic experiments can introduce additional levels of replication, such as multiple instances of sample preparations, and multiple mass spectrometry runs for a given sample preparation. What kind of replication will be most appropriate given a fixed number of mass spectrometry runs? When is it necessary to acquire both biological and technical replicates? We answer these questions by means of the mixed effects Analysis of Variance (ANOVA) model that describes the replicate structure of the data.

Design with no technical replicates is the simplest experimental design for a label-free experiment where one acquires a single mass spectrometry run per individual while randomly assigning the order of sample preparation and spectral acquisition. This is a completely randomized design as illustrated in Figure 3b. The statistical model corresponding to this design is described in Figure 4. If we denote σ^2 the variance of a spectral feature in our experiment, and I the number of individuals per group, then the variance of the estimated difference between disease groups 1 and 2 is

$$Var(\bar{y}_H - \bar{y}_D) = 2\sigma^2/I \quad (1)$$

Design with technical replicates contains multiple instances of sample preparations and multiple mass spectrometry runs for a same patient. The statistical model for this design is shown in Figure 5. The presence of multiple types of replicates allows us to partition σ^2 into $\sigma^2 = \sigma^2_{Indiv} + \sigma^2_{Prep} + \sigma^2_{Error}$. When the number of individuals I, sample preparations J and mass spectrum runs K is the same in both groups,

$$Var(\bar{y}_H - \bar{y}_D) = 2\left(\frac{\sigma^2_{Indiv}}{I} + \frac{\sigma^2_{Prep}}{IJ} + \frac{\sigma^2_{Error}}{IJK}\right) \quad (2)$$

One can see that $Var(\bar{y}_H - \bar{y}_D)$ in both eqs 1 and 2 are dominated by the number of individuals I. Thus, an increase in I results in a smaller variance, thereby making it easier to

$$\begin{array}{cccccccc} \text{Observed feature intensity} & = & \text{Systematic mean signal of disease group} & + & \text{Random deviation due to individual} & + & \text{Random deviation due to sample preparation} & + & \text{Random deviation due to measurement error} \\ y_{ijkl} & = & Group\ mean_i & + & Indiv(Group)_{j(i)} \sim N\left(0, \sigma^2_{Indiv}\right) & + & Prep(Indiv)_{k(ij)} \sim N\left(0, \sigma^2_{Prep}\right) & + & Error_{l(ijk)} \sim N\left(0, \sigma^2_{Error}\right) \end{array}$$

Figure 5. Statistical model for a mixed effects analysis of variance (ANOVA). *i* is the index of a disease group, *j*(*i*) the index of a patient within the group, *k*(*ij*) is the index of the sample preparation within the patient, and *l*(*ijk*) is the replicate run. $Indiv(Group)_{j(i)}$, $Prep(Indiv)_{k(ij)}$ and $Error_{l(ijk)}$ are all independent.

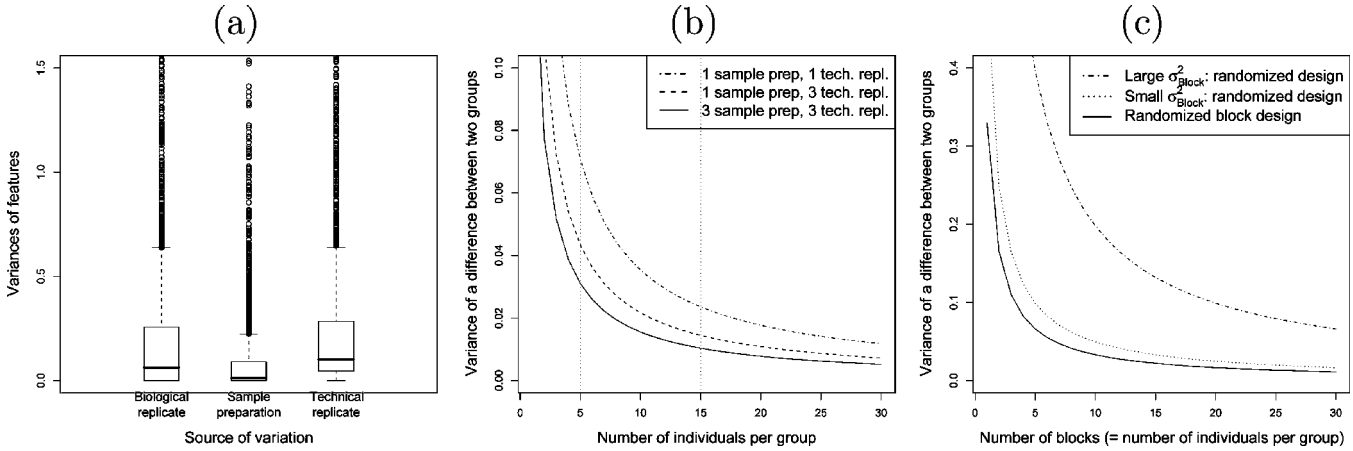


Figure 6. The pilot label-free experiment of patients with diabetes. (a) Variance components σ^2_{Indiv}, σ^2_{Prep} and σ^2_{Error} over all quantified features. Each box contains the middle 50% of the features, the horizontal line within a box is the median, dots are the outliers. (b) $Var(\bar{y}_H - \bar{y}_D)$ of a complete randomized design in eq 2. (c) $Var(\bar{y}_H - \bar{y}_D)$ for randomized block design in eq 3 and completely randomized design in eq 4, with no technical replicates. "Large" $\sigma^2_{Block} = 5(\sigma^2_{Indiv} + \sigma^2_{Error})$, and "small" $\sigma^2_{Block} = 0.5(\sigma^2_{Indiv} + \sigma^2_{Error})$.

detect true differences between the means. In eq 2, an increase in sample preparations *J* and mass spectrum runs *K* will only reduce a part of the variance, at the expense of an increasing number of runs. We conclude that in situations where the limiting factor is the total number of runs, a design with the maximum number of individuals and no technical replicates is the most efficient. This conclusion holds for all quantitative experiments combining measurements from biological and technical replicates. It was obtained theoretically for gene expression microarrays,[26] and empirically in the context of iTRAQ workflow by Gan et al.[33]

3.2.1. Example. The practical impact of selecting the number and type of replicates depends on the specific values of σ^2_{Indiv}, σ^2_{Prep} and σ^2_{Error}. We illustrate this using a pilot LC-MS experiment. Specifically, plasma samples from two patients with diabetes and two normal controls were subjected to two instances of glycopeptide enrichment,[34] and three replicate LC-MS runs were performed on each sample preparation in a label-free workflow using a Qstar Pulsar I Q-TOF MS spectrometer (Applied Biosystems). Features in LC-MS profiles were determined, quantified and aligned using SpecArray software suite,[35] and variance components σ^2_{Indiv}, σ^2_{Prep} and σ^2_{Error} were estimated separately for each feature by fitting the mixed model in Figure 5. Figure 6a shows the distributions of the three variance components over all features. The median values of the variances are $\sigma^2_{Indiv} = 0.0621$, $\sigma^2_{Prep} = 0.0118$ and $\sigma^2_{Error} = 0.1026$. As is frequently the case, the experimental error is the largest source of variation in this system. The combined median values of the technical variances are 1.84 times larger than the median of the biological variation.

