Five simple yet essential steps to correctly estimate the rate of false differentially abundant proteins in mass spectrometry analyses

Samuel Wieczorek, Quentin Giai Gianetto, Thomas Burger

Five simple yet essential steps to correctly estimate the rate of false differentially abundant proteins in mass spectrometry analyses

Samuel Wieczorek[1], Quentin Giai Gianetto[2,3], Thomas Burger[1,4] *

[1]Univ. Grenoble Alpes, CEA, INSERM, BIG-BGE, 38000 Grenoble, France
[2]Bioinformatics and Biostatistics Hub, C3BI, Institut Pasteur, USR 3756 IP CNRS, 75015 Paris, France
[3]Proteomics platform, Mass Spectrometry for Biology Unit, Institut Pasteur, USR 2000 IP CNRS, 75015 Paris, France
[4]CNRS, BIG-BGE, F-38000 Grenoble, France
*thomas.burger@cea.fr

Abstract

Results from mass spectrometry based quantitative proteomics analysis correspond to a subset of proteins which are considered differentially abundant relative to a control. Their selection is delicate and often requires some statistical expertise in addition to a refined knowledge of the experimental data. To facilitate the selection process, we have considered differential analysis as a five-step process, and here we present the practical aspects of the different steps. Prostar software is used throughout this article for illustration, but the general methodology is applicable with many other tools. By applying the approach detailed here, researchers who may be less familiar with statistical considerations can be more confident in the results they present.

---

1. Introduction

The combination of liquid chromatography with mass spectrometry has made it possible to identify and quantify large numbers of biomolecules at high-throughput and without requiring labeling. However, despite continuous improvements, this pipeline remains imperfect. Notably, the speed of analysis, which allows for more extensive coverage, comes at the price of lower-quality data, affected by both batch effects [1] and missing values [2], [3], [4]. In this context, determining with a given statistical significance level which biomolecules are differentially abundant (DA) between several biological conditions remains challenging. To help omics researchers with this task, numerous software tools have been developed, including our own contribution, Prostar [5]. Prostar was originally developed to deal with peptide and protein data, but the methodology it relies on is essentially the same as that used for various classes of metabolites [6], [7].

Whatever the software tool, we have noticed that, the preprocessing steps are generally quite straightforward. Most tools provide intuitive interfaces for any user to successfully (1) filter out irrelevant or low-quality analytes; (2) normalize the ion intensities to account for batch effects; and (3) impute missing values thanks to off-the-shelf algorithms. Although a deeper understanding of the underlying mathematics is always helpful when tuning the algorithms, an analyst with a refined knowledge of their data can generally preprocess them in a nearly optimal way. In contrast, the final differential analysis requires some statistical expertise when seeking to identify subsets of proteins that can be confidently deemed DA in at least one biological condition relative to the others.

With this in mind, this article details five steps which are essential to the differential analysis. We identified these steps some time ago, and therefore included the appropriate visualization interfaces and routines in Prostar to make them as easy and intuitive to apply as possible. For this reason, Prostar will be used throughout this article for illustration purposes, but these steps have a methodological backbone that should apply for any differential analysis, whatever the software tool -- and to some extent, whatever the type of biomolecule studied, be it proteomics, peptidomics, metabolomics or lipidomics; even thought this article is focused on proteomics. The five steps (and their associated questions) are the following:

1. At early stages of the differential analysis, it is customary to discard proteins presenting an excessively low fold-change. How should the corresponding cutoff value be tuned?
2. The data are classically a mixture of observed values (direct mass spectrometry measurements) and recovered values (run alignments, imputations, etc.), which should not be trusted equally. However, discarding recovered values is not always possible when making multiple simultaneous comparisons. How can the impact of condition-related differences in recovered intensity values be controlled on the final differential analysis?
3. To calculate a reliable False Discovery Rate (FDR) requires well-calibrated p-values. How can the calibration correctness be easily assessed?
4. The FDR computation is highly sensitive to the overall proportion of non-DA proteins. How can this sensitivity be exploited to improve the calibration?
5. Similarly, the quality of the FDR computation depends on the total number of proteins measured (whether DA or not). How can this dependence be exploited to improve the FDR quality control?

2. Tuning the log-fold-change

The logarithmized fold-change of protein x from condition A to condition B, formally defined as

$$\log_2\left(\frac{mean_{i\in A}(x_i)}{mean_{i\in B}(x_i)}\right) = \log_2(mean_{i\in A}\, x_i) - \log_2(mean_{i\in B}\, x_i)$$

(where x_i denotes the intensity of protein x in sample i), is classically approximated (See Supplemental Material, Sec. 5, where this approximation is discussed) by the following quantity:

$$\mathrm{logFC}_{A/B}(x) = mean_{i\in A}(\log_2 x_i) - mean_{i\in B}(\log_2 x_i).$$

At the beginning of any differential analysis, proteins for which the logFC is below a certain threshold are generally filtered out. There are two motivations for this filtering: First, such proteins are often difficult to exploit in post-proteomics biological investigations, so that researchers prefer to focus on proteins for which intensities change to a greater extent. Second, knowing the cumulated sources of variations (biological, technical, analytical) and the relatively low number of observed intensity values (either because of a small number of samples, or because of a high proportion of missing values), one fears inaccurate estimated fold-changes. Notably, proteins with similar concentrations between the compared conditions can lead to non-zero logFC (see Supplemental Material, Sec. 1). In this context, applying a more stringent logFC threshold appears to be a good way to reduce false discoveries (i.e., non-DA proteins that are falsely selected as DA).

