Chapter 1

Important Issues in Planning a Proteomics Experiment: Statistical Considerations of Quantitative Proteomic Data

Karin Schork, Katharina Podwojski, Michael Turewicz, Christian Stephan, and Martin Eisenacher

Abstract

Mass spectrometry is frequently used in quantitative proteomics to detect differentially regulated proteins. A very important but unfortunately oftentimes neglected part in detecting differential proteins is the statistical analysis. Data from proteomics experiments are usually high-dimensional and hence require profound statistical methods. It is especially important to already correctly design a proteomic experiment before it is conducted in the laboratory. Only this can ensure that the statistical analysis is capable of detecting truly differential proteins afterward. This chapter thus covers aspects of both statistical planning as well as the actual analysis of quantitative proteomic experiments.

Key words Experimental design, Data preprocessing, Normalization, Sample size calculation, Statistical hypothesis test, Multiple testing, Fold change, Volcano plot

1 Introduction

The development and improvement of experimental procedures and technical equipment for protein detection and quantification have resulted in increasing numbers of conducted experiments each yielding vast amounts of data. Mass spectrometry-based quantitative proteomics is frequently used for the search for biomarkers for many different diseases, e.g., different cancer types or neurological diseases, and different types of tissue or body fluids, e.g., blood, urine or cerebrospinal fluid [1–3]. Often, the aim is to be able to understand biological processes behind the diseases and offer personalized therapies in the future [4]. However, the strength of most proteomic methods of measuring hundreds or thousands of protein abundances in complex biological samples within one experiment is also the soft spot of these methods. The derived data are high-dimensional and thus require sophisticated statistical analyses in order to draw the right conclusions. Unfortunately, in mass

spectrometry proteomic experiments, often only little attention is given to correctly planning and performing the statistical analysis.

Typically, differential proteomic experiments compare several sample groups and aim at finding differences in the protein abundances between the groups. Statistical methods to deal with this kind of scenario and the amount of hypothesis test conducted were often originally developed for genomics data and can in most cases also be applied to proteomics data. However, it is important to consider statistical issues already during the planning phase of a proteomic experiment. Only this can ensure an unobstructed statistical analysis that has the power to detect truly differential proteins.

Issues and pitfalls in experimental design are detailed in the next section. This is followed by a short section on preprocessing of the derived data. While preprocessing is a very important issue to derive reliable measures of protein quantification, it is far too broad to be covered in detail in this section. Instead, emphasis is put on the actual statistical analysis based on the derived quantitative measures. In Subheading 4, basic statistical principles are explained. These comprise statistical testing, adjusting for multiple testing, as well as sample size planning. Finally, the actual applications of statistical analyses to proteomic experiments are detailed in Subheading 5.

2 Planning a Proteomic Experiment

Typically, scientists take great care in planning all experimental procedures to be used in their experiments. Every step of the workflow is assessed, and the different steps combined in such a way that they are capable of answering the scientific questions. Sometimes even pilot experiments are performed to optimize individual steps of the workflow. All in all, the experimenters usually spend a lot of time in planning and of course also in conducting their experiments.

Once the experiment is done and the scientists have returned from the lab and are back at their computers, a “quick” statistical evaluation is sought. But all too often, the experimenters will find out that the questions they posed are not answerable with the data obtained from the experiment. And very often the cause for the failure of the experiment will be inadequate or even completely missing considerations of statistical issues during the planning phase of the experiment [5, 6].

2.1 Experimental Design for Proteomic Experiments

Quantitative proteomic experiments usually aim at detecting differentially regulated proteins between different sample groups. In the simplest differential experiment, biological samples from two different groups, e.g., tumor tissue samples vs. healthy control samples, are compared. In this case, there is one experimental factor

with two possible categories (i.e., tumor or healthy) that is studied. Of course, this example can be arbitrarily expanded both to a factor with more than two categories (i.e., samples from several tumor stages) and/or several experimental factors (i.e., gender, treatment with different substances, etc.) of interest. In the case of several experimental factors, there exist two basic types of designs. In a cross-classification design, each category of one factor is combined with each category of the other factor. In hierarchical designs, the possible categories of one factor depend on the category of the other factor. Thus not all possible combinations can be studied.

The first step in planning a proteomic experiment is thus the specification of all factors of interest and their possible categories that are to be studied within the experiment. It is of course also possible to incorporate continuous variables (i.e., age) into the design. Once all factors of interest are specified, several samples should be obtained per category of each factor. Using multiple samples ensures that detected differences are actually attributable to the studied factors and are not due to technical variation or intragroup biological variations. The necessary number of subjects per group can be assessed by sample size planning which is further detailed in Subheading 4.2.