Figure 6b demonstrates that the most substantial decrease in variance in a future study will be obtained by increasing the overall number of patients in each group. For example, an experimental design with 5 individuals per group and 3 technical replicates (a total of 15 mass spectrometry runs per group) will result in a larger value of $Var(\bar{y}_H - \bar{y}_D)$ than a design which allocates all 15 runs to 15 distinct individuals. The technical replicates are most helpful when working with small sample sizes. If the number of biological samples is fixed at 5 and can not be increased, three technical replicates of each sample will reduce the variance by more than a half. Since σ^2_{Prep} in this experiment is relatively small, the additional decrease in variance due to sample preparation is also small. Supporting Information contains figures similar to Figure 6b, obtained with other ratios of experimental versus biological variation.

3.3. Blocking: Reducing Variance Due to Known Sources of Undesirable Variation. When sources of undesirable variation are known but cannot be fixed throughout the experiment, one can improve the efficiency of the design by taking these sources into account. For example, changes in technical support, sample processing plates, separation columns and days of spectral acquisition create *experimental blocks*, that is, groups of mass spectrometry runs that are potentially more homogeneous within than between groups. In a labeling workflow, the labeling process introduces an additional blocking factor in that the samples that are labeled with different reagents and mixed together undergo the same MS experimental procedures. Observations on samples undergoing MS simultaneously (i.e., in one block) are more similar than observations on samples undergoing MS separately. Thus, one can assess the difference between disease groups more efficiently within each block than between blocks.

The statistical model in the presence of blocking variables is described in Figure 7. The model makes explicit the contributions of different sources of variation, and decomposes the total variation σ^2 from eq 1 into $\sigma^2 = \sigma^2_{Block} + \sigma^2_{Indiv} + \sigma^2_{Error}$. Although the model is valid generally, the variance of a comparison between disease groups depends upon the specific

Observed feature intensity	=	Systematic mean signal of disease group	+	Random deviation due to block (e.g. plate or day)	+	Random deviation due to individual	+	Random deviation due to measurement error
y_{ijkl}	=	$Group\ mean_i$	+	$Block_k$ $\sim N\left(0, \sigma^2_{Block}\right)$	+	$Indiv(Group)_{j(i)}$ $\sim N\left(0, \sigma^2_{Indiv}\right)$	+	$Error_{l(ijk)}$ $\sim N\left(0, \sigma^2_{Error}\right)$

Figure 7. Statistical model for a mixed ANOVA with random blocks. *i* is the index of a disease group, *j*(*i*) is the index of a patient within the group, *k* is the index of the block, and *l*(*ijk*) is the replicate run. $Block_k$, $Indiv(Group)_{j(i)}$ and $Error_{l(ijk)}$ are all independent.

layout of allocations of individuals to blocks. The goal of experimental design is therefore to optimally allocate treatment groups and subjects to blocks, in order to remove the bias and minimize the variances of comparisons of interest.

We start the discussion of strengths and weaknesses of blocking in multigroup experiments using a simple example of a label-free experiment. We will then provide examples of more complex designs that are often applicable to labeling workflows. According to the discussion in Section 3.2, we will assume that the experiment has no technical replicates, that is, spectra from each individual are acquired in a single MS run.

3.3.1. Blocking in Label-Free Workflows. In a label-free workflow, it is often possible to identify as blocks experimental units that contain an equal number of individuals per disease group. For example, such units can be batches of sample processing, sample plates, or a person manipulating the plates. When a source of variation has a continuous temporal or spacial trend, blocks can be selected somewhat arbitrarily as convenient. For example, in the case of instrumental drift in time, blocks can be defined as sets of adjacent runs containing an equal number of individuals from each disease group.

Randomized complete block design includes one sample from each disease group in a block. A block can be generally viewed as an independent replicate of the experiment (or nearly independent when blocking by spectral acquisition time, as discussed in Section 3.2). Randomized sample allocations are still necessary for this design; however, blocking imposes a restriction upon the randomization: we randomize the allocation of individuals to runs separately within each block, but not between blocks.

According to the model in Figure 7, the variance of the difference between any two groups is[14]

$$Var(\bar{y}_H - \bar{y}_D) = 2\left(\frac{\sigma^2_{Indiv} + \sigma^2_{Error}}{I}\right) \quad (3)$$

where *I* denotes the total number of individuals in a group. To make the comparison with the completely randomized design, we can rewrite the variance of a comparison in eq 1 as

$$Var(\bar{y}_H - \bar{y}_D) = 2\left(\frac{\sigma^2_{Block} + \sigma^2_{Indiv} + \sigma^2_{Error}}{I}\right) \quad (4)$$

Thus, the advantage of the block design is that it reduces the variance in eq 4 by removing σ^2_{Block}.

The practical advantage of the block design depends upon the relative importance of σ^2_{Block}, and on the number of individuals *I* in each group. When σ^2_{Block} is large as compared to $\sigma^2_{Indiv} + \sigma^2_{Error}$, blocking is more efficient since it minimizes the variance of the comparison. However, when σ^2_{Block} is small, blocking does not produce a substantial reduction in variation. Moreover, statistical theory indicates that in this situation blocking is inefficient due to the loss of the effective number of observations (i.e., *degrees of freedom*). For example, analysis in Figure 3c studies 4 differences within each block, that is, the effective number of 4 observations, as opposed to 8 observations in Figure 3b. When not compensated by a strong reduction in variance, the loss of degrees of freedom undermines the efficiency of the design.[12] Finally, blocking becomes irrelevant when *I* is large.

3.3.1.1. Example. We illustrate the efficiency of blocking using the diabetes pilot study in Section 3.2. We set σ^2_{Indiv} and σ^2_{Error} to the medians of the experimental values as before, and investigate the use of instances of sample preparation (or any other experimental step) as experimental blocks. We assume for simplicity that each block contains one individual from each disease group, and consider two hypothetical scenarios of between-block variation: "small" $\sigma^2_{Block} = 0.5(\sigma^2_{Indiv} + \sigma^2_{Error})$, and "large" $\sigma^2_{Block} = 5(\sigma^2_{Indiv} + \sigma^2_{Error})$. Figure 6c displays the variances in eqs 3 and 4 for these experimental configurations, and demonstrates that blocking can increase the efficiency in the case of large between-block variance and moderate experiment size. In our diabetes study, the median variance $\sigma^2_{Prep} = 0.07(\sigma^2_{Indiv} + \sigma^2_{Error})$. This small variance indicates that blocking by sample preparation is not necessary in this setting, unless the number of individuals is small. Supporting Information contains figures similar to Figure 6c, obtained with other ratios of experimental versus biological variation.

3.3.2. Blocking in Labeling Workflows. Stable isotope labeling workflows, and also multichannel gels, combine samples from multiple individuals within the same run. This introduces an additional blocking factor. For example, with workflows such as ICAT, measurements on samples undergoing MS simultaneously are more similar than measurements on samples undergoing MS in separate runs. With iTRAQ, measurements obtained within a single MS/MS spectrum are more comparable than measurements obtained across different MS/MS spectra and runs (for simplicity of presentation, in the following we will assume that a single MS/MS spectrum is used to quantify the abundance of a peptide in each run).