Thus, there are both practical and theoretical motivations for logFC filtering. As statistical software developers, we are not qualified to discuss the rationale behind any biological or analytical reasoning and we assume researchers can use their prior knowledge on the experiment to adequately tune the logFC cutoff. However, we can present our case against the statistical one. First, in absence of any rule to optimize the logFC cutoff, one may face uncontrolled boundary effects (i.e. filtering proteins which

are significantly DA, while keeping proteins with slightly higher logFC, yet endowed with a variance which makes them non-significant), not to speak about the approximation underlying the logFC (see above). Second, several statistically-relevant solutions exist to cope with the aforementioned sources of variations and their subsequent volatile estimates: the first is, obviously, to increase the number of replications in view to get a better statistical power, although we are aware that this is not always possible in practice. The second is to rely on more sophisticated imputation methods, which generally do not shrink the variance as much as simple ones (such as mean or fixed-value imputations [4]). The third is to rely on improved variance estimates, such as proposed in Limma [8], SAM [9] or Cyber-T [10]. All these methods should be preferred to logFC thresholding, so that we can formulate our first recommendation:

Recommendation 1*: One should not consider the use or the tuning of the logFC cutoff as statistically-motivated. On the contrary, whether it is used or not - and where appropriate - its precise tuning must be motivated by non-statistical arguments relating only to the practitioner's background knowledge of the biological experiment or the post-analytic validations.*

In differential analysis, statistical tools are mostly used as safeguards against false discoveries. As a result, positioning the logFC threshold outside the scope of such statistical guidelines indirectly provides a supplementary tool to perform unintentional "p-value hacking" [11], [12]. "p-value hacking" is what we call any attempt to collect, prepare or preprocess the data, so that an expected result appears more statistically significant than it really is. This hacking may be unintentional - an ill-informed practitioner adjusting the processing options to get the best from their data may be unaware that these adjustments might lead to inconsistencies or claims with no statistical support. In the case of logFC filtering, unintentional p-value hacking may result from either an overfitted logFC threshold (i.e., the tuned value), or an inappropriate combination between the logFC filtering step and other steps in the differential analysis. We will focus on the latter first.

We are aware of three ways to combine logFC filtering with the other steps in differential analysis. The classical approach is to define an (adjusted) p-value filtering step [13] producing a volcano plot representation where joint "vertical" and "horizontal" cutoffs can be independently applied. An alternative approach, popularized by PatternLab [14] and Perseus [15], is to rely on hyperbolic curved thresholds on the volcano plot. The final approach is to include the logFC in the hypothesis tested, as proposed in the TREAT option in Limma [16].

In our view, the "curved threshold" option corresponds to a misuse of the SAM test it is based on [9], making it particularly likely to result in unintentional p-value hacking [17], regardless of the type of data treated [18]. For this reason, it is not implemented in Prostar, and explains why, more generally, we do not recommend it. In contrast, the TREAT approach is statistically supported. Instead of a classical test, for which the null hypothesis for each protein x reads:

$$\mathbf{H0}(x)\colon |\mathrm{logFC}(x)| = 0,$$

another hypothesis incorporating the logFC threshold T (with $T > 0$), is simply tested. This hypothesis reads:

$$\mathbf{H0}'(x)\colon |\mathrm{logFC}(x)| < T,$$

However, this approach has a practical drawback: It requires the user to change their interpretation of the test results (the p-values and the meaning of DA). As an illustration, consider an example where the logFC threshold is set to $T = 2$. Then, a protein with logFC=2.001 and a relatively small variance will have a large p-value when $\mathbf{H0}'$ is tested; as a result, it will most likely be considered non-DA. In contrast, if only proteins for which logFC $\geq T$ are tested under $\mathbf{H0}$, the p-value for this protein will be much smaller and it will be clearly considered to be DA. We found that this second situation, where the logFC cutoff is considered as a safety margin rather than a strict modification of the hypothesis tested clearly reflects the main use of differential abundance analysis in mass spectrometry-based omics studies. Moreover, when using $\mathbf{H0}$, the p-values are not shifted to an extent which depends on

T (as they are with $\mathbf{H0}'$), and consequently the same (adjusted) p-value thresholds can be applied whatever the dataset. This transversality makes some standardization possible in the quality control procedures applied. Combining all these reasons, we make the following recommendation:

Recommendation 2*: Among the various ways in which logFC filtering can be combined with subsequent steps in the differential analysis, the most appropriate one clearly depends on the practitioner's level of expertise:*

- *The most rigorous way is to include the logFC threshold in the statistical test, however, result interpretation will require advanced statistical skills, and this approach is not necessarily compatible with the daily work of an analytical platform where standardized quality control procedures are easier to implement.*
- *For non-experts in statistics, it is generally easier to rely on a combination of vertical and horizontal thresholds on the volcano plot, as it helps to prevent the risk of erroneous interpretations.*
- *Finally, curved thresholds were not originally designed to account for logFC; as a result, using them for this purpose dramatically increases the risk of p-value hacking.*

As Prostar was developed for users that are not necessarily experts in statistics, we chose to implement the combination of vertical and horizontal thresholds only. However, depending on the situation and the level of expertise, any other approach can be justified.