It is however also possible that other factors apart from the ones of interest (e.g., factors coming from the experimental procedure) have an influence on the measured protein abundance. Problems arise especially, when a factor of interest is completely overlapping with such an uninteresting factor. Consider, for example, an experiment where the difference between two sample groups is measured. The samples have been processed in the laboratory in two batches on two different days. Now imagine the first batch only consisted of samples from the first experimental group, while the other batch only consisted of samples from the second group. If a differential protein is detected, one cannot tell whether the difference is due to the different group or because the samples have been processed on different days. In this case the time of processing is called a confounding factor. Typically all factors that contribute to sample handling may confound the results of an experiment. Also variables like age, gender, or additional underlying diseases may not be the factors of interest and have not been incorporated into the experimental design but still may confound the analysis.

A way to avoid confounding factors in an experiment is to already incorporate appropriate methods during experimental design. First of all, samples should be assigned randomly to the different categories of the factors to study. This should be done for all factors the researcher can influence (i.e., treatments or measurement order). This way, systematic errors from variables like age or gender can be diminished. Second, when different centers, instruments, batches, or the like are used within an experiment, samples from each category should be allocated ideally with equal sample

sizes to each of the different batches [7]. This experimental design is called block design [8]. Finally, an experimental procedure or protocol should never be changed halfway through an experiment as this is an almost certain source of error. The same is valid for the technical equipment as a whole as well as the individual parts. For example, the change of the LC column can already have an enormous effect that possibly obscures any true biological effect.

2.2 Design Considerations in Labeled Mass Spectrometry Experiments

In label-free mass spectrometry experiments [9, 10], the aforementioned considerations often are sufficient for planning the experiment. However, in mass spectrometry using isotopic labeling, additional aspects have to be considered. Through the use of isotopic labeling, it is possible to measure two or more samples within the same mass spectrometry run [11]. In general, these techniques reduce variances between the measurements of samples in the same run, as the samples are exposed to the same experimental conditions after the pooling. On the other hand, experiments with many samples still need to be distributed over multiple runs.

The natural question is which samples should be measured together in one run. Generally, two cases should be discerned. First there is the case of paired samples. Paired samples are present if each sample from one group is connectable to exactly one sample from the other group. This is, for example, the case, when two samples are drawn from the same subject, e.g., before and after a treatment. Also, pairings between more than two sample groups are possible, e.g., in time-course experiments. In these cases, each sample pair or group should naturally be measured together within the same mass spectrometry run.

In the second case when samples are independent or unpaired, there are generally two possibilities to advance. The first possibility is to randomly pair up one sample from each group. For this approach it is however necessary to have equal sample sizes in each group.

The other possibility to measure unpaired samples with isotopic labeling is to incorporate an internal standard (also called master mix), which can be established by pooling one aliquot of each sample. In each mass spectrometry run, samples are measured together with this internal standard, which can later be used to standardize the different MS runs. This reduces technical variations and thus makes all the different samples well comparable, no matter in which run they were measured. This procedure can especially also be used in the case of unequal sample sizes. On the other hand, this may require more MS runs and also more labeling reagents.

Another important aspect to consider is the possible effect of different isotopic labels on the measured protein abundances. By incorporating a label swapping strategy in the experimental design, measuring one experimental group always with the same label is avoided. This prevents bias that may be introduced by the different labels [12, 13].

3 Data Preprocessing

Suppose the proteomic experiment has been planned and finally conducted in the laboratory. The obtained raw results from mass spectrometry are available in the form of binary or text files representing the corresponding mass spectra. However, a direct measure for each peptide or protein is not readily available. To derive such measures, several steps of preprocessing have to be performed that connect the raw data to the different biological entities like peptides or proteins. This, for example, includes spectrum preprocessing, peptide identification, and protein inference and quantification. The preprocessing steps generally comprise methods from computer science or information technology but also from statistics. As there is a multitude of necessary steps and a great variety of suitable methods for each step, the reader is referred to corresponding reviews on this matter for more information [14–17]. Both commercial and free software solutions are available for each step or as complete workflows [18–21]. At this point, it is assumed that a suitable software solution is used for the derivation of quantitative values for peptides or proteins. In the following paragraphs, preprocessing steps for a quantitative proteomics data set are discussed, which are an essential preparation for the following statistical analysis.

3.1 Missing Values

In quantitative proteomics datasets, missing values (*see* Chap. 27) are very common and can make up a large fraction of the whole dataset. The way missing values are handled can have a huge impact on the results of the following analysis.

A data point can be missing for various reasons. For example, a missing value will occur if the corresponding protein is not present in the sample or was present, but its abundance was below the detection limit. Other reasons are based on the nature of mass spectrometry measurement, e.g., the fact that the selection of precursor ions for fragmentation is stochastic or that noisy MS/MS spectra may not be identified with high enough confidence. But also additional filters of the quantitative data, as requiring at least two unique peptides, or normalization steps may lead to missing values. These different missing value types may require different treatment, but unfortunately, often for a certain missing value in the data, it is not clear why exactly it is missing.

There are three different methods to handle missing values: (1) remove proteins with missing values, (2) perform analysis only on valid values, and (3) impute missing values. Removing all proteins that contain missing values can lead to a substantial information loss, while replacing missing values with a valid value (missing value imputation) has also a lot of drawbacks, especially in the case of many missing values or unknown origin of them [22]. For

instance, imputing a constant value (e.g., the mean or median of the valid values) for a protein will lead to an underestimated variance that can easily lead to false-positive findings.