MS runs (and MS/MS spectra) form blocks of relatively small size, that is, the number of disease groups that can be jointly allocated within a run is relatively small as compared to the number of the groups. Thus, it may be impossible to include individuals from all disease groups in one block. The goal of an experimental design is to allocate individual samples to labeling reagents and runs in a way that avoids bias, and reduces the variance of comparisons between groups. A variety of such designs exist. They all proceed by systematically creating a minimal *replicate set*, that is, a set of blocks that contains the minimal number of individuals from all groups, and repeat the sets multiple times for experiments of larger size. These designs are typically used with more than two groups, and estimates of μ (in notation of Figure 1) may differ from the sample mean $\bar{y}$. Thus, we will denote the variance of a comparison between two disease groups as $Var(\hat{D}_1 - \hat{D}_2)$.

(a) Randomized Complete Block

Disease group	Replicate set 1 Block 1	Replicate set 2 Block 2	⋯
D_1	X	X	⋯
D_2	X	X	⋯
D_3	X	X	⋯
D_4	X	X	⋯

(b) Balanced Incomplete Block

Disease group	Replicate set 1 Block 1	Block 2	Block 3	Block 4	Block 5	⋯
D_1	X	X	X	X		⋯
D_2	X	X	X		X	⋯
D_3	X	X		X	X	⋯
D_4	X		X	X	X	⋯
D_5		X	X	X	X	⋯

Figure 8. Experiments with a four-label workflow; "X" indicates a unique biological sample. (a) Randomized complete block design with four disease groups: each block contains one individual from each disease group. (b) Balanced incomplete block design with four labels and five disease groups: individuals from each pair of groups appear in the same block an equal number of times.

(a) Balanced Incomplete Block

Disease group	Replicate set 1 Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8	Block 9	Block 10	⋯
D_1	X_{L_1}	X_{L_2}	X_{L_1}	X_{L_2}							⋯
D_2	X_{L_2}				X_{L_1}	X_{L_2}	X_{L_1}				⋯
D_3		X_{L_1}			X_{L_2}			X_{L_1}	X_{L_2}		⋯
D_4			X_{L_2}			X_{L_1}		X_{L_2}		X_{L_1}	⋯
D_5				X_{L_1}			X_{L_2}		X_{L_1}	X_{L_2}	⋯

(b) Reference

Disease group	Replicate set 1 Block 1	Block 2	Block 3	Block 4	Block 5	⋯
R	R_{L_1}	R_{L_1}	R_{L_1}	R_{L_1}	R_{L_1}	⋯
D_1	X_{L_2}					⋯
D_2		X_{L_2}				⋯
D_3			X_{L_2}			⋯
D_4				X_{L_2}		⋯
D_5					X_{L_2}	⋯

(c) Loop

Disease group	Replicate set 1 Block 1	Block 2	Block 3	Block 4	Block 5	⋯
D_1	X_{L_1}				X_{L_2}	⋯
D_2	X_{L_2}	X_{L_1}				⋯
D_3		X_{L_2}	X_{L_1}			⋯
D_4			X_{L_2}	X_{L_1}		⋯
D_5				X_{L_2}	X_{L_1}	⋯

Figure 9. Five-group experiments with a two-label workflow; "X" indicates a unique biological sample, "R" indicates a reference sample, and "L_1" and "L_2" indicate the (optional) systematic labeling scheme. (a) Balanced incomplete block: individuals from each pair of groups appear in a same block once. (b) Reference design: each block contains a reference sample which is the same in all blocks, and one additional unique individual. (c) Loop design: pairs of individuals from different groups cycle through blocks.

The designs differ in exactly how one allocates individuals to blocks. The following two designs are generally used, and are particularly appropriate when the block size (i.e., the number of labels in the experiment) is moderate or large.

Randomized complete block design is described in the previous section, and is applicable in labeling workflows where the number of labels equals the number of disease groups. For example, Figure 8a illustrates a randomized complete block design in a four-label workflow studying four disease groups. Since one individual from each disease group can be included in a block, each block corresponds to its own independent replicate set. The variance of comparisons between groups $Var(\hat{D}_1 - \hat{D}_2)$ in this case is as in eq 3, and I in the formula can be interpreted simultaneously as both the number of individuals per group and the number of minimal replicate sets.

Balanced incomplete block design is applicable when the number of disease groups exceeds the number of individuals in a block. This is the case, for example, of a two-label experiment studying three or more disease groups, four-label experiment studying five or more groups, and so on. A minimal replicate set of the design allocates individuals from all pairs of disease groups in a same block an equal number of times. It is possible to either randomize, or to systematically rotate, the label allocations within each block. We illustrate this design with an example of a five-group experiment using a four-label workflow (Figure 8b), and a two-label workflow (Figure 9a). The variance of comparisons between any pair of groups is[12]

$$Var(\hat{D}_1 - \hat{D}_2) = 2\frac{n_b}{n_g n_p n_s}(\sigma^2_{Indiv} + \sigma^2_{Error}) \quad (5)$$

where n_b is the block size (i.e., the number of labels), n_g is the number of groups, n_p is the number of times that individuals from a given pair of disease groups occur in a same block within a minimal replicate set, and n_s is the number of minimal replicate sets.

When the block size is small, such as in the case of two-label workflows, a variety of more specialized designs can be considered. These designs were introduced for use with two-color gene expression microarrays,[21,26] and were extensively evaluated in subsequent publications. We illustrate these designs using two examples of label allocations. More details regarding these designs, as well as variance calculations for specific experimental workflows can be found in Dobbin and Simon.[28] Woo et al.[36] extend these designs to the 3- and 4-label microarrays.

Reference design controls for between-block variation by means of a common reference sample allocated to all blocks. The reference sample itself is generally not of interest, but helps eliminate the between-block variation when comparing the remaining samples and groups. It is possible to randomly assign the labels within a block, or to always allocate the same label to the reference sample in experiments of small size. An example of the reference design in a four-group and two-label experiment is illustrated in Figure 9b, and the variance of comparison between any two nonreference groups is

$$Var(\hat{D}_1 - \hat{D}_2) = \frac{2}{I}(\sigma^2_{Indiv} + 2\sigma^2_{Error}) \quad (6)$$

where I can be interpreted as both the number of individuals per group and the number of minimal replicate sets. This expression is general and does not depend on the number of groups under investigation. Kerr et al.[37] found that the choice of the reference sample in the case of gene expression microarrays has no practical impact on the efficiency of the design.