We will now focus on the precise tuning of the logFC threshold value. As already indicated, overfitting the threshold to the data available may result in unintentional p-value hacking. Luckily, applying a threshold of 0 (i.e., no logFC filtering) is always acceptable. Thus, the only risk is to tune it to too high a value as it would help isolating some proteins that are known to be biologically relevant. To avoid doing so, we recommend the following fitting procedure:

Recommendation 3(a)*: The logFC threshold can be appropriately tuned by observing the volcano plot and asking whether the filtered-out protein with the largest absolute logFC would still be filtered out if its p-value were much smaller. If the answer is "yes", the logFC is tuned to an appropriate value. If the answer is "no", the logFC threshold has been tuned to an excessively high value and some "p-value hacking" has probably occurred, albeit unintentionally.*

Recommendation 3a is illustrated in Fig. 1. When observing the volcano plot with the objective of tuning the logFC cutoff, it is essential to: (1) Identify the last filtered-out protein for a given threshold; i.e., the protein with the largest absolute logFC. (2) Imagine that this point is shifted vertically to mimic an excellent p-value. (3) Consider whether, at this theoretical p-value it would still be acceptable to filter it out.

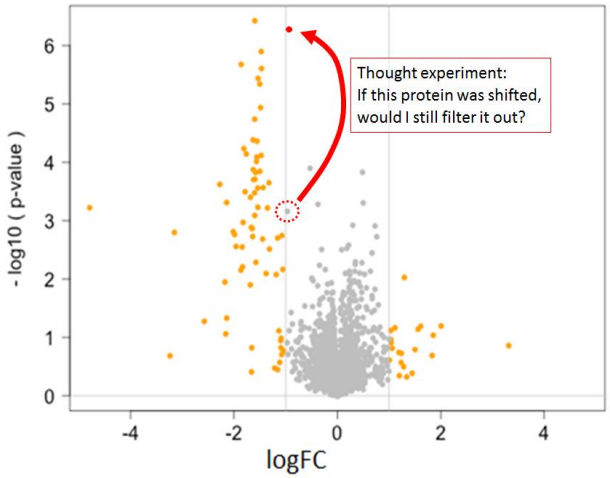

Figure 1: Illustration of the thought experiment that is useful when attempting to tune the logFC without interfering with the p-value distribution, and thus, to safely avoid p-value hacking.

Finally, Recommendation 3(a) helps the practitioner to ignore the p-value information available on the volcano plot, and to focus only on the logFC distribution. To help implement this approach, in the latest Prostar releases the logFC can be tuned without recourse to the volcano plot, even though it initially appears counter-intuitive. In addition, this method respects the next recommendation.

***Recommendation 3(b)**: For datasets with N>2 conditions, where several pairwise comparisons are to be performed, all the pairwise differential analyses should be conducted using the same logFC threshold.*

When comparing *A* vs. *B*, *A* vs. *C*, and *B* vs. *C*, three logFC thresholds must be defined. If one of them differs from the two others, it probably means that it has been overfitted to filter in/out some specific proteins. The reasoning behind this recommendation is that analytic experiments which result in interpretation intrinsically require different fold-change references should be rather scarce. Practically, a large number of superimposed volcano plots can be difficult to visualize and compare: For example, with five conditions, 10 comparisons must be made, as a result the thought experiment from Fig. 1 to tune a single logFC threshold across all 10 volcano plots simultaneously becomes intractable. In contrast, as many logFC distributions can be superimposed without problems (see Supplemental Material, Sec. 1). Finally, we make the following recommendation:

***Recommendation 3**: Getting used to tuning the logFC threshold on logFC distributions rather than on a volcano plot is a good practice, as it helps respect both Recommendations 3(a) and 3(b).*

3. Dealing with recovered intensity values

In this article, "recovered" refers to all the intensity values that cannot be fully trusted as they were obtained through recovery processes. For instance, any intensity value which was missing before an imputation step; or which was retrieved from map alignment rather than direct mass spectrometry evidence.

When comparing two different biological conditions, all proteins containing an overly large proportion of recovered values are customarily filtered out at an early stage, as the evidence for them is too weak to allow them to be confidently considered DA. However, when there are more than two conditions, such early filtering can lead to the loss of biologically-relevant proteins.

Table 1: Hypothetical example of a protein with 12 intensity values spread across three conditions where six values are recovered (noted R).

Condition *A*				Condition *B*				Condition *C*			
24.3	26.2	25.7	24.9	R	23.0	R	R	R	R	R	13.4

In the example presented in Table 1, there are three conditions with four replicates each, and the intensity values are shown for a given protein. This protein will be tested in three different pairwise comparisons:

- In *A* vs. *C*, the protein is expected to be DA, at least on the basis of the non-recovered intensities. It is most likely that in condition *C*, intensity recovery was necessary because the protein abundance was below the lower sensitivity limit for the instrument. Thus, if the recovered intensity values concur with this hypothesis, the differential analysis should lead to a very small p-value.
- In *A* vs. *B*, several scenarios are possible, each more or less plausible, depending on the various recovered values. Notably, if the recovered values are quite high, the protein will be deemed non-DA, so that there is little risk of falsely claiming DA. Conversely, if the recovered values are really lower than the only value directly observed, the protein will be considered DA. This outcome is risky as the claim would only be based on a single non-recovered value, which, in addition, does not perfectly concur with the recovered ones.
- In *B* vs. *C*, only one directly observed value is available per condition, and both are of different magnitudes. There has been a glaring lack of reproducibility of measured values in both conditions, so that the difference in intensity between the two conditions does not seem reliable. Thus, a claim that the protein is DA lacks support, it is therefore best to classify it as non-DA, by default, without considering the recovered values.