Often the strategy is to first remove proteins with too few valid values (e.g., <50%) and then continue either with the remaining valid values or choose an appropriate imputation strategy.

3.2 Normalization

Quantitative proteomics data contains biological variation that is interesting to investigate but also unwanted technical variation. Even if possible confounding factors were considered during experiment planning (*see* Subheading 2.1), there will still be small variations in experimental conditions and sample handling (e.g., temperature, pipetting), which lead to technical bias. Often, the exact reasons for this bias are unknown.

Normalization strategies reduce the technical bias while keeping the interesting biological differences. After normalization, the samples are in general more comparable which makes the following statistical analysis more reliable. In general, high-throughput data and the assumption that the majority of proteins do not change between the experimental groups are required (*see* Subheading 6.1).

There are many different normalization methods that were originally developed for genomics data or especially tailored at proteomics. Although there are several comparisons of these methods on proteomics data [23–25], in most cases it is not clear beforehand which method will work best for a certain dataset. Because of this, it is advisable to try different normalization methods and evaluate those using different types of plots [26].

Boxplots can give an overview over the whole dataset and are suitable to compare non-normalized with normalized data. The length of the boxes can show differences in variance between samples and may indicate outliers that stick out even after normalization.

An MA plot compares two samples with each other and helps to assess whether these samples have been normalized appropriately. On the x -axis the average of the $\log_2$ -transformed intensities (A -value) and on the y -axis the difference (M -value) is shown. Low-abundant proteins will show up on the left side and high-abundant proteins on the right side. Proteins with a high difference between the two samples are shown on the top and the bottom. Often, a local linear regression (loess) curve is drawn into the MA plot. For well-normalized data, this line will be very close to the horizontal line at $M = 0$ (*see* Fig. 1). Deviations from this line might indicate unnormalized data or a normalization method that does not fit to the data (e.g., when median normalization is not able to adjust differences in variance).

Even if the experiment is carefully planned, batch effect sometimes cannot be completely avoided. A principal component analysis plot (PCA plot) can uncover possible batch effects by

Fig. 1 Examples for MA plots for the comparison of two samples. Panel (a) shows unnormalized data with a high bias between the two samples. The normalization used for (b) is not ideal as the MA plot still shows a deviation of the local regression line from $M = 0$. Panel (c) shows a suitable normalization method where the regression line is almost equal to $M = 0$

representing the whole dataset in a two-dimensional plot. As batch effects are sometimes not completely removed by standard normalization methods, special batch normalization methods have been proposed [27].

4 Basic Statistical Concepts for Difference Detection

When performing quantitative proteomics experiments, one naturally wants to measure the abundances of the individual proteins within the samples. However, the calculated peptide or protein intensities are not absolute abundance measures in a standard

proteomics experiment [28]. This means that one cannot directly confer the absolute amount of each protein (either molecule number or volume). This is because the different peptides do not ionize equally well (i.e., they have different ionization efficiencies) and the factor by which their intensities are reduced differs from peptide to peptide.

What is however possible is to compare the abundances in one sample to those of another sample. This is also called *relative quantification* (see Chaps. 9, 10, 13–21, 23, 24, 26).

Through this procedure it is especially possible to quantify differences in protein expression between different sample groups. Typically, the abundance ratio or the fold change is used to quantify the relative expression change between two samples or sample groups. Basically these measures compare the mean expression in one sample group to that of the other sample group. The exact formulae are given in Subheading 6.2.

4.1 Statistical Hypothesis Tests

The mere quantification of the magnitude in abundance change between different groups by abundance ratio or fold change is only the first step toward identifying differentially regulated proteins. Declaring a protein differentially expressed if the corresponding ratio is different from one is the solution that first comes to mind. However, this strategy neglects some of the typical characteristics of experiments in general.

First, every experimental procedure has its typical variability and imprecision resulting in measurements that slightly deviate from the true values. Second, an experiment normally only uses a small subset of subjects from the whole population of interest.

There are also always variations of the measured values (e.g., protein abundances) between different subjects of a population. In summary, there will always be deviations between the observed values of the sample subset and the true mean value of the complete population of interest. All these variations result almost certainly in observed abundance ratios differing from one for every single measured protein even though most of the proteins in truth are not differentially expressed between the different groups. To detect truly differential proteins, one will thus have to distinguish proteins with ratios truly different from one from those proteins where the measured difference is only due to experimental imprecision.

To solve this problem, a statistical test can be performed. A statistical test is based on hypotheses about the characteristics of both the populations. The *null hypothesis* usually describes the state one would like to rebut. In differential proteomics this would be “the protein *is not* differentially expressed.” The *alternative hypothesis* states the opposing characteristic, i.e., “the protein *is* differentially expressed.” To assess which of the two hypotheses holds true, a test statistic based upon the observed measurements is calculated. The derived score of this test statistic is then used to decide which hypothesis is to be chosen.