Loop design is an alternative to the reference design. The design cycles samples through the blocks in a systematic manner as shown in Figure 9c. It is possible to either randomize, or to systematically rotate, the label allocations within each block. The general formula of the variances of a comparison

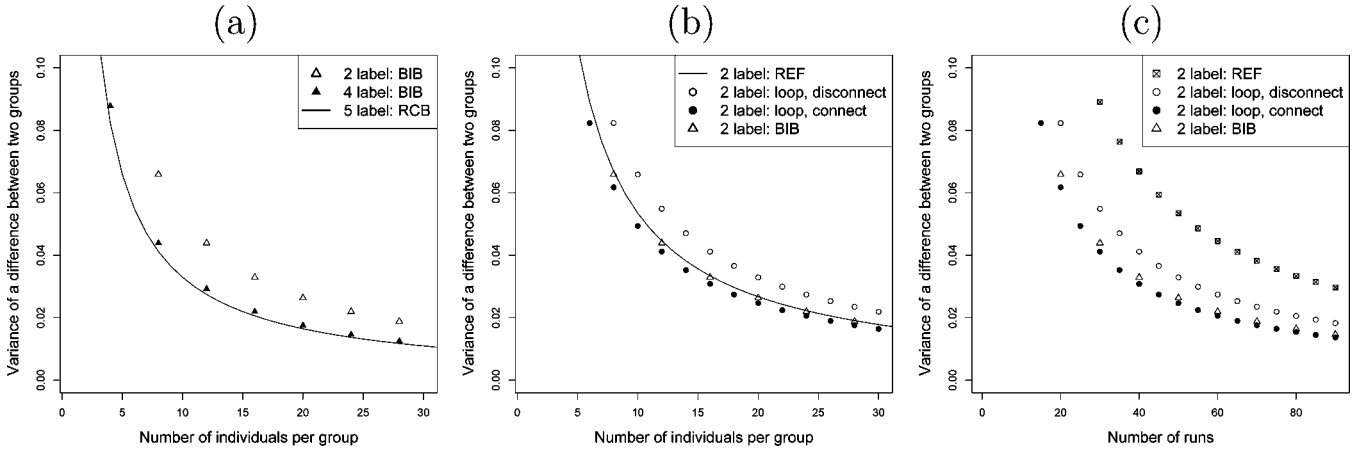


Figure 10. Variances $Var(\hat{D}_1 - \hat{D}_2)$ of a comparison between two disease groups in a 5-group experiment. (a) A 5-label workflow with a randomized complete block design in eq 4, and 4- and 2-label workflows with a balanced incomplete block design in eq 5. (b and c) A 2-label workflow with a balanced incomplete block in eq 5, reference design in eq 6 and loop design in eqs 7 and 8.

between two groups is more complicated, and depends on both the number of disease groups, and the number of blocks that separate the two disease groups under comparison in the design. In the special case of 5 disease groups in Figure 9c, the variances are

$$Var(\hat{D}_1 - \hat{D}_2) = \frac{8}{5n_s}(\sigma^2_{Indiv} + \sigma^2_{Error}) \quad (7)$$

if two disease groups are "connected" in the same block, and

$$Var(\hat{D}_1 - \hat{D}_3) = \frac{12}{5n_s}(\sigma^2_{Indiv} + \sigma^2_{Error}) \quad (8)$$

if two groups are "disconnected", i.e. are one block apart. When more blocks separate the two groups of interest in the design, the variance of the comparison increases even more.

Each of these designs can be useful in some experimental circumstances. In particular, each design only exists for a certain combination of number of disease groups, number of runs, and block size. For example, in the case of a 4-label workflow, the randomized complete block design only exists for studies of at most 4 disease groups, and the balanced incomplete block with 5 groups will require a number of biological replicates that is an exact multiple of 4 (Figure 8). Similar constraints apply in the case of a 2-label workflow.

In an experimental situation where multiple designs can be used, a good design minimizes the variances in eqs 3−8. However, one should also consider other features of the designs, such as the total number of MS runs, and robustness of the design to failure or loss of some runs. The randomized complete block design, when feasible, requires the smallest number of runs, and is the most robust to run failures. The loss of a run effectively amounts to the loss of one set of biological replicates, but it will not otherwise affect the balanced structure of the experiment and the variance of the comparison in eq 3. Both balanced incomplete block and loop designs require more runs, and are not robust to run failures. The loss of a run destroys the balanced structure of a minimal replicate set, and will result in an increase in the variance of the comparison. An additional drawback of the loop design is that it is not equally precise when comparing pairs of disease groups, and one must allocate samples from comparisons of primary interest into neighboring blocks. Finally, the reference design is robust to run failures and is easy to implement. However this simplicity comes at the price of an excessive number of runs, and half of the resources are spent on acquisition of spectra for a reference sample that is not of interest.

3.3.2.1. Example. Figure 10 illustrates the performance of the designs in an experiment with 5 disease groups, assuming the same values of the variance components as in Section 3.2. Figure 10a compares the variances of a hypothetical 5-label workflow which can accommodate all 5 disease groups in a randomized complete block design (e.g., in the case of 8-label iTRAQ where only 5 labels are used) with the balanced incomplete block designs for 4- and 2-label workflows. The figure illustrates the advantage of larger experimental blocks. The 5-label workflow, if feasible, produces the smallest variance of a comparison, and exists for any number of biological replicates. The least efficient design is a 2-label workflow. Both 4-label and 2-label workflows exist fully replicated only for a number of biological replicates that is a multiple of 4.

Figure 10b compares balanced incomplete block, loop and reference designs in a situation where the 2-label workflow is the only possible choice. One can see that the balanced incomplete block design is an optimal choice since it minimizes the variance of the comparison, however the design does not always exist. The reference design exists for any number of biological replicates, but results in a slightly larger variance. The variance of a comparison for the loop design is similar to that of the balanced incomplete block when the disease groups are connected (i.e., appear in the same block), but closer to the reference design when the disease groups are disconnected (i.e., one block apart). Figure 10c further compares these designs in terms of the number of runs. Although the variance of a disconnected pair of groups is similar to that of the reference design, the reference design is less efficient in that it requires a larger number of runs. Supporting Information contains figures similar to Figure 10b,c, obtained with a series of other ratios of experimental versus biological variation.

The variances in Figure 10 are theoretical in that they assume that all workflows have the same σ^2_{Indiv} and σ^2_{Error}. While adding more labels increases the efficiency and power of the experiment, in practice, the increase in efficiency can be offset by a

reduced dynamic range and increased variability in the multilabel system.[36] Thus, in order to fully compare various experimental designs, variances used in Figure 10 should be estimated separately for each workflow from the corresponding prior experiments.

3.3.3. Additional Comments. The designs above can be extended to handle multiple and inter-related sources of variation. For example, one can specify several types of blocks, and combine blocking with technical replication. Alternatively, it may be necessary to account for a statistical interaction between disease group and block. Another extension can be required in a labeling workflow in the presence of unequal efficiency of labeling reagents across features, a case documented and analyzed in studies of gene expression.[16,38] Most extensions will change the corresponding ANOVA models, and affect the expressions of $Var(\hat{D}_1 - \hat{D}_2)$. However, the principles of replication, randomization and blocking will always apply.