Based on these comparisons, it is clear that in *A* vs. *C*, the protein must not be filtered out before differential analysis. Conversely, in *B* vs. *C*, it is wiser to discard it. Finally, in *A* vs. *B*, it will depend on the recovered values.

To apply different filtering thresholds for each pairwise comparison requires only two conditions to be loaded into the software tool, the entire analysis to be performed on this comparison, and the process repeated for other pairwise comparisons. This approach is time-consuming and makes it impossible to have global normalizations across all the conditions, which are sometimes necessary. We therefore formulate the following recommendation:

Recommendation 4*: If possible, it is better to load the replicates of all the biological conditions together, to run a single combined preprocessing.*

A consequence of Recommendation 4 is that a unique filtering pattern must be chosen for all the pairwise comparisons. Thus, in the example above, either the protein is retained in the three comparisons, or it is discarded from all three. With this in mind, we advise:

Recommendation 5*: When loading more than two conditions, very loose filters should be applied which retain any protein which could be of interest in one of the comparisons.*

The rationale behind Recommendation 5 is that if the protein is filtered out, it is lost for good, whereas if the filtering is too loose, it is still possible to correct it *a posteriori*: To do so, the tool simply has to propose a second round of filtering for each pairwise differential analysis. This processing can be implemented in many ways. In Prostar, it corresponds to the "push p-value" option which allows modification of any p-value that is low enough to cause a protein for which the evidence is too weak to be returned as DA. In real terms, these proteins are identified by applying a series of filters, but instead of being discarded, the corresponding p-value is replaced by 1 for the pairwise comparison of interest. Thus, these proteins are still present in the dataset, but no longer interfere with the rest of the differential analysis. To summarize, we formulate a last recommendation:

***Recommendation 6**: Any preliminary loose global filtering can be completed by a more stringent filter operating at a pairwise comparison level. This second round of filtering can be implemented in various ways. It is referred to as "push p-values" in Prostar, where it replaces unreliable p-values by the value 1.*

4. Verifying p-value calibration

Based on the recommendations made above, a certain number of proteins will have been discarded due to an excessively low logFC, or to an overly large proportion of less reliable recovered values among the conditions compared. The remaining proteins will be associated with a p-value reflecting the reliability of results from a differential abundance test (e.g., *t*-test family, Limma, etc.). From this point on, we are interested in selecting the first N proteins with the smallest p-values which will be considered DA. Generally, no prejudice is applied to values of N, and as a result it is not directly tuned. Similarly, N is not defined as the number of proteins that pass a manually tuned p-value threshold, for it would not account for the necessary multiple test correction. The selection process is therefore classically based on a FDR threshold.

As explained in [13], in a differential analysis, an FDR corresponds to a value estimating the False Discovery Proportion (FDP), which is the proportion of proteins that are non-DA but which were erroneously selected as DA. Defining the number of selected proteins based on an FDR is a good way to control the quality of the final proteomics analysis step. However, as with any statistical estimate, the computed FDR can be more or less close to the true value, i.e., the FDP. Thus, if the FDR is to be reliable, it is necessary to check that the statistical assumptions on which its computation is based are fulfilled. With this in mind, the most important assumption is that the p-values for non-DA proteins distribute uniformly, whereas those for the DA proteins should be as small as possible (see [13] for details). If the total number of proteins tested is large enough, this can be easily illustrated on a histogram, as proposed in [19], and presented in Fig. 2. A p-value histogram can be straightforward to interpret. However, when the dataset is too small, the binning may hinder the decision on whether the histogram fits the expectations or not. In these cases, a calibration plot [20] can be used. A calibration plot is a decade-old alternative to the histogram which is not as intuitive to interpret, but which does not depend on a number of bins. Fig. 2 shows how the information on the calibration plot relates to that on the histogram.