Table 1**Possible results from statistical tests and measures for assessing a statistical testing procedure**

		Test result		Measures
		Protein differential	Protein <i>not</i> differential	
Unknown reality	Protein differential	True Positive	False Negative (type II error (β))	Sensitivity = TP/ (TP + FN)
	Protein <i>not</i> differential	False Positive (type I error (α))	True Negative	Specificity = TN/ (TN + FP)

The procedure for deciding which of the hypotheses is true is comparable to a criminal proceeding where the accused is presumed to be innocent until the contrary can be proven. In the context of statistical testing, the null hypothesis is assumed to be true unless it can be proven to be false. To this end, assumptions are made about the distribution of the test statistic under the null hypothesis. The distribution particularly takes the variability of the data into account. If the obtained value of the test statistic is too extreme in terms of the distribution under the null hypothesis, then the null hypothesis is rejected in favor of the alternative. The test result is then called *significant*. Otherwise there is not enough evidence, and the null hypothesis has to be retained.

Oftentimes the so called *p-value* is calculated to assess whether a test is significant. The *p-value* is the probability of obtaining a test statistic at least as extreme as the calculated one under the assumption that the null hypothesis is true. If the *p-value* is below or equal to the pre-specified *α -level*, the null hypothesis can be rejected.

As the test decision is based on probabilities, it is possible to decide erroneously. All possible test decisions are outlined in Table 1 together with the two types of errors that may occur depending on the decision made. The *α -error* occurs if the null hypothesis is rejected even though it is true. The *β -error* is present if the null hypothesis is retained even though it is false. Unfortunately, the two possible errors are conflictive. If one of the error types is decreased and all other experimental characteristics are kept unchanged, the other error type will automatically increase. Returning to the court example, the presumption of innocence is based on the concept that it is generally deemed worse to convict someone innocent than failing to convict someone actually guilty. In the context of statistical testing, it is worse to declare a difference where in fact there is none than to miss one true difference. In accordance with this concept, the probability for the type I error is controlled through specifying the *α -level* of the test to ensure that the probability for wrongly rejecting the null hypothesis is small. A typically chosen *α -level* in statistical testing is 5%. If one wants to be

very restrictive, this can be decreased further, e.g., to 1%. In any case, the α -level has to be specified before the performance of the experiment and its statistical evaluation.

4.2 Power and Sample Size

So far, the importance of keeping the type I error small has been shown. However, a researcher will generally also want to make sure to be able to detect a true difference.

The probability of rejecting the null hypothesis in favor of the alternative hypothesis when the alternative actually is true is called the *power* of the test. It is the opposite of the probability of a type II error (power = $1 - \beta$). The power depends on the true effect size, the variance, and the sample size. A large effect size, a small variance, and a high sample size will increase the power. Additionally, the power also depends on the significance niveau (α -level). Once the α -level has been fixed, the only remaining aspect a researcher can usually influence is the sample size of the experiment.

By increasing the sample size, it is possible to increase the power of a test. The relationship between power and sample size under several scenarios for effect size and variance is depicted in Fig. 2 for the special case of the two-sample t -test. It shows that the sample size will have to be increased with decreasing true effect size to obtain a certain power of the test. Also, for increasing variance the sample size will have to be increased to keep a certain power of the test.

As all the described characteristics are interdependent, it is necessary to specify all except one beforehand to be able to calculate the value of the remaining one. It is hence necessary to know and

Fig. 2 Relationship of power and sample size for the two-sample t -test. *Left:* Fixed α -level and standard deviation σ with three different choices for the true effect size. *Right:* Fixed α -level and effect size Δ with three different choices for the standard deviation

specify the expected variance and the true effect size one desires to be able to detect before the experiment (*see* Subheading 6.3). The expectable true effect size depends on the samples used in the experiment and especially on how different the studied groups are. It is generally necessary to make a compromise between the possibility to detect very small expression changes and keeping the sample size reasonable.

However, one should try to specify reasonable expectable effect sizes one would like to detect, because an overestimation of expectable effect sizes can easily impede a complete experiment. Then the detectable effect size will be too large and reasonable, but smaller differences are detectable any more.

Specifying the expectable variance is even more difficult than specifying the effect size. Information about the variance are generally only be accessible through earlier experiments or from the literature. In this case it is necessary to make sure the experimental procedure to be used will be (almost) identical to the one described in the literature. Only this ensures that the variances are transferable. If no information is found in the literature, one could perform a small pilot experiment beforehand and estimate the variance from this experiment [29].

Finally, the anticipated power to detect a differential protein needs to be specified. The power is generally not set as strict as the α -level of a test. Typically, 80% are used in many statistical applications. Dropping the power much further would decrease the probability of detecting a true expression change accordingly.

4.3 Multiple Testing

Especially when measuring complex biological samples like blood or tissue, up to several hundreds or even thousands of peptides or proteins are measured simultaneously. Typically, a separate statistical test is performed for each protein to evaluate, if it is differentially expressed. For each of these tests, there is the possibility of a false test decision. The expected number of false-positive test decisions increases linearly with the number of tests performed (Table 2). And the probability for at least one false-positive test decision increases even more dramatically and converges to a 100% very fast.