The higher the complexity of the ANOVA model, the more replicates are necessary to estimate the associated terms and variance components from the data. Therefore, when the number of available replicates is small, this limits the feasible statistical models and designs. Several strategies can be pursued when working with a small sample size. First, it is important to conduct pilot studies in a similar setting to identify the major sources of variation, and the sources that can be discarded. Second, if a major source of variation is identified, it is preferable to keep this source fixed throughout the experiment, as discussed in Section 3.1. While this strategy limits the scope of inference, it simplifies the model and improves our ability of detecting differences. Third, when a major source of variation cannot be fixed, randomization can produce suboptimal allocations in experiments with small sample size. For example, in a labeling workflow, it can unintentionally assign more samples from a disease group to a particular label, and introduce a bias as discussed in Section 2. This can be avoided by enforcing a systematic allocation of individuals to all labels as illustrated in Figure 9a,c, and discussed in Dobbin et al.[16,38] The allocation will not reduce $Var(\hat{D}_1 - \hat{D}_2)$, but will help avoid bias. Finally, Empirical Bayes ANOVA combines information on feature-specific variances over all features to improve the power of detecting a fold change in small experiments. Such models have been proposed for gene expression microarrays,[39,40] and are made available for proteomic research, for example, through the Corra framework.[41]

It is important to note that these strategies provide only a partial remedy to the problem of a small sample size, and will not substitute the insight that can be gained from increasing the number of biological replicates.

4. Pooling: Reducing Biological Variation and Number of Runs

An experimental design can require decisions regarding pooling biological specimens from the same disease group prior to mass analysis. Pooling is often considered out of necessity, for example, in order to increase the volume of the sample to achieve proper performance of the assay, or to reduce the overall cost of the experiment. Another potential motivation of pooling is to improve the ability of finding differences between groups by reducing the overall subject-to-subject variation. One can envision two scenarios: combining all samples from a disease group into a single pool and combining several samples at a time to produce several pools. As always in questions of experimental design, pooling strategies should be characterized on the basis of bias and variance of the comparisons between disease groups.

First, combining all subjects from a disease group in a single pool seriously undermines the usefulness of the experiment. This strategy results in a single measurement per spectral feature per disease group, as in the situation shown in Figure 2a. Such pooling makes it impossible to assess the underlying variability in an experiment without technical replicates (σ^2 in eq 1), or in the case of multiple technical replicates (σ^2_{Indiv} in eq 2). As we have seen, an experimental design that does not allow characterization of the biological variation can not produce a valid inference, and should not be done.

An alternative strategy involves creating multiple pools from different subsets of individuals. If each subject belongs to a single pool, the pools can be viewed as independent biological replicates, and the general inferential procedure in Figure 1 applies. However, feature quantification from the pooled samples is susceptible to multiple sources of bias. The first obvious source is due to technical aspects of the experiment. For example, pipetting errors may cause specimens to have unequal contributions to the pool, resulting in a biased signal. It is also more difficult to detect outlying or contaminated samples. If contamination in a specimen is not detected prior to mixing, the entire pool is affected and it will be difficult to determine which member of the pool was affected. The second, less obvious source of bias, is due to the choice of the measurement scale used at the statistical analysis stage. As discussed in Section 2, the true biological measurements are generally viewed as multiplicative in nature, and peak intensities resulting from quantitative experiments are log-transformed prior to statistical analysis. Yet physically mixing the samples amounts to averaging the signal on the raw scale, and the averages on these two scales are not equivalent. To summarize, the application of pooling strategy is based on the *assumption of biological averaging*, that is, on the assumption that the sources of bias listed above are small and can be neglected.

If the assumption of biological averaging holds, there is a theoretical advantage in creating multiple groups of pools for class comparison. Consider, for example, a label-free LC-MS experiment and the statistical model in Figure 5, and assume for simplicity that a single sample preparation and a single technical replicate is acquired per biological mixture. In the absence of pooling, the variance of a comparison of feature intensities between groups, where each group has I individuals, is

$$Var(\bar{y}_H - \bar{y}_D) = 2\left(\frac{\sigma^2_{Indiv} + \sigma^2_{Prep} + \sigma^2_{Error}}{I}\right) \quad (9)$$

If instead each biological mixture is a pool of r individual specimens, the corresponding variance is

$$Var(\bar{y}_H - \bar{y}_D) = 2\left(\frac{\sigma^2_{Indiv}}{Ir} + \frac{\sigma^2_{Prep} + \sigma^2_{Error}}{I}\right) \quad (10)$$

Thus, increasing the number of biological replicates by pooling allows one to reduce the variance of the comparison, without changing the total number of runs. The theoretical advantage of pooling has been studied extensively in the context of gene expression microarray.[42−44] In particular, Zhang and Gant[44] calculated efficiencies of pooling designs while incorporating

Table 1. Outcomes of Testing the Null Hypothesis $H_0 : \mu_H = \mu_D$ for a Single Experimental Feature

	no detected difference	detected difference
true equal abundance		type I error
true different abundance	type II error	

the analysis of associated costs in terms of money, time, or other resources, and developed an online application that performs on-demand calculations. They concluded that the largest gain from a pooled design with respect to the cost is when the cost per subject is relatively low as compared to the cost per assay.

The practical utility of pooling in a proteomic experiment will depend on the extent to which the assumptions of biological averaging holds. Shih et al.[45] and Kendziorski et al.[46] empirically evaluated this assumption in the context of gene expression. They found that the measurement scale affected approximately 25% of the genes in their experiment, and different conclusions regarding differential expression of these genes were made in presence or in absence of pooling. Additional work is needed to evaluate the appropriateness of the biological averaging assumption in the context of various proteomic workflows.

Our final comment is that pooling limits the type of analysis that can be performed on the resulting spectra. Although combining samples into pools does not prevent class comparison with ANOVA, it prevents investigations that involve class discovery and class prediction that are necessary, for example, in the context of diagnostics and prognostics. Such analyses cannot be performed without a quantitative measurement for each separate individual.

5. Sample Size Calculations for a New Experiment

An important aspect of experimental design is calculation of the number of biological and technical replicates necessary for a future study. The number of replicates should be relatively large to ensure that the experiment can detect important differences with a high enough probability. It should also be relatively small to avoid prohibitively large costs. Sample size calculations depend on a series of experimental characteristics which include allocation of resources according to the previously described designs, anticipated difference in feature intensities, and the overall number of features. In the following, we discuss sample size calculations for one feature, and then extend the calculations to the more realistic multiple-feature situation.

5.1. Sample Size Calculation for Comparing Group Means of One Feature. The goal of a comparison between two groups is to test the null hypothesis of equality of average peak intensities between healthy and disease patients $H_o : \mu_H = \mu_D$ (where μ_H and μ_D are the population means described in Figure 1), versus the alternative hypothesis $H_a : \mu_H \neq \mu_D$. Results of the hypothesis testing procedure belong to one of the four scenarios summarized in Table 1. We define the significance level of the test as $\alpha = \text{Prob}\{\textit{making Type I error}\}$, and the power of the test as the probability of obtaining the evidence supporting the research, that is, $1 - \beta = 1 - \text{Prob}\{\textit{making Type II error}\}$.