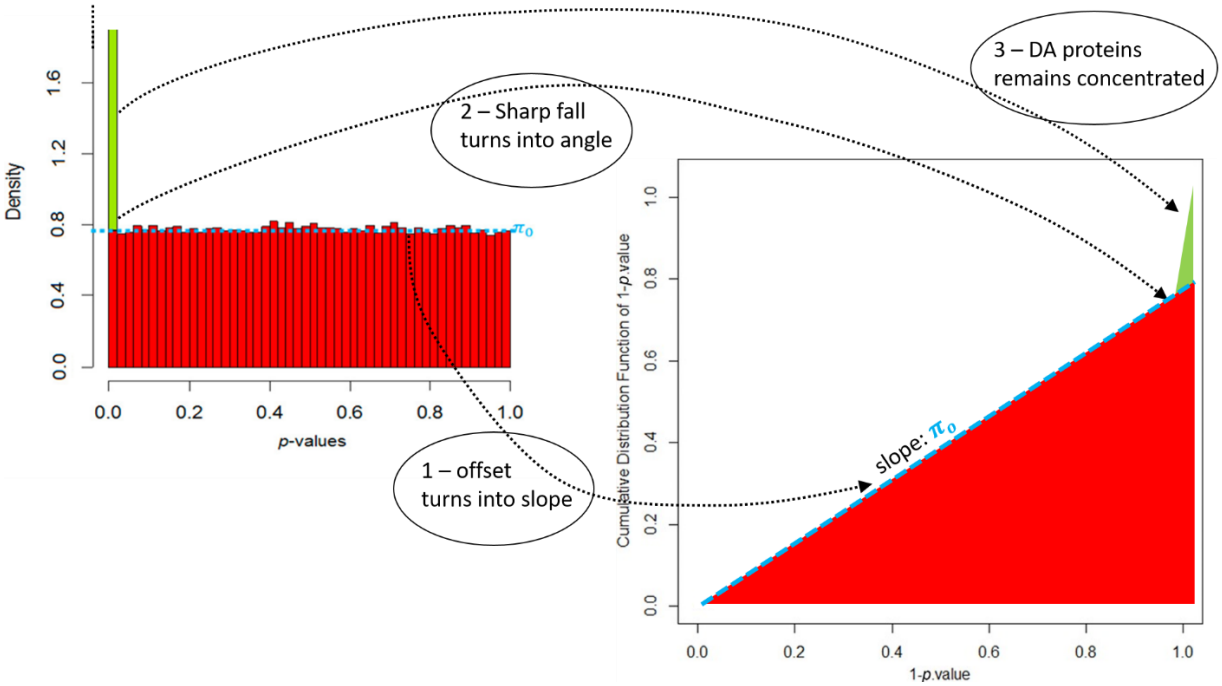

Figure 2: On the left, a histogram of simulated p-values, where a perfectly uniform distribution is observed for non-DA proteins (in red), and small p-values are found for DA proteins (in green). Unfortunately, in practice, histograms are rarely so easy to interpret (see [9] for examples). On the right, a schematic representation of how a histogram was transformed into a calibration plot. As the calibration plot relies on cumulative distributions rather than binning, it displays smoother patterns which smooth out stochastic variations.

We recently adapted calibration plots to the proteomics context [21], to help their interpretation and inclusion in bioinformatics pipelines. As illustrated in Fig. 3, we have added three quality control criteria that are helpful in this context:

- The first one is referred to as “Non-DA protein proportion”. This percentage corresponds to the π_0 value that is indicated in Fig. 2. It is displayed in the same shade of blue as the straight line on the graph (in Fig. 3). As this π_0 value is also a user-defined parameter, it will be fully described in the next section.
- The second one is “DA protein concentration”, displayed in green. It corresponds to a measure of how the green triangle (in the upper right corner of Fig. 2) has a narrow basis. Concretely it indicates the extent to which the DA proteins have their p-values concentrated in a near zero interval. A concentration value close to 100% indicates that there is a clear break between the distribution of non-DA proteins (represented in red in Fig. 2) and that of the DA proteins (in green). Practically, concentrations lower than 85-90% should be considered with caution, as in such cases, the overall distribution deviates too much from that needed to compute a reliable FDR. In these conditions, although an FDR value will be produced by the algorithm, there will be no guarantee that the value obtained is realistically close the true FDP value.
- The last criterion is termed “Uniformity underestimation”, it tends to qualify the extent to which non-DA proteins have a distribution which deviates from uniform. Once again, a certain level of uniformity is necessary to produce a reliable FDR value. This criterion is shown in red in Fig. 3 as the areas under the cumulative distribution (in black) which cross the straight blue line. Thus, any area colored in red above the blue line indicates regions where the uniformity is questionable. Due to random effects, small deviations, such as those illustrated in Fig. 3 are acceptable. However, very large deviations will produce spurious FDR computations. Note that to compute this criterion, one only accounts for the parts of the curve which is above the straight line depicting uniformity. The reason is deviations from below do not question the reliability on the FDR (in fact, increasing the gap from below between the curve and the straight line only makes the FDR more conservative). On the contrary, overshooting the line makes the FDR unreliable as possibly smaller than the FDP More details can be found in [21].

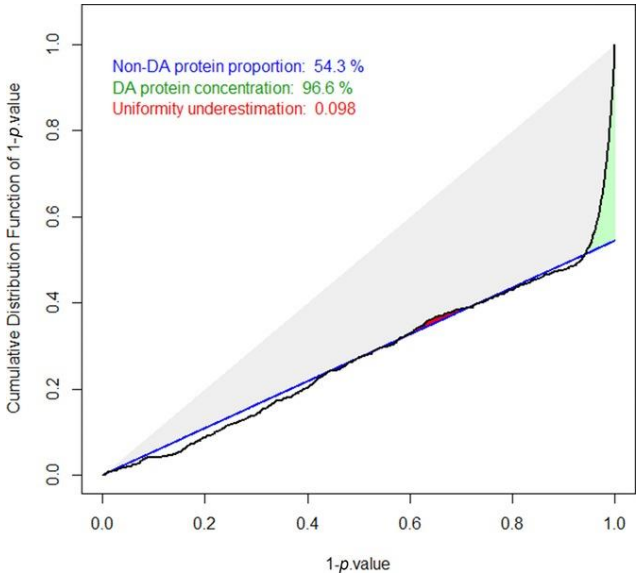

Figure 3: Calibration plot such as provided by CP4P (borrowed from [21]), presenting three quality control criteria