Especially when considering the time and effort generally put into downstream validation of each single detected protein, one would like to keep the number of false positives to a minimum and avoid unnecessary work spent on truly unregulated proteins. But also when no further work is performed in the lab, a researcher will make sure that the published results do indeed reflect truly regulated proteins. These circumstances make it obvious that the number of false positives has to be controlled.

The typical way to control the number of false-positive test decisions in a multiple testing setting is to apply multiple testing correction methods that control certain error rates [30], the family-wise error rate (FWER) or the false discovery rate (FDR).

Table 2
Relation between false-positive test decisions and number of tests performed

Number of tests performed	Probability for at least one false-positive test decision ($\alpha = 0.05$)	Expected number of false-positive test decisions ($\alpha = 0.05$)
1	5.00%	≈ 0
2	9.75%	≈ 0
5	22.62%	≈ 0
10	40.13%	0–1
20	64.15%	1
100	99.41%	5
1000	100.00%	50

The FWER is the probability of having at least one false-positive test decision among all test decisions. The so-called Bonferroni correction adjusts the p -values so that the FWER is controlled by multiplying the original p -values with the number of performed tests. An adjusted p -value thus reflects the probability of having at least one false-positive test result. The formula for the adjusted p -values is given in Subheading 6.4. However, this method is very conservative as it basically allows *no* false-positive test decision among all tests.

The FDR is defined as the fraction of false-positive test decisions among all positive test decisions. Thus, one can allow several false-positive tests, given that there are enough true positive results. Because false positives are allowed to a certain degree, controlling the FDR generally is less strict. Several algorithms have been developed that control the FDR under different assumptions. Especially, methods for deriving adjusted p -values (called q -values) have been introduced. A q -value directly reflects the FDR, as it gives the fraction of false positives among all positives under the condition that the corresponding protein is still considered significantly regulated. A formula for deriving q -values is given in Subheading 6.4.

4.4 Statistical Significance and Biological Relevance

In proteomics literature often a *fold change* (FC, ratio of means of the two compared groups) cutoff is used in addition to the p -value. From a statistical point of view, this is not necessary. The p -value already contains the complete information if a protein is significantly regulated. This means, if the p -value is small enough (i.e., less than 0.05), then there is statistical evidence that the mean abundance of a protein between the experimental groups is different. However, from a biological point of view, even though a difference is significant, it might not be relevant, because it is too small.

Classically the terms *statistical significance* and *biological relevance* are used to characterize these different concepts. The fold change cutoff is a means of introducing this latter concept of biological relevance.

Now the question remains of how to choose such a fold change cutoff. Even though a fold change cutoff is regularly used, its derivation is rarely described in the proteomics literature. To deduce the cutoff, one might try to consider what is a biologically relevant difference in mean abundance. Opposing to the connotation of “biological,” the relevance of a difference in the context of proteomic experiments is not connected to biological processes that might be controlled by it. Rather, this concept takes into consideration the technical circumstances of the experiment. Each experimental procedure has an intrinsic variability or inaccuracy in measuring outcomes. This is the so-called *technical variation* of the experiment. In proteomics experiments such variation is introduced, for example, through sample handling and through the technical equipment like LC systems or mass spectrometer. Due to this technical variation, different outcomes are observed when measuring the same sample several times. When considering this variability, it might be possible that, even though significant, a difference in mean abundance is below the typical technical variation of the experiment. Then it cannot be assessed, whether the difference found is in fact due to biological difference or just due to the experimental procedure. Biologically relevant is thus a result which lies above the technical variation of the experimental procedure, and a fold change cutoff should be chosen accordingly.

The technical variation of one’s experimental procedure is dependent on the sample type, handling, and equipment used. It can, for example, be estimated through technical replicates. Details on this estimation are given in Subheading 6.5. Additionally, the fold change may be used as a second criterion to reduce the list of significant proteins for further validation. By choosing a high enough fold change cutoff, it is also ensured that abundance differences are visible, e.g., in Western blots, and that are often used for validating biomarker candidates with another experimental method.

P -values and fold changes can be simultaneously depicted in a so-called *volcano plot* (Fig. 3). On the x -axis the $\log_2(\text{FC})$ and on the y -axis the $-\log_{10}(p\text{-value})$ are shown. Cutoffs for fold changes and p -values can also be added. The most interesting biomarker candidates (those with a low p -value and a fold change away from one) will show up in the left and right upper corner of the volcano plot. It is also possible to combine fold changes and p -values into a single measure (see Subheading 6.6).

Fig. 3 Example of a volcano plot combining information of biological and statistical relevance. Cutoffs of 0.05 for the p -value and 0.5 resp. 2 for the fold change are applied. Proteins that do not reach these cutoffs are depicted in gray. Significantly regulated proteins after FDR correction using the Benjamini-Hochberg procedure are highlighted in orange

5 Statistical Tests in Proteomic Experiments

As described above, a statistical test suitable to detect the differences between different sample groups has to be applied once the proteomic experiment has been performed, data are preprocessed, and abundance measures are obtained for each peptide or protein. Standard statistical tests generally can be used for difference detection in proteomics. Which test is suitable is directly dependent on the design of the conducted experiment.