To calculate the sample size, we fix the significance level α and power $1 - \beta$ of the test at the desired levels, and specify the smallest difference between population means $\Delta = |\mu_H - \mu_D|$ that we would like to detect. Finally, we need to have access to information from previous investigations conducted under similar physiological and experimental conditions to calculate representative values of σ^2_{Indiv} and σ^2_{Error}. Separate estimations of σ^2_{Indiv} and σ^2_{Error} are only possible in pilot experiments containing technical replicates. Experiments with a single technical replicate provide a joint estimate of $\sigma^2_{Indiv} + \sigma^2_{Error}$; however, this is sufficient for many planned designs.

When a large number of biological replicates is anticipated, the following formula[14] approximately relates Δ and $Var(\hat{D}_1 - \hat{D}_2)$:

$$Var(\hat{D}_1 - \hat{D}_2) \leq \left(\frac{\Delta}{z_{1-\beta} + z_{1-\alpha/2}}\right)^2 \tag{11}$$

where $z_{1-\beta}$ and $z_{1-\alpha/2}$ are respectively the $100(1-\beta)$th and the $100(1-\alpha/2)$th percentiles of the standard Normal distribution. The formula can be applied to variances of comparisons from various experimental designs (such as in eqs 1−8), and solved to determine the minimal sample size. In particular, for the simple two-group comparison in eq 1, the number of biological replicates per disease group I is calculated as

$$I \geq 2\left(\frac{z_{1-\beta} + z_{1-\alpha/2}}{\Delta/\sigma}\right)^2 \tag{12}$$

When the anticipated sample size of an experiment is small, the quality of the approximation in eq 11 can be poor. An alternative iterative computational procedure based on the Student distribution can be used instead. The procedure has been described,[12] and its implementation is available from a variety of statistical software systems such as SAS and R. Extensive examples, including SAS code, are given in ref 47. Finally, an alternative procedure for sample size calculations involves computer simulations, whereby one generates synthetic data representing a variety of biological conditions and experimental designs, and observes the frequency of type I error and the power of the test. Such simulations have been applied in the context of proteomic experiments in ref 48.

5.1.1. Example. We use the label-free diabetes data set to calculate the sample size for a single feature in a future experiment. We set the probability of type I error to $\alpha = 0.05$ and the power of the test to $1 - \beta = 0.8$, and vary Δ between 0.1 and 0.4. Because of the logarithm transformation, the fold change of the signal on the raw scale is defined as e^{Δ} (or 2^{Δ} when using logarithm with base 2).

Figure 11a illustrates the sample size calculation with different types of replicates discussed in Section 3.2, using $Var(\bar{y}_H - \bar{y}_D)$ in eq 2. We fix the number of sample preparations J and technical runs K for each patient to 1 or 3, set σ^2_{Indiv}, σ^2_{Prep} and σ^2_{Error} to the medians of the experimental values as in Section 3.2, and solve the approximate formula in eq 11 for the number of individuals per group I. One can see that if the number of runs is fixed, allocating all the runs to the biological replicates allows one to detect the smallest fold change. On the other hand, if the number of individuals cannot be increased, multiple technical replicates will also help reduce the detectable fold change, but to a smaller extent and at the expense of the number of runs. Supporting Information contains figures similar to Figure 11a, obtained with different values of experimental and biological variation.

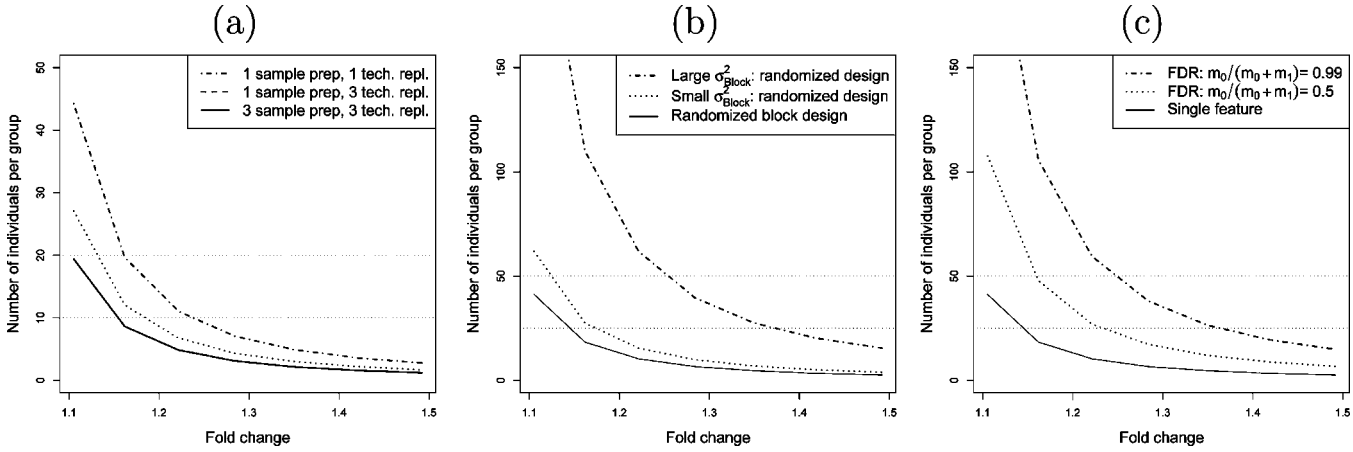


Figure 11. Number of individuals per disease group for a new label-free experiment. (a) Single feature: a completely randomized design, $\alpha = 0.05$. (b) Single feature: a randomized block design and a completely randomized design with no technical replicates. "Large" $\sigma^2_{Block} = 5(\sigma^2_{Indiv} + \sigma^2_{Error})$, and "small" $\sigma^2_{Block} = 0.5(\sigma^2_{Indiv} + \sigma^2_{Error})$, $\alpha = 0.05$. (c) Multiple features: a randomized block design as in (b), $q = 0.05$.

Figure 11b illustrates the sample size calculation in the presence of blocking, using "small" $\sigma^2_{Block} = 0.5(\sigma^2_{Indiv} + \sigma^2_{Error})$, and "large" $\sigma^2_{Block} = 5(\sigma^2_{Indiv} + \sigma^2_{Error})$ as in Section 3.3.1. The figure further demonstrates the advantage of the block design in that it requires a smaller number of biological replicates to detect a small-to-moderate fold change, and the difference is particularly important when the between-block variation is large.

5.2. Sample Size Calculation for Comparing Group Means of Multiple Features. In proteomic experiments, one is rarely interested in changes in abundance of a single feature in the data. Instead, we are interested in comparing the abundances of a potentially large number of features which are simultaneously detected as part of the experiment. A variety of multivariate generalizations of the type I error rate and of the power of the test exist, along with the statistical procedures for their control. The use of these generalizations in the context of gene expression microarrays is reviewed in ref 49. The choice of the multivariate type I error rate and of multivariate power will affect the testing procedures, and will also affect the calculations of the sample size.

A powerful multivariate generalization of the type I error is the *False Discovery Rate (FDR)*, defined as the expected proportion of unduly detected differences in the list of rejected null hypotheses.[50] In other words, FDR is the average false positive rate that would be obtained under multiple repetitions of the same experiment. Listgarten and Emili[7] and Karp et al.[51] emphasize the importance of controlling the FDR when comparing abundances of multiple features in quantitative proteomics.