If the "green" and "red" criteria are satisfied, a calibration plot such as the one shown in Fig. 3 is obtained. Several other calibration plots can be found in the Supplemental Material, to illustrate as many cases as possible, ranging from excellent to catastrophic calibrations. Figures depicting calibration problems are accompanied by explanations that can help to understand data producing similar plots. To make things easier, the continuum of calibration quality can be clustered into three groups:

1. **Well-calibrated** (Supplemental Material, Sec. 2): Although quite rare when dealing with real data, these situations are the most comfortable. The experimental data are likely to be of excellent quality, the FDR will be accurately estimated and generally, any biological claim will be clearly supported by the data.
2. **Ill-calibrated** (Supplemental Material, Sec. 3): This situation arises most frequently. In general, the plot does not depict well-calibrated p-values due to one or several of the following: (1) underestimation of uniformity (in red); (2) overly-low DA protein concentration (in green); (3) discontinuities in the cumulative distribution function (steps). However, the magnitude of the problem is small enough to be dealt with.
3. **Miscalibrated** (Supplemental Material, Sec. 4): the data distribution presents one or several of the above-mentioned problems at a magnitude that is beyond what can be corrected.

In this context, we formulate the following recommendations:

***Recommendation 7**: First, determine whether the p-values are well-calibrated, ill-calibrated, or miscalibrated: If well-calibration, no specific action is required. If ill-calibrated, apply **Recommendation 8**. If miscalibrated, apply **Recommendation 9**.*

***Recommendation 8**: An FDR can still be computed in the presence of ill-calibrated p-values. However, the risk it underestimates the FDP increases due to ill-calibration. As such underestimation may lead to unsupported claims, "larger margins" are necessary: one has to increase the FDR (such FDR is said "conservative"). The tricks presented in the next sections to recover correct calibration have such effect. As a result, the data owner may be tempted to minimize the calibration correction, so as to obtain a smaller FDR; this temptation should be avoided.*

***Recommendation 9**: For miscalibrated p-values, the data owner is faced with a real problem: on the one hand, it remains possible to blindly compute an FDR (even if it should not be trusted) and to proceed with the analysis, at the risk of providing unsupported conclusions. On the other hand, scientific rigor would suggest that the data cannot be exploited, an outcome which is not always acceptable for a variety of reasons (no other samples available, experimental cost, etc.). However, a compromise is possible, to reconcile the project constraints and statistical validity: in fact, numerous miscalibrations are not related to the raw data themselves, but to the various bioinformatic post-analysis steps (ranging from peptide identification to the first step of the differential analysis). Thus, we advise the data owner to review the whole data processing pipeline, taking the opportunity to increase its robustness. If carefully conducted, a clear reduction of the miscalibration can be expected, hopefully sufficient for the data to be considered only as ill-calibrated.*

5. Estimating the proportion of non-DA proteins

In the histogram in Fig. 2, π_0 corresponds the heights of the red uniform distribution. It represents the average proportion of non-DA proteins, whether selected or not (i.e., it is different from the FDP and the FDR). This proportion is essential in the computation of the FDR: it need not be precisely known, but it must not be underestimated, as otherwise it would produce a spurious (anti-conservative) FDR. In contrast, its overestimation only leads to FDR overestimation (which amounts to making cautious or conservative claims). In the original Benjamini-Hochberg procedure [22], π_0 was tuned to the maximum value (*i.e.* 1) which amounts to the most cautious posture.

Naturally, everyone wants to get the most from their data, and it is therefore justified to limit the FDR overestimation. To do so, π_0 must be tuned to a value that is strictly smaller than 1, while nevertheless avoiding underestimation. In fact, this practice is already widely used in biostatistics, notably when controlling the FDR at peptide-identification-level by using target-decoy competition (see for instance [23], [24]). However, for unknown reasons, it has not been applied in quantitative mass spectrometry analysis, even though numerous π_0 estimators have been proposed in the statistical literature (see [21] for a proteomics-oriented overview).

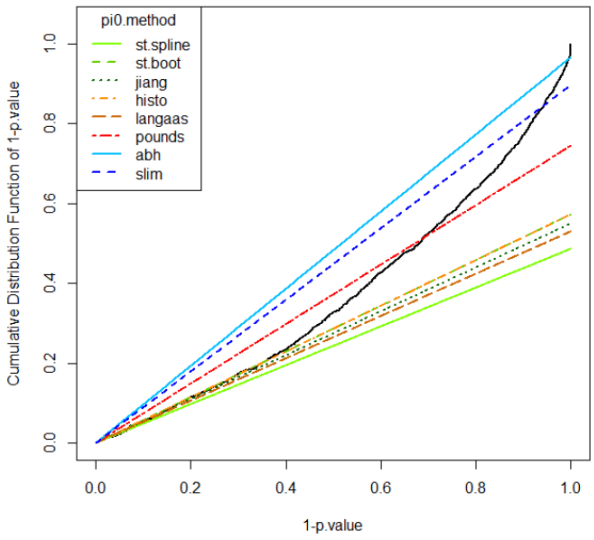

Figure 4: A calibration plot for which all the π_0 estimates are rather different, which was to be expected, for a cumulative distribution function that appears ill-calibrated.