5.1 Comparing Two Sample Groups

Commonly, proteomic experiments are performed to find differences in the proteome between two different sample groups. The t -test is used most often in the literature to detect differential proteins. The t -test is the traditional parametric test for difference between two sample means. However, the t -test requires normally distributed measurements, which may not be the case with proteomic measurements. Abundance measurements based on either peak areas or heights can only take positive values, and their distribution is often skewed, i.e., they do not follow a normal distribution.

However, log-transformed abundance values will approximately be normally distributed. In any case, the normality assumption should be checked before the t -test is applied. This can be done visually through QQ plots, for example. For these plots, the theoretical quantiles of the standard normal distribution are plotted against the measured quantiles. If the points lie approximately on a straight line, then normal distribution can be assumed.

Once the conditions for the application of the t -test have been verified, there are further things to consider. Data analysis software usually offers several options for the t -test. The first option is to either use a *paired* or an *unpaired* t -test. This decision depends on the experimental design. In the case of unpaired samples, mean values are calculated separately for each group. The means are afterward compared to detect differential proteins. In the paired samples setting, the difference for each pair of samples is calculated first, and afterward the mean of these differences is calculated and compared to zero (i.e., no difference between groups).

The second selectable option is to use a *one-sided* vs. a *two-sided* alternative hypothesis. A two-sided alternative is equivalent to testing if there is any change in abundance between the groups regardless of the direction. This test is thus able to detect both up- and downregulation. If it can be assumed that the experiment will result in *only* down- or *only* upregulated proteins, the one-sided test may be chosen as it has a higher power in the specified direction. However, it is never able to find any changes in the opposite direction (that may still occur unexpectedly).

If the data is not normally distributed (even after log transformation), there are a number of nonparametric alternatives to the t -test that do not rely on normally distributed data. Especially the two-sample Kolmogorov-Smirnov test (K-S-test) or the Mann-Whitney- U test can be used for difference detection when the data is not normal.

5.2 Analysis of Multiple Sample Groups and Additional Factors

Of course, it is also possible that more than two sample groups are to be compared in a proteomic experiment. A typical example is the comparison of samples from different tumor stages. But also more advanced experimental designs where additional factors are incorporated into the analysis are possible and have been introduced in Sect. 2. In these cases analysis of variance (ANOVA) and related methods can be used. The basic idea behind ANOVA is to study the effect of each incorporated factor on the peptide or protein abundance. This is done by allocating portions of the overall variance to the different factors and performing a statistical test for each factor based on the apportioned variances. Each test evaluates if the corresponding factor has a significant influence on the abundance level. If there is a significant influence, the abundance will differ for different levels the factor takes on. In the case of comparing different tumor stages, for example, a significant

influence of the factor “tumor stage” means that the abundance is different between at least one pair of tumor stages. It is also possible to assess interactions between different factors with ANOVA methods. When an interaction is present, the influence of one factor is different for the different levels of the interacting factor.

Similarly to the t -test, classical linear models and ANOVA methods generally assume normally distributed values. Thus the log transformation is again advisable. Aside from that, ANOVA methods are very flexible. Hence they are applicable to a multitude of different experimental designs with complicated factor combinations. However, the reader is strongly recommended to seek advice from a statistician *before* performing such complicated experimental designs. Good knowledge of the analysis method is needed to ensure the factors of interest are really assessable by ANOVA. This can only be guaranteed by careful planning of such sophisticated experiments with several factors and possibly even interactions.

Once a factor with more than two levels is found to have a significant influence, the natural question is which pairs of levels actually are differential. This can be assessed by additionally performing a test for each possible pair of factor levels. There exists special post hoc tests like Tukey’s honestly significant difference (HSD) test, which is based on the standard t -test. Additionally, a multiple testing correction is incorporated into the post hoc test by taking into account the total number of pairwise tests performed. This way the α -level is conserved across all tests.

6 Notes

6.1 Normalization in Case of Many Changing Proteins

For some datasets the assumption for normalization that the majority of proteins do not change between experimental groups may not hold. It may come to mind to do a groupwise normalization, i.e., normalizing every group separately and combining the data for the analysis. This type of normalization is dangerous and not recommended, as it can introduce an artificial bias that can make proteins look significantly differentially expressed if there are not and increase the number of false positives. Alternatively, there exists normalization methods that can cope with a larger portion of changing proteins between groups than standard normalization methods. An example is the least trimmed squares (LTS) normalization [31].