Consider an experiment where we simultaneously compare the abundance of m features, m_0 out of which do not differ in the underlying populations. The outcome of the comparison is summarized in Table 2. Conditional on the feature identification, m and m_0 can be considered fixed. However, R, S, T, U, V in the table are random variables that depend on the observed data, and only R is actually observed. FDR is defined mathematically as

$$q = E\left[\frac{V}{max(R, 1)}\right] \tag{13}$$

where E[] denotes the expected value. Benjamini and Hochberg[50] propose to control the FDR at the desired level q

Table 2. Outcomes of Testing m Null Hypotheses $H_0 : \mu_H = \mu_D$ Simultaneously for m Experimental Features, Conditionally on the Features Detected and Quantified by a Signal Processing Procedure[a]

	no. of features with no detected difference	no. of features with detected difference	total
no. true nondiff. features	U	V	m_0
no. true diff. features	T	S	$m_1 = m - m_0$
Total	$m - R$	R	m

[a] R, S, T, U and V are random quantities, but only R is observed.

according to the following procedure. First, order the p-values of the m comparisons from the largest $p_{(m)}$ (i.e., the least significant) to the smallest $p_{(1)}$ (i.e., the most significant). Next, vary j from m down to 1, and compare $p_{(j)}$ to $(j/m)q$. Once we encounter the first p-value such that $p \le (j/m)q$, we reject the null hypothesis for this test, as well as all other null hypotheses that correspond to lower p-values. It can be shown that the FDR in the resulting list of rejected null hypotheses does not exceed q. A variety of statistical software tools, including SAS and R, contain packages performing this procedure.

The Benjamini−Hochberg procedure can be used to calculate the number of biological replicates in a future experiment with multiple features, while controlling the FDR. The authors show[50] that the procedure controls the average type I error α_{ave} over all features in the experiment at

$$\alpha_{ave} \le (1 - \beta)_{ave} \cdot q \frac{1}{1 + (1 - q) \cdot m_0 / m_1} \tag{14}$$

where $(1 - \beta)_{ave}$ is the average power over all features. Sample size estimates can be made by first estimating α_{ave} using eq 14, and then calculating the sample size required to achieve α_{ave} and $(1 - \beta)_{ave}$ for a single feature according to eq 11. The procedure requires the specification of an additional quantity, the anticipated ratio m_0/m_1. An example of application of this procedure in proteomic research can be found in ref 22. An

alternative approach to estimating sample size in the case of multiple features involves computer simulation.

5.2.1. Example. We continue the example of the label-free diabetes experiment. We use the variance components discussed in Section 3.2, and assume a randomized block design with no technical replicates. We now consider the sample size necessary for a simultaneous comparison of abundance of multiple features. We set the FDR $q = 0.05$ and the average power $(1 - \beta)_{ave} = 0.8$, and vary Δ (and the corresponding fold change on the original scale) as above. In addition, we consider two experimental scenarios. The first corresponds to a relatively large proportion of features that do not change in abundance between two disease groups, where $m_0/(m_0 + m_1) = 0.99$. This translates into $m_0/m_1 = 99$, and $\alpha_{ave} = 0.0004$. The second scenario corresponds to a moderate proportion of unchanging features where $m_0/(m_0 + m_1) = 0.5$. This translates into $m_0/m_1 = 1$, and $\alpha_{ave} = 0.0205$. Thus, the simultaneous testing requires a more conservative α_{ave} when controlling FDR at $q = 0.05$.

Figure 11c displays the number of biological replicates per disease group for the label-free randomized block design. A more conservative α_{ave} results in a larger sample size. Studies with larger proportions of unchanging features require a larger number of biological replicates to control the False Discovery Rate.

6. Discussion

Statistical design of experiments is unfortunately often mistaken for calculations of the number of replicates. This notion is incorrect since, as we have seen, different experimental designs can yield very different power of a comparison, given the same number of biological replicates or runs. Thus, sample size calculation is only one out of many aspects of planning a future experiment.

Experiment planning should start by clearly stating the scientific question of interest, and identifying the population(s) of interest that will allow one to answer the question. Although this sounds obvious, failure to state the question and to identify the underlying populations is one of the most common causes of inadequate study designs. The next step is to translate the question of interest into specific statistical hypotheses, for example, by specifying which comparisons between which groups will be considered. With these comparisons in mind, one should determine a procedure for recruiting individuals into the study that ensures an unbiased and accurate representation of the underlying populations.

Once the biological specimens are obtained, a decision should be made regarding their allocation in space and time at different stages of sample processing and spectral acquisition. As we have seen, allocation strategies should be judged on the basis of a statistical model that reflects the similarities and differences in the resulting spectra. Choice of the appropriate statistical model is one of the most difficult aspects of the design, and should be based on the detailed knowledge of the experimental procedure, and on quantitative data from pilot studies or from previous similar experiments.

When selecting a model, one should consider all potential sources of experimental variation, use data from previous experiments to calculate variance components such as σ^2_{Indiv} and σ^2_{Error}, and ensure that both the statistical model and experimental design account for the major sources of variation. We would like to emphasize that variance components vary between experimental workflows and laboratories, as well as according to computational signal processing tools that are subsequently applied, and should be calculated anew for each experimental setting. Moreover, more than one statistical model can be potentially appropriate. For example, an alternative to the model in Figure 5 can contain both technical replicates and blocking factors; an alternative to the model in Figure 7 can relax the additive structure, and contain a statistical interaction between disease groups and blocks. Data from previous experiments should be used to select a model that is most appropriate for each experimental workflow.

Finally, all the statistical models are based on distributional assumptions, such as normality of the random quantities involved. Applying a logarithm transformation to peak intensities is usually necessary to ensure that these assumptions are plausible. Considering the log transform at the design stage is important since variance components and model choice can be dramatically different when analyzing the data on the raw or on the log scale. Once one or several candidate models are set up, one proceeds with sample size calculations such as in Figure 11 in order to select the most desirable design.

In summary, many mistakes can be avoided if experimentalists work with statisticians at the early design of experiment stage, prior to collecting biological specimens and acquiring data. Inviting a statistician when the data are already collected may be too late to correct these mistakes.

Acknowledgment. The authors would like to thank Dr. Julian Watts (Institute for Systems Biology, Seattle, WA) for providing the LC-MS data for the human diabetes analysis, and for proof-reading the manuscript.