Classically these estimators are based on a variety of methodologies, producing several estimates for the same dataset. Fig. 4 shows a calibration plot with various π_0 estimates. Intuitively, when choosing the estimator providing the greatest π_0 value, i.e., that producing the steepest line, two risks are reduced: that of a non-uniform distribution (as increasing the slope reduces the red-shaded area on

the calibration plot) and that of lack of concentration (as increasing the slope makes the base of the green triangle thinner).

***Recommendation 10**: Increasing π_0 is a simple way to improve calibration. Notably, it can efficiently recover calibration from ill-calibrated data as a result of either (1) uniformity underestimation, or (2) insufficient DA concentration.*

At this point, we provide a warning: estimating π_0 requires a full p-value distribution, which is not available anymore after logFC filtering. Thus, combining both processing steps requires some precautions (as implemented in CP4P [21], see Supplemental Material, Sec. 1, for details).

For some data, all the π_0 estimators provide rather similar estimates, so that in practice, choosing one or another estimate will make little difference, and as a result the possibility of improving the calibration is limited. However, this is inherently not a problem as convergent estimates tend to be a sign of good calibration:

***Recommendation 11**: When comparing various π_0 estimates, the extent to which they concur or not is an additional indicator of the quality of the calibration.*

In spite of this trend, there remain datasets for which it is difficult to obtain good calibration. For these last cases, the following recommendation can be applied:

***Recommendation 12**: In case of doubt regarding ill-calibration, the default rule is to rely on Benjamini-Hochberg estimate, i.e. π_0=1. This corresponds to the most conservative case. If application of this value is not sufficient to reach correct calibration, the data are definitively miscalibrated.*

It should be noted that, to improve readability, the line corresponding to π_0=1 is not presented on the calibration plot. Indeed, this line never changes and should follow the diagonal of the plot, so that using the Benjamini-Hochberg procedure does not require it to be explicitly displayed.

6. Optimizing of the numbers of DA and non-DA proteins

Previously, we hinted that calibration plots depicting a cumulative distribution function with steps should raise concerns (see Supplemental Material, Sec. 3). One reason for this is that, obviously, this shape impedes plot interpretation, in the same way that binning interferes with histogram reading. On histograms, binning remains visible unless enough data are available. However, calibration plots are less sensitive to data size, as a result, stepped cumulative distribution functions are a sign of a rather small number of tested proteins. In itself, a very small number of tested proteins is an issue for several reasons.

The most obvious one is that an FDR corresponds to a percentage. Thus, to remain meaningful, the dataset should be large enough for the chosen percentage to have a physical counterpart. For instance, if 30 proteins are selected as DA, while applying an FDR of 1%, an interpretation problem arises: less than 0.333 falsely attributed DA proteins are expected and the FDP cannot be guaranteed to be different from 0%. With such results, the FDR control provides little support. However, this is not necessarily a problem: just because the FDR is meaningless does not mean the proteomics experiment is also meaningless. The point is mainly that such a small FDR threshold cannot be used as a quality control metric for small datasets. Classically, the FDR must be controlled on datasets that are too big to allow individual inspection of each DA protein. In these cases, an overall quality control such as FDR is mainly used as an acceptable surrogate for individual quality control on each putative DA protein. Thus, if the dataset is too small to compute an FDR safely, it makes sense to resort to other metrics, such as the Bonferroni family-wise error rate (FWER, [25]), which is more appropriate for small sets of proteins; or the Posterior Error Probability (PEP, [26]), which provides a metric specific to each selected protein.

***Recommendation 13**: If the number of DA proteins is too small to make the chosen percentage meaningful, then it makes sense to increase the threshold. Alternatively, the FDR can be replaced by other metrics such as Bonferroni FWER or PEP. These metrics are not provided in the current Prostar release, however, they can be computed a posteriori with any tool, using the list of p-values as input.*

The appearance of steps in the cumulative distribution function is not only a reflection of the number of DA proteins selected: DA proteins are found in the upper right corner of the plot and cannot explain steps appearing elsewhere on the plot. The latter appear when the number of non-DA proteins is too small. In fact, if too few non-DA proteins are present, then the same issue as the binning of the red histogram (on Fig. 2) emerges. This can lead to inaccurate estimate of π_0, which would then impact the FDR quality. Therefore, confidently estimating an FDR requires a sufficiently large number of non-DA proteins. This sounds counter-intuitive in the first place, but one has to keep in mind that the non-DA proteins should not necessarily look like DA proteins (and be selected by mistake): they are necessary to help identify the pattern-change (between DA and non-DA proteins), as illustrated by point 2 in Fig. 2.

The easiest way to retain sufficient numbers of non-DA proteins in the differential analysis is to avoid discarding them at earlier steps. To do so, previous recommendations, regarding stringent logFC cutoff, as well as loose filters (balanced by using a processing akin to “push p-value”) can be applied:

***Recommendation 14**: Steps in the left-hand side of calibration plots can be reduced or removed by using looser cutoff values in preliminary filtering steps (on logFC or on recovered values).*

Once the calibration plot depicts sufficiently well-calibrated p-values, a set of DA proteins can be selected for display on a volcano plot. At this point, some instability may still be observed in the FDR computation. Notably, a slight change in the horizontal threshold can lead to considerable variations in the FDR values. This variability should be interpreted similarly to a binning effect: The volume of data is insufficient to produce smooth variations of the FDR as a function of the p-value threshold. To balance this, the logFC threshold can be adjusted to looser cutoff.