6.2 Fold Change

Consider the expression of a protein is to be compared between two sample groups. The first group consists of n samples and the second group of m samples. Let x_i be the protein abundance measured in the i th sample in the first group and let y_j be the protein abundance in the j th sample in the second group. If the

samples in one group are independent from the samples in the other group (unpaired samples), the fold change is calculated as

$$FC_{\text{unpaired}} = \frac{1}{n} \sum_{i=1}^n x_i / \frac{1}{m} \sum_{j=1}^m y_j = \frac{\bar{x}}{\bar{y}}$$

In the case of paired samples, e.g., x_i and y_i originate from the same sample, or they have been measured in the same LC-MS run in a labeled experiment, the fold change is defined as

$$FC_{\text{paired}} = \frac{1}{n} \sum_{i=1}^n \frac{x_i}{y_i} = \overline{(x/y)}$$

Note that with paired samples, the sample size has to be the same in both groups, i.e., $n = m$.

6.3 Expected Effect Size and Variance for Choosing the Sample Size

To specify a correct effect size and variance for the determination of the optimal sample size, several things have to be considered. In the classical scenario where protein abundances from two different sample groups are compared, usually the two-sample t -test is used. The true effect size in this case is a difference based on (the logarithm of) the true abundances of the two groups (*see* also Subheading 5.1):

$$\Delta = \mu_{\log(x)} - \mu_{\log(y)}$$

This true effect size Δ can be estimated through the arithmetic means of the measured log-transformed intensities:

$$d = \overline{\log(x)} - \overline{\log(y)}$$

This difference is *not* identical with a fold change. Thus, while using the fold change as a measure for the size of abundance change between the two groups, the specifiable effect size for the determination of the correct sample size needs to be the above difference. Keep in mind that the variance also has to be specified based on the logarithmized scale.

Another problem to consider is that there are usually hundreds of proteins to be tested within one proteomic experiment. Each protein will have its own effect and variance. However, only one value can be used in the power analysis to derive the correct sample size. When a pilot experiment is used for the derivation, one can, for example, take the mean, median, or a specific quantile of all the observed variances. Generally one should keep in mind that it is reasonable to be conservative (i.e., overestimate the true variance). This will result in slightly higher sample sizes but at the same time ensure a true effect of the specified size is actually detectable with the conducted experiment.

6.4 Adjusted p -Values

Assume a differential proteomic experiment has been performed where n proteins have been quantified and tested for differential expression. For each test i , a p -value p_i has been derived. Adjusted p -values that control the FWER are derived by applying the following procedure of Bonferroni to each p -value:

$$p_i^* = \min(n \cdot p_i, 1)$$

A protein is then assumed to be differentially expressed, if the corresponding adjusted p -value p_i^* is below a pre-specified α . This procedure ensures that the probability for at least one false-positive test decision among all n test decisions is less or equal to α .

An alternative and less strict multiple testing correction is the control of the FDR. In the case of independent test statistics, the procedure by Benjamini and Hochberg [32] uses the n ordered p -values, $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(n)}$ to derive adjusted q -values through

$$q_{(i)} = \min_{k=i, \dots, n} \left(\min \left(\frac{n}{k} p_{(k)}, 1 \right) \right)$$

Again, a protein is assumed to be differentially expressed, if the corresponding q -value is below a pre-specified α . The expected fraction of false-positive test decisions among all positive test decisions then is less or equal to α .

6.5 Fold Change Cutoff

To estimate the technical variation of a procedure, several technical replicates should be measured, for example, within a pilot experiment. The easiest way to assess the variation is to use the coefficient of variation. The coefficient of variation is a measure of the percentage variation and is defined as

$$CV = \frac{\text{standard deviation}}{\text{mean}}$$

To derive a fold change cutoff for a proteomic experiment, compute the coefficient of variation for each protein. Then take the maximum or, for example, the 95% quantile of all coefficients. Finally set the fold change cutoff for positive fold changes to

$$cut = \frac{1 + \widetilde{CV}}{1}$$

and to $-cut$ for negative fold changes, respectively. $\widetilde{CV}$ is the chosen maximum or quantile of the measured coefficients from all proteins.

6.6 Euclidean Distance Measure in Volcano Plots

Fold changes and p -values can be combined into a single measure using the volcano plot. The Euclidean distance measures the distance of a point in the volcano plot to the origin of the coordinate system [33]. The higher the Euclidean distance, the more interesting is the corresponding protein. However, it has to be taken care that the x - and y -axis of the volcano plot are comparable and the

plot is close to quadratic. If not, either fold changes or p -values will gain a higher weight. Adjustment of the volcano plot can be achieved by changing the base for the logarithm of either p -values or fold changes.

Acknowledgments

This work was supported by de.NBI, a project of the German Federal Ministry of Education and Research (BMBF) [grant number FKZ 031 A534A], as well as P.U.R.E. (Protein Research Unit Ruhr within Europe) and ProDi (Center for protein diagnostics), Ministry of Innovation, Science and Research of North-Rhine Westphalia, Germany.