Supporting Information Available: Additional procedures and figures. 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) Goshe, M. B.; Smith, R. D. *Curr. Opin. Biotechnol.* **2003**, *14*, 101–109.
(3) Yan, W.; Chen, S. S. *Briefings Funct. Genomics Proteomics* **2005**, *1*, 27–38.
(4) Domon, B.; Aebersold, R. *Science* **2006**, *312*, 212–217.
(5) Mueller, L. N.; Brusniak, M.-Y.; Mani, D. R.; Aebersold, R. *J. Proteome Res.* **2008**, *7*, 51–61.
(6) Gillette, M. A.; Mani, D. R.; Carr, S. A. *J. Proteome Res.* **2005**, *4*, 1143–1154.
(7) Listgarten, J.; Emili, A. *Mol. Cell. Proteomics* **2005**, *4*, 419–434.
(8) Ransohoff, D. F. *Nat. Rev.* **2005**, *5*, 142–149.
(9) Ransohoff, D. F. *J. Natl. Cancer Inst.* **2005**, *97*, 315–319.
(10) Coombes, K. R.; Morris, J. S.; Hu, J.; Edmonson, S. R.; Baggerly, K. A. *Nat. Biotechnol.* **2005**, *23*, 291–292.
(11) Rifai, N.; Gillette, M. A.; Carr, S. A. *Nat. Biotechnol.* **2006**, *24*, 971–983.
(12) Kutner, M. H.; Nachtsheim, C. J.; Netter, J.; Li, W. *Applied Linear Models*; McGraw-Hill/Irwin: Columbus OH, 2005.
(13) Montgomery, D. C. *Design and Analysis of Experiments*,6th ed.; Wiley: New York, 2005.
(14) Rao, P. V. *Statistical Research Methods in the Life Sciences*; Brooks/Cole Publishing Company: Pacific Grove, CA, 1998.
(15) Simon, R.; Radmacher, M. D.; Dobbins, K. *Genetic Epidemiol.* **2002**, *23*, 21–36.
(16) Dobbin, K.; Shih, J. H.; Simon, R. *J. Natl. Cancer Inst.* **2003**, *95*, 1362–1369.
(17) Churchill, G. A. *Nat. Genet.* **2002**, *32 Suppl*, 490–495.
(18) Yang, Y. H.; Speed, T. *Nat. Rev. Genet.* **2002**, *3*, 579–588.
(19) Dupuy, A.; Simon, R. M. *J. Natl. Cancer Inst.* **2007**, *99*, 147–157.
(20) Fisher, R. A. *The Design of Experiments*, 5th ed.; Oliver and Boyd: Edinburgh, 1937.
(21) Kerr, M. K.; Martin, M.; Churchill, G. A. *J. Comput. Biol.* **2000**, *7*, 819–837.
(22) Patil, S. T.; Higgs, R. E.; Brandt, J. E.; Knierman, M. D.; Gelfanova, V.; Butler, J. P.; Downing, A.-C. 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.
(23) Daly, D. S.; Anderson, K. K.; Panisko, E. A.; Purvine, S. O.; Fang, R.; Monroe, M. E.; Baker, S. E. *J. Proteome Res.* **2007**, *7*, 1209–1217.
(24) 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, H. R. *J. Proteome Res.* **2008**, *7*, 225–233.
(25) Wolfinger, R. D.; Gibson, G.; Wolfinger, E. D.; Bennett, L.; Hamadeh, H.; Bushel, P.; Afshari, C.; Paules, R. S. *J. Comput. Biol.* **2001**, *8*, 625–637.
(26) Kerr, M. K.; Churchill, G. A. *Biostatistics* **2001**, *2*, 183–201.
(27) Kerr, M. K.; Afshari, C. A.; Bennett, L.; Bushel, P.; Martinez, J.; Walker, N. J.; Churchill, G. A. *Stat. Sin.* **2002**, *12*, 203–217.
(28) Dobbin, K.; Simon, R. *Bioinformatics* **2002**, *18*, 1438–45.
(29) Banks, R. E.; Stanley, A. J.; Cairns, D. A.; Barrett, J. H.; Clarke, P.; Thompson, D.; Selby, P. J. *Clin. Chem.* **2005**, *51*, 1637–1649.
(30) Hu, J.; Coombes, K. R.; Morris, J. S.; Baggerly, K. A. *Briefings Funct. Genomics Proteomics* **2005**, *3*, 322–331.
(31) Mann, C. J. *Emerg. Med. J.* **2003**, *20*, 54–60.
(32) Bailey, R. A. *Int. Stat. Rev.* **1985**, *53*, 171–182.
(33) Gan, C. S.; Chong, P. K.; Pham, T. K.; Wright, P. C. *J. Proteome Res.* **2006**, *6*, 821–827.
(34) Zhang, H.; Li, X.-J.; Martin, D. B.; Aebersold, R. *Nat. Biotechnol.* **2003**, *21*, 660–666.
(35) Li, X.-J.; Yi, E. C.; Kemp, C. J.; Zhang, H.; Aebersold, R. *Mol. Cell. Proteomics* **2005**, *4*, 1328−1340; http://sourceforge.net/project/showfiles.php?group_id=69281.
(36) Woo, Y.; Krueger, W.; Kaur, A.; Churchill, G. *Bioinformatics* **2005**, *21* (Suppl 1), i459−67.
(37) Kerr, K. F.; Serikawa, K. A.; Wei, C.; Peters, M. A.; Bumgarner, R. E. *OMICS* **2007**, *11*, 152–165.
(38) Dobbin, K. K.; Kawasaki, E. S.; Petersen, D. W.; Simon, R. M. *Bioinformatics* **2005**, *21*, 2430–2437.
(39) Smyth, G. *Stat. Appl. Genet. Mol. Biol.* **2004**, *3*, Article 3.
(40) Smyth, G.; Michaud, J.; Scott, H. S. *Bioinformatics* **2005**, *21*, 2067–2075.
(41) Brusniak, M.-Y.; Bodenmiller, B.; Campbell, D.; Cooke, K.; Eddes, J.; Garbutt, A.; Lau, H.; Letarte, S.; Mueller, L. N.; Sharma, V.; Vitek, O.; Zhang, N.; Aebersold, R.; Watts, J. D. *BMC Bioinf.* **2008**, *9*, 542.
(42) Kendziorski, C. M.; Zhang, Y.; Lan, H.; Attie, A. D. *Biostatistics* **2003**, *4*, 465–477.
(43) Peng, X.; Wood, C. L.; Blalock, E. M.; Chen, K. C.; Landfield, P. W.; Stromberg, A. J. *BMC Bioinf.* **2003**, *4*, 1–9.
(44) Zhang, S. D.; Gant, T. W. *Bioinformatics* **2005**, *21*, 4378–83.
(45) Shih, J.; Michalowska, A.; Dobbin, K.; Ye, Y.; Qiu, T.; Green, J. *Bioinformatics* **2004**, *20*, 3318–3325.
(46) Kendziorski, C.; Irizarry, R.; Chen, K.; Haag, J.; Gould, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 4252–4257.
(47) Stroup, W. W. Mixed model procedures to assess power, precision, and sample size in the design of experiments. *Proceedings of the Biopharmaceutical Section of the American Statistical Association*, Baltimore, MD, 1999.
(48) Paulovich, A. G.; Whiteaker, J. R.; Hoofnagle, A. N.; Wang, P. *Proteomics: Clin. Appl.* **2008**, *2*, 1386–1402.
(49) Dudoit, S.; Shaffer, J. P.; Boldrick, J. C. *Stat. Sci.* **2003**, *18*, 71–103.
(50) Benjamini, Y.; Hochberg, Y. *J. R. Stat. Soc., Ser. B* **1995**, *57*, 289–300.
(51) Karp, N. A.; McCormick, P. S.; Russell, M. R.; Lilley, K. S. *Mol. Cell. Proteomics* **2007**, *21*, 1354–1364.

PR8010099