***Recommendation 15**: Increasing the number of non-DA proteins is also an efficient way to reduce FDR instability with respect to the p-value threshold that might be observed on the volcano plot.*

7. Conclusions

Differential analysis is classically the final stage in any mass spectrometry-based proteomics experiment. As such, it combines all the imprecisions or mistakes accumulated from the sample preparation to the mass spectrometry measurements and bioinformatics involved in biomolecule identification and quantification. The resulting data may have characteristics and qualities that vary according to the problems studied and instruments used, making the strict application of statistical guidelines difficult. With this in mind, statistical protocols may be blindly and routinely applied, with the risk of generating doubtful results; or the data and the theory can be matched as much as possible. However, this alternative requires watchful eyes, as well as expertise unrelated to experience in analytical chemistry. Fortunately, modern bioinformatics tools provide visualization interfaces for large proteomics datasets that make it possible for the analyst to scrutinize each step in data processing. In this context, this article splits the differential analysis into five essential steps, the quality of which can be visually assessed throughout the process. Using these steps as a guideline will help proteomics and other mass spectrometry-based omics communities to communicate more reliable biological results. Moreover, they will help familiarize users with the underlying statistical concepts, leading researchers to better handle the ever-growing experimental data they produce.

Acknowledgments

This work was supported by grants from the French National Research Agency: ProFI project (ANR-10-INBS-08), GRAL project (ANR-10-LABX-49-01) and SYMER project (ANR-15-IDEX-02).

References

[1] Callister, S. J., Barry, R. C., Adkins, J. N., Johnson, E. T., Qian, W. J., Webb-Robertson, B. J. M.,... & Lipton, M. S. (2006). Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *Journal of proteome research*, *5*(2), 277-286.

[2] Hrydziuszko, O., & Viant, M. R. (2012). Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. *Metabolomics*, *8*(1), 161-174.

[3] Webb-Robertson, B. J. M., Wiberg, H. K., Matzke, M. M., Brown, J. N., Wang, J., McDermott, J. E.,... & Waters, K. M. (2015). Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *Journal of proteome research*, *14*(5), 1993-2001.

[4] Lazar, C., Gatto, L., Ferro, M., Bruley, C., & Burger, T. (2016). Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *Journal of proteome research*, *15*(4), 1116-1125.

[5] Wieczorek, S., Combes, F., Lazar, C., Giai Gianetto, Q., Gatto, L., Dorffer, A.,... & Burger, T. (2016). DAPAR & ProStaR: software to perform statistical analyses in quantitative discovery proteomics. *Bioinformatics*, *33*(1), 135-136.

[6] Stratton, K. G., Webb-Robertson, B. J. M., McCue, L. A., Claborne, D., Stanfill, B., Godinez, I.,... & Bramer, L. M. (2019). pmartR: Quality Control and Statistics for Mass Spectrometry-based Biological Data. *Journal of proteome research*.

[7] Xia, J., Psychogios, N., Young, N., & Wishart, D. S. (2009). MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic acids research*, *37*(suppl_2), W652-W660.

[8] Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor* (pp. 397-420). Springer, New York, NY.

[9] Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, *98*(9), 5116-5121.

[10] Baldi, P., & Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, *17*(6), 509-519.

[11] https://en.wikipedia.org/wiki/Data_dredging

[12] Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS biology*, *13*(3), e1002106.

[13] Burger, T. (2017). Gentle Introduction to the Statistical Foundations of False Discovery Rate in Quantitative Proteomics. *Journal of proteome research*, *17*(1), 12-22.

[14] Carvalho, P. C., Yates III, J. R., & Barbosa, V. C. (2012). Improving the TFold test for differential shotgun proteomics. Bioinformatics, 28(12), 1652-1654.

[15] Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T.,... & Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote) omics data. *Nature methods*, *13*(9), 731.

[16] McCarthy, D. J., & Smyth, G. K. (2009). Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*, *25*(6), 765-771.

[17] Giai Gianetto, Q., Couté, Y., Bruley, C., & Burger, T. (2016). Uses and misuses of the fudge factor in quantitative discovery proteomics. *Proteomics*, *16*(14), 1955-1960.

[18] Larsson, O., Wahlestedt, C., & Timmons, J. A. (2005). Considerations when using the significance analysis of microarrays (SAM) algorithm. *BMC bioinformatics*, *6*(1), 129.

[19] Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, *100*(16), 9440-9445.

[20] Schweder, T., & Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, *69*(3), 493-502.

[21] Giai Gianetto, Q., Combes, F., Ramus, C., Bruley, C., Couté, Y., & Burger, T. (2016). Calibration plot for proteomics: A graphical tool to visually check the assumptions underlying FDR control in quantitative experiments. *Proteomics*, *16*(1), 29-32.

[22] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300.

[23] Käll, L., Storey, J. D., MacCoss, M. J., & Noble, W. S. (2007). Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of proteome research*, *7*(01), 29-34.

[24] Keich, U., Kertesz-Farkas, A., & Noble, W. S. (2015). Improved false discovery rate estimation procedure for shotgun proteomics. *Journal of proteome research*, *14*(8), 3148-3161.

[25] Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual review of psychology*, *46*(1), 561-584.

[26] Käll, L., Storey, J. D., MacCoss, M. J., & Noble, W. S. (2007). Posterior error probabilities and false discovery rates: two sides of the same coin. *Journal of proteome research*, *7*(01), 40-44.