References

1. Crutchfield CA, Thomas SN, Sokoll LJ et al (2016) Advances in mass spectrometry-based clinical biomarker discovery. *Clin Proteomics* 13:1
2. Thomas S, Hao L, Ricke WA et al (2016) Biomarker discovery in mass spectrometry-based urinary proteomics. *Proteomics Clin Appl* 10:358–370
3. Bharucha T, Gangadharan B, Kumar A et al (2019) Mass spectrometry-based proteomic techniques to identify cerebrospinal fluid biomarkers for diagnosing suspected central nervous system infections. A systematic review. *J Infect* 79:407–418
4. Kowalczyk T, Ciborowski M, Kislik J et al (2020) Mass spectrometry based proteomics and metabolomics in personalized oncology. *Biochim Biophys Acta Mol Basis Dis* 1866:165,690
5. Hu J, Coombes KR, Morris JS et al (2005) The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Br Funct Genomic Proteomic* 3:322–331
6. Karp NA, Lilley KS (2007) Design and analysis issues in quantitative proteomics studies. *Proteomics* 7(Suppl 1):42–50
7. Cairns DA (2011) Statistical issues in quality control of proteomic analyses: good experimental design and planning. *Proteomics* 11:1037–1048
8. Tocher KD (1952) The design and analysis of block experiments. *J R Stat Soc Ser B* 14:45–91
9. Neilson KA, Ali NA, Muralidharan S et al (2011) Less label, more free: approaches in label-free quantitative mass spectrometry. *Proteomics* 11:535–553
10. Megger DA, Bracht T, Meyer HE et al (2013) Label-free quantification in clinical proteomics. *Biochim Biophys Acta* 1834:1581–1590
11. Chahrour O, Cobice D, Malone J (2015) Stable isotope labelling methods in mass spectrometry-based quantitative proteomics. *J Pharm Biomed Anal* 113:2–20
12. Park S-S, Wu WW, Zhou Y et al (2012) Effective correction of experimental errors in quantitative proteomics using stable isotope labeling by amino acids in cell culture (SILAC). *J Proteome* 75:3720–3732
13. Maes E, Valkenburg D, Baggerman G et al (2015) Determination of variation parameters as a crucial step in designing TMT-based clinical proteomics experiments. *PLoS One* 10: e0120115
14. Shteynberg D, Nesvizhskii AI, Moritz RL et al (2013) Combining results of multiple search engines in proteomics. *Mol Cell Proteomics* 12:2383–2393
15. Griss J (2016) Spectral library searching in proteomics. *Proteomics* 16:729–740
16. Audain E, Uszkoreit J, Sachsenberg T et al (2017) In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. *J Proteome* 150:170–182
17. Blein-Nicolas M, Zivy M (2016) Thousand and one ways to quantify and compare protein abundances in label-free bottom-up proteomics. *Biochim Biophys Acta* 1864:883–895
18. Sinitcyn P, Rudolph JD, Cox J (2018) Computational methods for understanding mass spectrometry-based shotgun proteomics data. *Annu Rev Biomed Data Sci* 1:207–234
19. Navarro P, Kuharev J, Gillet LC et al (2016) A multicenter study benchmarks software tools

- for label-free proteome quantification. *Nat Biotechnol* 34:1130–1136
20. Misra BB (2018) Updates on resources, software tools, and databases for plant proteomics in 2016–2017. *Electrophoresis* 39:1543–1557
 21. Turewicz M, Kohl M, Ahrens M et al (2017) BioInfra.Prot: a comprehensive proteomics workflow including data standardization, protein inference, expression analysis and data publication. *J Biotechnol* 261:116–125
 22. Webb-Robertson B-JM, Wiberg HK, Matzke MM et al (2015) Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J Proteome Res* 14:1993–2001
 23. Callister SJ, Barry RC, Adkins JN et al (2006) Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J Proteome Res* 5:277–286
 24. Karpievitch YV, Dabney AR, Smith RD (2012) Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics* 13(Suppl 16):S5
 25. Välikangas T, Suomi T, Elo LL (2018) A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief Bioinform* 19:1–11
 26. Willforss J, Chawade A, Levander F (2019) NormalyzerDE: online tool for improved normalization of omics expression data and high-sensitivity differential expression analysis. *J Proteome Res* 18:732–740
 27. Lazar C, Meganck S, Taminiau J et al (2013) Batch effect removal methods for microarray gene expression data integration: a survey. *Brief Bioinform* 14:469–490
 28. Unwin RD, Evans CA, Whetton AD (2006) Relative quantification in proteomics: new approaches for biochemistry. *Trends Biochem Sci* 31:473–484
 29. Cairns DA, Barrett JH, Billingham LJ et al (2009) Sample size determination in clinical proteomic profiling experiments using mass spectrometry for class comparison. *Proteomics* 9:74–86
 30. Ge Y, Sealfon SC, Speed TP (2009) Multiple testing and its applications to microarrays. *Stat Methods Med Res* 18:543–563
 31. Huber W, Von Heydebreck A, Sültmann H et al (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18:S96–S104
 32. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 57:289–300
 33. Großerueschkamp F, Bracht T, Diehl HC et al (2017) Spatial and molecular resolution of diffuse malignant mesothelioma heterogeneity by integrating label-free FTIR imaging, laser capture microdissection and proteomics. *Sci Rep* 7:44,829

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.